FAKE

COMO AS PLATAFORMAS ENFRENTAM A DESINFORMAÇÃO

NEWS

FAKE

COMO AS PLATAFORMAS ENFRENTAM A DESINFORMAÇÃO

NEWS

realização
INTERVOZES
COLETIVO BRASIL DE COMUNICAÇÃO SOCIAL

Multifoco

intervozes
coletivo brasil de
comunicação social

GRUPO MULTIFOCO
Rio de Janeiro, 2021

CREATIVE COMMONS (CC BY-SA 4.0) © 2021 INTERVOZES – COLETIVO BRASIL DE COMUNICAÇÃO SOCIAL

DIREÇÃO EDITORIAL Grupo Multifoco
EDIÇÃO Equipe Multifoco
AUTORES Bia Barbosa, Helena Martins e Jonas Valente
CONSELHO DIRETOR (2020-2021) Alex Hercog, André Pasti, Bruno Marinoni, Marcos Urupá, Mônica Mourão, Paulo Victor Melo e Tâmara Terso
COORDENAÇÃO EXECUTIVA Gyssele Mendes (Comunicação), Iara Moura (Desenvolvimento Institucional e Articulação Interna), Maria Mello (Incidência e Comunicação), Marina Pita (Incidência), Olívia Bandeira (Pesquisa, Formação e Articulação Internacional) e Pedro Ekman (Litigância, Sustentabilidade e Administrativo)
REVISÃO Isabel Rodrigues
REVISÃO TÉCNICA Olívia Bandeira
PROJETO GRÁFICO E CAPA Caroline Silva
FOTOGRAFIA DE CAPA Sergey Zolkin e Tom Ramalho | Unsplash
IMPRESSÃO Gráfica Multifoco

APOIO

FORD FOUNDATION

DIREITOS RESERVADOS A
GRUPO MULTIFOCO
Av. Mem de Sá, 126 - Centro / 20230-152
Rio de Janeiro, RJ / Tel.: (21) 2222-3034
contato@editoramultifoco.com.br
www.editoramultifoco.com.br

TODOS OS DIREITOS RESERVADOS.
Nenhuma parte deste livro pode ser utilizada ou reproduzida sob quaisquer meios existentes sem autorização por escrito dos editores e autores.

Dados Internacionais de Catalogação na Publicação (CIP).

I61	Intervozes – Coletivo Brasil De Comunicação Social. Fake News: como as plataformas enfrentam a desinformação / Intervozes – Coletivo Brasil De Comunicação Social. – Rio de Janeiro : Multifoco, 2021. 309 p. ; 23 cm. Inclui bibliografia ISBN: 978-65-5611-099-8 1. Comunicação social 2. Fake News 3. Mídias sociais I. Título
21-0034	CDD: 302.234

Fernanda Silvino – Bibliotecária – CRB-7 7230/O

SUMÁRIO

INTRODUÇÃO: O QUE É DESINFORMAÇÃO E COMO O PROBLEMA SE APRESENTA NO BRASIL **9**

1. A ASCENSÃO DAS PLATAFORMAS DIGITAIS E O PROBLEMA DA DESINFORMAÇÃO **19**

1.1. Perspectiva teórico-metodológica e categorias de análise da pesquisa 20

1.2. Conceito e características de funcionamento das plataformas digitais 28

1.3. Panorama do contexto regulatório sobre desinformação 34

1.4. Abordagens sob a perspectiva da defesa dos direitos humanos 41

1.4.1 Conceito utilizado e enquadramento do problema 44

1.4.2 Abordagem regulatória e/ou agente(s) ao(s) qual(is) as propostas são dirigidas 46

1.4.3 Medidas recomendadas 52

1.4.4 Avaliação da possível moderação de conteúdos por parte das plataformas 61

1.5. Enfrentando a desinformação: conceito e práticas defendidas pelo Intervozes 65

2. MEDIDAS ADOTADAS PELAS PLATAFORMAS DIGITAIS 75
2.1. Facebook 75
2.1.1 Abordagem do fenômeno 77
2.1.2 Moderação de conteúdo 79
2.1.3 Promoção de informações e transparência 90
2.1.4 Medidas correlatas 96
2.2. Instagram 106
2.2.1 Abordagem do fenômeno 108
2.2.2 Moderação de conteúdo 110
2.2.3 Promoção de informações e transparência 113
2.2.4 Medidas correlatas 115
2.3. WhatsApp 121
2.3.1. Abordagem do fenômeno 123
2.3.2. Moderação de conteúdo 126
2.3.3. Promoção de informações e transparência 127
2.3.4. Medidas correlatas 135
2.4. YouTube 145
2.4.1. Abordagem do fenômeno 146
2.4.2. Moderação de conteúdo 152
2.4.3. Promoção de informações e transparência 161
2.4.4. Medidas correlatas 169
2.5. Twitter 175
2.5.1. Abordagem do fenômeno 176
2.5.2. Moderação de conteúdo 179
2.5.3. Promoção de informações e transparência 195
2.5.4. Medidas correlatas 201

3.ANÁLISE DAS MEDIDAS ADOTADAS PELAS PLATAFORMAS DIGITAIS 211
3.1 Facebook 211
3.1.1 Abordagem do fenômeno 214
3.1.2 Moderação de conteúdo 215
3.1.3 Promoção de informações e transparência 220
3.1.4 Medidas correlatas 224

3.2 Instagram 228
3.2.1 Abordagem do fenômeno 229
3.2.2 Moderação de conteúdo 229
3.2.3 Promoção de informações e transparência 231
3.2.4 Medidas correlatas 233
3.3 WhatsApp 234
3.3.1 Abordagem do fenômeno 236
3.3.2 Moderação de conteúdo 240
3.3.3 Promoção de informações e transparência 244
3.3.4 Medidas correlatas 247
3.4 YouTube 252
3.4.1 Abordagem do fenômeno 253
3.4.2 Moderação de conteúdo 254
3.4.3 Promoção de informações e transparência 260
3.4.4 Medidas correlatas 263
3.5 Twitter 265
3.5.1 Abordagem do fenômeno 267
3.5.2 Moderação de conteúdo 269
3.5.3 Promoção de informações e transparência 274
3.5.4 Medidas correlatas 275
3.6 Análise comparativa das medidas de enfrentamento à desinformação 279
3.6.1 Abordagem do fenômeno 280
3.6.2 Moderação de conteúdo 284
3.6.3 Promoção de informações e transparência 290
3.6.4 Medidas correlatas 293

BIBLIOGRAFIA 303
DOCUMENTOS INTERNACIONAIS 308
OUTROS DOCUMENTOS 309

INTRODUÇÃO

O QUE É DESINFORMAÇÃO E COMO O PROBLEMA SE APRESENTA NO BRASIL

Se a disseminação de conteúdos enganosos e deliberadamente forjados sempre existiu, ao longo dos tempos, no debate público, as plataformas digitais elevaram-na a um novo patamar. Em primeiro lugar, porque permitem um alcance sem precedentes a tais conteúdos, na medida em que possibilitam o contato simplificado com perfis criados de maneira simples e que reúnem milhares ou até milhões de conexões (amigos ou seguidores). Em segundo, pela velocidade da circulação desses conteúdos, gerada pela progressão geométrica dos compartilhamentos, tornando muito complexo o acompanhamento e a adoção de medidas de contrainformação ou de reparação. Em terceiro lugar, porque o tratamento massivo e abusivo de nossos dados pessoais abre espaço para uma segmentação no direcionamento do conteúdo, o que torna mais efetivas as táticas de manipulação da opinião. Por fim, porque as arquiteturas informacionais e parâmetros adotados no funcionamento das plataformas potencializam conteúdos extremos e que buscam engajamento e reações dos usuários, que, por sua vez, são bem

recompensados via recursos publicitários dentro dos modelos de negócio dessas empresas.

Assim, ao mesmo tempo em que permitiram maior debate público, produção e circulação de informações por não especialistas, as plataformas também criaram um modelo de negócios em que determinadas informações são mais bem disseminadas e ganham mais alcance do que outras. Ou seja, as plataformas, em benefício de seus modelos de negócios, "barateara m" e "democratizaram" a possibilidade de estratégias enganosas, de manipulação do debate público, do discurso de ódio e dos ataques a adversários políticos, de maneira a transformá-las, também, num lócus por excelência das práticas desinformativas. Tendo em vista a ascensão das plataformas como arenas privilegiadas do debate público conectado, mas também a falta de transparência por parte das empresas e de ausência de controle real ou priorização, por parte dos usuários, dos conteúdos que recebem, há uma mistura explosiva, que se transformou num risco à própria democracia e às suas instituições. Em tempos de pandemia, os danos da desinformação evidenciam que a transferência do debate público para as plataformas digitais traz, ainda, riscos reais à saúde e à vida de pessoas.

No contexto da pandemia causada pelo novo coronavírus (Covid-19), o cenário de emergência na saúde pública fez com que as plataformas passassem a adotar filtros mais duros sobre o fluxo de informações no ambiente on-line. Mas quando dizem que adotam medidas de enfrentamento às chamadas "fake news", termo que foi popularizado no debate público sobre o problema, a que tipo de conteúdo as plataformas digitais se referem?

Na academia e nos organismos internacionais cujos mandatos envolvem a área das comunicações ou de direitos humanos, desde 2018 até 2020, período abarcado pela pesquisa, a expressão "fake news" passou a ser abandonada, sobretudo por ser considerada insuficiente para explicar o fenômeno. Em seu lugar, pesquisadores e especialistas em liberdade de expressão passaram a trabalhar

prioritariamente com o conceito de desinformação, trazendo para sua definição aspectos como consciência do conteúdo forjado ou enganoso e intencionalidade na sua distribuição.

Uma das fragilidades da abordagem do fenômeno tido como "fake news", portanto, é que ele oculta outras situações relacionadas à produção de notícias que podem levar a um distanciamento do que é considerado verdadeiro. Nesse sentido, Wardle e Derakhsan (2017) fazem uma distinção entre "*mis-information*", que seria a difusão de informação falsa, embora sem intuito de prejudicar, e "*des-information*", que envolveria conteúdo sabidamente falso, fabricado ou manipulado com intenção de gerar dano. Ou seja, não basta apenas a falta de base factual ou a intenção de causar dano, seja por motivações políticas e/ou econômicas. Os dois aspectos, para os autores, devem estar combinados.

Em 2018, em *Uma abordagem multidimensional à desinformação – Relatório do Grupo de Alto Nível independente sobre notícias falsas e desinformação on-line,* o grupo de especialistas da Comissão Europeia criado para debater o fenômeno definiu desinformação como "todas as formas de falsidade, informações inexatas ou enganosas concebidas, apresentadas e promovidas para causar intencionalmente danos públicos ou para fins lucrativos" (HLEG, 2018, p. 05), excluídas erros jornalísticos, comentários políticos e distorções deliberadas como a sátira e a paródia. Dentro da ideia de danos ou prejuízos públicos, a União Europeia inclui ameaças a processos políticos democráticos e a bens públicos como a proteção da saúde, o ambiente e a segurança (CE, p. 04).

Para o Intervozes – Coletivo Brasil de Comunicação Social, a desinformação não se fundamenta, entretanto, somente em informações deliberadamente falsas, mas também em distorções geradas por exagero, omissão ou descontextualização, como já ocorria e segue ocorrendo na mídia tradicional. A diferença, como dito anteriormente, são as formas de produção, circulação, mediação e

acesso à informação características das plataformas digitais (MARINONI, GALASSI, 2020, p. 37). Esta pesquisa trabalha, assim, com esse conceito para as análises que se seguem.

Se a ideia de desinformação como tal também não é nova na arena pública de debates do Brasil, é importante destacar que o fenômeno mudou de proporções a partir da campanha eleitoral de 2018. Meses antes do pleito que elegeu deputados, senadores, governadores e o Presidente da República, o presidente do Tribunal Superior Eleitoral (TSE), ministro Luiz Fux, declarou, durante seminário do órgão sobre "Impactos Sociais, Políticos e Econômicos das Fake News", que se o resultado das eleições fosse fruto das notícias falsas, as mesmas poderiam ser anuladas. Na época, a afirmação soou exagerada, mas o volume de desinformações que tomou conta do processo confirmou os receios de quem já acompanhava o funcionamento das redes.

Pesquisa IDEIA Big Data/Avaaz[1] divulgada em outubro de 2018, logo após o segundo turno, apontou que mais de 98% dos eleitores de Bolsonaro foram expostos a uma ou mais notícias falsas durante a campanha – e que mais de 89% deles acreditaram que tratava-se de verdades. Pesquisa da Atlas Político[2] mostrou um número menor, mas que mesmo assim é altíssimo: 36% do eleitorado brasileiro tinha acreditado que Fernando Haddad, candidato do PT, havia distribuído um "kit gay" para crianças nas escolas quando ocupou o Ministério da Educação. Mesmo desmentida pelo TSE, a desinformação seguiu sendo compartilhada nas redes e foi ao ar na última semana de propaganda eleitoral no rádio e na TV. Mais de 50 milhões de brasileiros compraram a história. Somente no final de semana do primeiro turno, a agência de checagens Lupa classificou como inverídicas 12 informações que, somadas, tiveram mais de 1,17 milhão de compartilhamentos no

1. https://secure.avaaz.org/act/media.php?press_id=917
2. https://agenciabrasil.ebc.com.br/geral/noticia/2018-11/noticias-falsas-influenciaram-eleicoes-deste-ano-dizem-pesquisadores

Facebook. Não se sabe a quantas pessoas a verificação dos fatos classificados como inverídicos chegou.

A chefe da missão da Organização dos Estados Americanos (OEA), que acompanhou as eleições daquele ano no Brasil, sintetizou como um "fenômeno sem precedentes" sua análise sobre o uso das redes sociais para disseminação de desinformação no debate político no país[3]. A maior preocupação de Laura Chinchilla foi com o WhatsApp. Se, de fato, em se tratando de Facebook e YouTube o TSE tomou diversas decisões – mesmo que insuficientes –, determinando remoções de conteúdos comprovadamente falsos, no WhatsApp o caráter privado das comunicações impediu qualquer esforço de frear as disseminações. Levantamento dos professores Pablo Ortellado (USP) e Fabrício Benevenuto (UFMG), em parceria com a agência de checagem Lupa, feito em 347 grupos na plataforma durante o primeiro turno de 2018 revelou que apenas 8% das principais imagens compartilhadas no período no aplicativo de mensageria eram verdadeiras[4].

Matéria da *Folha de S. Paulo* também denunciou o financiamento, via caixa 2, de contratos de dezenas de milhões de reais para disparos de também milhões de mensagens por empresas de marketing digital[5]. A ilegalidade denunciada, cujo rumo das investigações hoje segue desconhecido, pode não ter influenciado apenas o resultado da disputa presidencial, mas também para os governos estaduais. Ações pedindo a impugnação de chapas como a do presidente Jair Bolsonaro e seu vice Hamilton Mourão com base nessas informações aguardam julgamento no TSE. Os indícios de ação automatizada em múltiplos grupos públicos de WhatsApp, constatados em estudos da época como o realizado pelo Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS),

3. https://congressoemfoco.uol.com.br/de-olho-no-combate-as-fake-news/
4. https://piaui.folha.uol.com.br/lupa/2018/10/17/whatsapp-lupa-usp-ufmg-imagens/
5. https://congressoemfoco.uol.com.br/opiniao/colunas/caso-whatsapp-bolsonaro-omissao-aprofunda-impacto-de-noticias-falsas/

estão atualmente no centro das análises da Comissão Parlamentar Mista de Inquérito (CPMI) da Câmara e do Senado sobre o tema, instalada em 2019.

Após o alerta vermelho aceso durante as eleições, os últimos dois anos foram de muito debate em busca de medidas para enfrentar, na escala e velocidade necessárias, a desinformação. No Congresso Nacional, além da abertura da CPMI, surgiram dezenas de projetos de lei tratando do tema. Em sua maioria, eles propõem a criminalização da produção e distribuição de "notícias falsas" com penas de reclusão que chegam a oito anos para qualquer pessoa que compartilhar este tipo de conteúdo. A minoria se propôs a definir o conceito de desinformação. Outra parte das propostas legislativas transfere para as plataformas o dever de definir o que são conteúdos falsos e removê-los, assim como as contas que os propagam, imediatamente. Para os relatores internacionais para a liberdade de expressão, a prática pode institucionalizar a censura privada na Internet.

Em agosto de 2019, o Congresso definiu pela inclusão de um novo dispositivo para a Lei Eleitoral que prevê pena de dois a oito anos de reclusão para quem, comprovadamente ciente da inocência de um candidato, divulgar uma notícia falsa sobre o mesmo durante as eleições. A Lei Eleitoral também já proíbe o uso de perfil falsos e robôs (contas automatizadas, geridas por programas de computador, que simulam ou substituem atividades humanas na distribuição de conteúdo) para propaganda eleitoral, a compra de banco de cadastros de usuários de Internet e telecomunicações e o envio massivo, automatizado e/ou orquestrado de mensagens, como aconteceu no caso do WhatsApp. A mesma lei também traz punições para casos de divulgação de informações inverídicas e o Código Penal criminaliza casos de injúria, calúnia e difamação.

Em 2020, foi apresentada e votada no Senado, em menos de dois meses, a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet, que segue em discussão na Câmara no

momento em que finalizamos esta pesquisa. Inicialmente apresentada como uma lei de combate à desinformação, tem sido objeto de intensos debates e de diferentes proposições, que vão desde medidas de cunho vigilantista a propostas de garantia de maior transparência e devido processo na adoção de medidas, especialmente por parte das plataformas digitais. O debate em torno da questão tem mostrado que o que precisamos é de um extenso debate público, de envolvimento de diferentes agentes, de conscientização da sociedade e da aplicação do marco normativo brasileiro já existente pela Justiça.

As autoridades eleitorais brasileiras não parecem, entretanto, dispostas a assumir sua tarefa nesse sentido. Em 2018, o então vice-procurador eleitoral, Humberto Jacques de Medeiros, chegou a falar a jornalistas que o problema da desinformação não era tão grave e não ajuizou ações contra perfis difusores de conteúdos enganosos, e a então presidente do TSE, ministra Rosa Weber, também disse que a Justiça Eleitoral estava fazendo seu papel. Antes de tomar posse como novo presidente do TSE, o ministro Luís Roberto Barroso, mesmo admitindo que a disseminação massiva de desinformação hoje ameaça as instituições e a democracia brasileira, declarou que o papel da Justiça Eleitoral em combatê-las "deve ser residual", jogando para as plataformas digitais, para as agências de checagem, a imprensa e, em última instância, para o cidadão comum a missão de enfrentar o problema.

As plataformas não ficaram paradas, sobretudo em função de processos extrajudiciais que sofreram em outros países e da própria opinião pública, que passou a questionar significativamente a responsabilidade das empresas nesse processo. A efetividade das medidas adotadas, entretanto, é questionável e a desinformação segue presente nas plataformas. No contexto da pandemia causada pelo coronavírus, pesquisa da Fundação Oswaldo Cruz (Fiocruz) divulgada em abril de 2020 mostrou que 73,7% das notícias falsas sobre o novo coronavírus circularam pelo WhatsApp. Ou-

tros 10,5% foram publicadas no Instagram e 15,8% no Facebook[6]. A pesquisa foi desenvolvida a partir da análise de denúncias e notícias falsas recebidas pelo aplicativo *Eu Fiscalizo*, entre 17 de março e 10 de abril de 2020.

Levantamento da Avaaz sobre desinformação e Covid-19 apontou que 57% dos inquiridos não viram nenhuma correção ou mensagem de alerta sobre conteúdo falso ou enganoso no Facebook. Apenas 34% dos entrevistados afirmaram ter visto tais informações. O presente estudo busca, justamente, jogar mais luz sobre as iniciativas das plataformas digitais neste campo, sobretudo porque, de acordo com o Digital News Report 2020[7], do Instituto Reuters, 54% dos brasileiros e brasileiras usam o Facebook para acessar notícias; 48% usam o WhatsApp; 45%, o YouTube; 30%, o Instagram e 17% utilizam o Twitter.

Se essas plataformas se transformaram em espaço fértil para a disseminação de desinformação, é o direito dos cidadãos à informação que segue sendo cotidianamente atingido. Não à toa, ainda de acordo com o Instituto Reuters, o Brasil é o país com o maior índice de preocupação da população (84%) quando se trata de saber o que é real ou falso na Internet em matéria de notícias. Para 40% da nossa população, os políticos são os principais responsáveis pela prática de distribuição de informações enganosas. Países onde a polarização política é menor – como Holanda, Alemanha ou Dinamarca – possuem índices bem menores.

Nesta pesquisa, objetivamos analisar as respostas que as plataformas digitais Facebook, Instagram, WhatsApp, YouTube e Twitter têm desenvolvido em relação à desinformação. Pesquisamos, conforme mencionado anteriormente, as medidas desenvolvidas entre os anos de 2018 e 2020, período marcado por eleições que suscitaram debates expressivos sobre desinformação, a exemplo

6. https://portal.fiocruz.br/noticia/pesquisa-revela-dados-sobre-fake-news-relacionadas-covid-19

7. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

de pleitos que ocorreram no Brasil, na Índia e nos Estados Unidos, e pela pandemia do coronavírus, marcada também por campanhas de desinformação e de medidas para tentar contê-las. Sabemos que outras ações foram tomadas posteriormente pelas diferentes corporações, mas acreditamos que o intervalo temporal analisado é bastante representativo do posicionamento delas frente aos problemas, bem como de seus limites. O desafio da atualização desse quadro está posto, mas com este estudo pretendemos oferecer um ponto de partida que facilite a apreensão desses fenômenos. Com isso, o Intervozes busca trazer novos subsídios para organizações da sociedade civil, movimentos sociais, reguladores, desenvolvedores de políticas públicas e para o setor privado melhor compreenderem os desafios que estão colocados para o enfrentamento à desinformação quando se trata das medidas adotadas pelas cinco principais plataformas digitais em operação no país. E espera contribuir para o avanço de uma luta que se mostra, a cada dia, mais estratégica para o futuro da nossa democracia.

CAPÍTULO 1

A ASCENSÃO DAS PLATAFORMAS DIGITAIS E O PROBLEMA DA DESINFORMAÇÃO

Como a introdução desta pesquisa evidencia, aqui adotamos um olhar crítico em relação à configuração atual da internet, marcada, entre outros elementos, pela conformação de monopólios digitais (VALENTE, PITA, 2018) que materializam neste campo em particular tendências históricas como a concentração e a centralização de capital. Diferente de abordagens teóricas de cunho tecnicista que entendem que as características intrínsecas da rede possibilitariam a superação das contradições sociais e conformação de um mundo mais horizontal e democrático, apontamos aqui que a forma de inserção social das tecnologias é alvo de disputas entre os diferentes agentes e respondem, em geral, a interesses dominantes, ainda que pressionados por outros interesses.

Nesse sentido, ao longo deste capítulo apresentaremos a perspectiva teórico-metodológica que orienta esta análise, discutiremos a conformação das plataformas digitais e avançaremos na discussão sobre a desinformação, tema central do estudo, tendo em vista o contexto regulatório atual, as formas de abordagem e medidas propostas por autoridades de direitos humanos e as

propostas do próprio Intervozes – Coletivo Brasil de Comunicação sobre o tema, a fim de detalhar e analisar, tendo em vista esse cenário mais amplo, nos capítulos seguintes, as medidas adotadas pelas próprias plataformas digitais em âmbito mundial.

1.1. PERSPECTIVA TEÓRICO-METODOLÓGICA E CATEGORIAS DE ANÁLISE DA PESQUISA

A inserção da Internet na sociedade sofreu uma mudança substancial em meados dos anos 1990, quando a rede mundial de computadores deixou de ser baseada em uma lógica estatal-militar-acadêmica para assumir uma configuração acadêmico-mercantil, mais diretamente vinculada à sua inserção na dinâmica social contemporânea. Para Bolaño e Vieira (2014, p. 74), "não se trata somente da passagem de uma lógica estatal para outra privada". A economia pública, centrada no investimento estatal, transitou para outra de mercado, de acordo com diferentes modalidades de mercantilização, ao mesmo tempo que também haveria a passagem "de uma lógica política militar, de defesa, para outra, de privatização, regulação e globalização econômica, de apoio à reestruturação capitalista e à manutenção da hegemonia norte-americana nas relações internacionais no campo econômico".

A Internet atraiu investidores de capital de risco, que encontravam oportunidades para grandes investimentos, "capazes de transformar pequenas empresas criadas por estudantes universitários em grandes corporações, alimentando o mito norte-americano do '*self-made man*'" (BOLAÑO, VIEIRA, 2014, p. 74). A onda especulativa resultou na bolha da Internet, no ano de 2000. Consequências disso foram demissões em massa e fusões que levariam à crescente concentração na rede. "As empresas que conseguiram sobreviver à crise são exatamente as gigantes que conhecemos hoje, as quais passaram a utilizar a Internet como plataforma e não

mais como ferramenta", apontam os autores (Ibidem, p. 74). Além dessas, outras empresas emergiram, essencialmente vinculadas ao próprio ambiente on-line, como Google, que também promoveram fusões e aquisições para se constituir em grandes plataformas.

Apesar de sua importância, todo um ideário da Internet como espaço aberto e livre construiu a tese de que ela não deveria ser regulada. É o que orienta a famosa Declaração da Independência do Ciberespaço, de John Perry Barlow (2006), onde se lê que a rede "[...] é intrinsecamente extranacional, intrinsecamente antissoberana e a soberania [dos países] não é aplicável a nós. Temos que descobrir as coisas por conta própria". A declaração é, até hoje, considerada ícone da linha teórica conhecida como excepcionalismo por recusar a legitimidade dos governos sobre a rede e afirmar a impossibilidade de regulá-la. O fato de funcionar à revelia das definições territoriais de um Estado nacional interpôs desafios ao âmbito jurídico, que também se defrontou com características tecnológicas dificilmente assimiladas pela dogmática jurídica tradicional. Além disso, pesou o argumento de que a regulação poderia criar obstáculos para a inovação no setor.

Não obstante, emergiram propostas como a de Lessig (2006), alertando sobre o código como o conjunto de regras que regem a rede e seus dispositivos, o que constituiria a "lei da Internet", e a de Wu (2003), que destaca o princípio da neutralidade de rede, o qual assegura a não discriminação dos pacotes na Internet, obrigando as operadoras a tratarem dados igualmente, o que manteria o caráter aberto da rede. Então, "[...] as ideias excepcionalistas foram confrontadas com a constatação inexorável de que a arquitetura de sistemas a partir da qual a Internet é construída constitui, em si, uma forma de regulação. A forma como seus sistemas são desenhados impõe regras de comportamento aos usuários, sendo a implementação de políticas de uso, informação e acesso inerentes ao desenho das redes e suas configurações" (KELLER, 2019, p. 139). A partir de tal constatação, ganhou força uma visão que

busca adaptar modalidades de regulação tradicional à Internet, emergindo, com isso, teses sobre harmonização e regulação específica da tecnologia.

Não é o caso, aqui, de detalhar tais perspectivas, mas de notar que a discussão sobre a regulação da Internet vem evoluindo junto com seu processo de transformação e expansão e de situar que a presente pesquisa adota posicionamento sobre a necessidade de abordagem multissetorial sobre a rede, tendo em vista seu papel social, os diferentes agentes envolvidos nela, suas características e os princípios que orientam a Internet. Tal abordagem orienta também o documento "Contribuições para uma regulação democrática das grandes plataformas que garanta a liberdade de expressão na Internet"[8], elaborado por organizações da sociedade civil latino-americana, entre elas o Intervozes, com o objetivo de propor parâmetros e diretrizes para a regulação do processo de moderação privada de conteúdos realizada pelas plataformas de Internet.

Partindo da crítica ao cenário polarizado de "autorregulação corporativa versus regulação autoritárias ou excessiva", as organizações defendem um terceiro caminho: "construir uma proposta de regulação democrática, adequada e inteligente, capaz de assegurar ambientes normativos adequados para proteger os direitos humanos das ações dos gigantes tecnológicos, respeitando os padrões internacionais de direitos humanos" (OBSERVACOM et al, 2019, p. 05). Nesse sentido, defendem limites para o poder das corporações, sejam plataformas ou aplicativos que operam serviços que permitem ou facilitam o acesso à informação disponível na Internet e/ou ser suporte para expressão, comunicação e troca de conteúdo entre seus usuários, sendo as regras mais rigorosas com os agentes de maior porte ou abrangência, nos marcos de

8. Disponível em: https://www.observacom.org/wp-content/uploads/2019/08/Contribui-c%cc%a7o%cc%83es-para-uma-regulac%cc%a7a%cc%83o-democra%cc%81tica-das--grandes-plataformas-que-garanta-a-liberdade-de-expressa%cc%83o-na-internet.pdf

uma regulação progressiva, ao passo que também os proteja de pressões ilegítimas de governos e outros atores, tendo em vista o papel das plataformas como intermediárias. Por isso, o documento aponta que a elas não cabe a responsabilidade legal pelo conteúdo de terceiros nem considera adequada a proibição de obrigá-las a acompanhamento genérico ou supervisão de conteúdo.

Em resumo, "tendo em vista as características especiais da internet, busca-se criar um ambiente normativo adequado ao funcionamento e às características da rede, que inclua mecanismos de autorregulação, corregulação e regulação pública, com o entendimento de que os desafios apresentados pelo novo cenário digital (dentre outros, a velocidade e volume de informações) não permitem a aplicação de soluções únicas e iguais do que em outros sistemas de informação e comunicação" (Ibidem, p. 05). Parte fundamental desse processo de regulação pública, o Estado, como em relação aos meios de comunicação hoje chamados de tradicionais, como rádio e TV, de acordo com a perspectiva do direito à comunicação que defendemos, tem o papel de eliminar as restrições econômicas e sociais impostas a diversos grupos para se comunicar por meio de veículos massivos – como é, hoje, a Internet. "Para tanto, é preciso que o Estado adote medidas contra as diferenças que limitam a condição de produtor e difusor de informações a tão poucos grupos e garanta o exercício do direito à comunicação de forma plena e em linha com o direito à informação e à expressão, já que os direitos humanos são complementares e indivisíveis" (INTERVOZES, 2015, p. 11). Um eixo central nessa perspectiva é a adoção de medidas antimonopólicas.

Apenas o Estado, contudo, não é capaz de garantir essas condições, especialmente no atual estágio da globalização capitalista e, parte disso, da operação de corporações transnacionais em todo o mundo, inclusive das empresas baseadas nas tecnologias da informação e da comunicação, responsáveis tanto por viabilizar a internacionalização da produção e das relações de trabalho

quanto por reconfigurar a estrutura de mediação social em geral. No momento presente, em que, como visto anteriormente, as plataformas digitais adquirem maior proeminência na economia mundial e se tornam verdadeiras esferas públicas, ao mesmo tempo em que as práticas desenvolvidas por elas impactam as próprias democracias, como a conformação da arquitetura dos sistemas e a configuração de algoritmos que impactam a circulação de conteúdos, tais corporações são também cobradas para que atuem com responsabilidade.

Outros agentes também participam desse processo, como as organizações da sociedade civil de cada país, bem como organismos internacionais. Desde os anos 1980, tendo em vista a preocupação com o aumento de poder das transnacionais, as Nações Unidas têm elaborado diretrizes sociais e ambientais no âmbito global que tenham como referência a Declaração Universal de Direitos Humanos. Agentes dos sistemas internacionais de direitos humanos caminham no mesmo sentido. Tais formulações, ainda que não necessariamente vinculantes, podem ser tomadas como parte da organização das atividades sociais e também como orientadoras de regras específicas. É nesse sentido que recomendações produzidas por relatores especiais que fazem parte dos Procedimentos Especiais do Conselho de Direitos Humanos da Organização das Nações Unidas (ONU), de órgãos de direitos humanos e do próprio Intervozes serão analisadas nesta pesquisa.

Formulamos a pergunta de partida desta pesquisa nos seguintes termos: quais políticas e práticas têm sido implementadas pelas plataformas Facebook, Instagram, WhatsApp, YouTube e Twitter para conter a disseminação de desinformação? Com esse estudo de caso, "[...] uma inquirição empírica que investiga um fenômeno contemporâneo dentro de um contexto da vida real, quando a fronteira entre o fenômeno e o contexto não é claramente evidente e onde múltiplas fontes de evidência são utilizadas" (YIN, 2010, p. 32), esperamos contribuir com o enfrentamento a um fenômeno

que marca e chega a ser decisivo da organização das relações sociais e de poder hoje.

A escolha daquelas plataformas deve-se ao fato de figurarem entre as mais importantes do ambiente digital, conforme será demonstrado ao longo da pesquisa. Feita essa escolha, passamos então para a fase de coleta e análise de informações. A primeira técnica utilizada foi a revisão de literatura, que viabilizou a definição do marco teórico e a construção conceitual da pesquisa. A segunda, a análise das próprias aplicações, tendo em vista a observação do design tecnológico que apresentam. A terceira, a análise documental de materiais produzidos pelas plataformas ou sobre elas. Utilizamos, então, tanto fontes primárias quanto secundárias. Entre as primárias, temos documentos oficiais (incluindo informações postadas em blogs, redes sociais e sites oficiais) e posicionamentos de seus representantes (considerados textos ou postagens nas redes). Entre as secundárias, informações produzidas por veículos da imprensa especializada e outros estudos.

A busca ativa junto aos vários espaços das plataformas e a utilização de diferentes fontes foram necessárias porque as plataformas não apresentam informações organizadas sobre o tema. Ao contrário, as informações oficiais estão dispersas. A falta de transparência nos leva a evitar afirmar categoricamente que aqui estão registradas todas as medidas sobre desinformação que elas efetivaram no período analisado, pois esses agentes não permitem, por exemplo, que saibamos como seus algoritmos efetivamente são delineados e de que modo isso interfere na circulação de conteúdos, inclusive desinformativos. Essa opacidade consiste em um desafio não só para esta, mas para o conjunto das pesquisas e das intervenções públicas relacionadas às plataformas.

Para a interpretação das informações coletadas, passamos à fase da categorização, que consiste em "[...] uma operação de classificação de elementos constitutivos de um conjunto, por diferenciação e, seguidamente, por reagrupamento segundo o gênero

(analogia), com os critérios previamente definidos" (BARDIN, 1977, p. 117). As categorias reúnem grupos de elementos sob um título genérico, agrupados em razão dos caracteres comuns desses elementos, considerando na conformação delas as seguintes características: exclusão mútua, homogeneidade, pertinência, objetividade e fidelidade, produtividade.

Partindo desses ensinamentos, formulamos quatro categorias, que estão detalhadas abaixo:

1. Abordagem do fenômeno

Busca identificar como as plataformas definem desinformação, se reconhecem a ocorrência do fenômeno da desinformação nos espaços que sustentam, se são estabelecidos critérios e indicadores utilizados para a avaliação do conteúdo desinformativo e quais são as estratégias gerais para lidar com o fenômeno, que serão discutidas de forma mais aprofundada nos itens 2 e 3.

Neste ponto, a análise contempla também se há estabelecimento de processos e/ou estrutura institucional para a integração e coordenação de ações sobre desinformação, seja internamente ou em parceria com outros agentes. A criação de equipes com atribuições voltadas à questão da desinformação, como grupos de verificadores, e a criação de conselhos são exemplos disso.

2. Moderação de conteúdo

Aborda estratégias que incidem sobre a circulação de conteúdos definidos como desinformativos nas plataformas, por meio da moderação exercida sobre contas e publicações em qualquer etapa do processo comunicativo na plataforma (postagem, compartilhamento, encaminhamento, interação entre usuários, etc.).

São ações de moderação de conteúdo, por exemplo, a proibição de criação ou suspensão de conta por identificação como promotora da desinformação; retirada de conteúdos identificados como desinformativos; limitação do alcance por meio de algoritmos; e estabelecimento de limites para compartilhamentos ou encaminhamentos.

3. Promoção de informações e transparência

Comporta medidas direcionadas aos receptores dos conteúdos, a fim de alertar sobre a identificação de um conteúdo como desinformativo, oferecer informações adicionais relacionadas a essas mensagens, fomentar compreensão do funcionamento das plataformas – por meio do fomento a iniciativas como a educação midiática – e/ou disponibilizar informações claras e organizadas sobre suas regras e ações.

Medidas de verificação de conteúdos diretamente ou por meio de parcerias com outros agentes, a exemplo de agências de checagem e disponibilização de informações relacionadas, também são entendidas aqui como de promoção de informações, caso não resultem em moderação direta de conteúdo, sendo assim listadas na categoria anterior.

Também estão incluídas nesta categoria: disponibilização de informações adicionais, inclusive "desmentidos", dentro da plataforma; fomento ao jornalismo com vistas à promoção de informações tidas como confiáveis; ações de formação sobre a mídia, como oficinas em escolas e palestras; realização de campanhas de comunicação para a sociedade; apoio à pesquisa relacionada à desinformação e publicação de relatórios sobre suas ações.

4. Medidas correlatas

Ações das plataformas que, embora não sejam voltadas diretamente à desinformação, têm o potencial de impactar a dinâmica de tal processo. Exemplos são: remoção de contas e conteúdos em função de violações às diretrizes da comunidade; ações para promoção de pluralidade e diversidade de forma geral e identificação de conteúdos patrocinados.

As categorias servirão para a organização das ações das plataformas, bem como para analisar as medidas que elas têm adotado, permitindo ao leitor a visualização das respostas dos agentes ao problema da desinformação, ao longo do segundo capítulo. Antes dele, o texto apresentará o contexto regulatório e formulações de órgãos de direitos humanos sobre o tema, a fim de definir o problema e averiguar como ele vem sendo tratado por diferentes agentes. Também será apresentada posição do Intervozes quanto à abordagem do fenômeno da desinformação. A partir desses referenciais e tendo em vista as categorias criadas para a descrição das ações das plataformas, as medidas adotadas pelas plataformas serão analisadas no terceiro capítulo desta pesquisa. Para tanto, é preciso que também saibamos como funcionam as próprias plataformas que aqui analisamos, por isso trataremos deste tema no próximo tópico.

1.2. CONCEITO E CARACTERÍSTICAS DE FUNCIONAMENTO DAS PLATAFORMAS DIGITAIS

A pandemia do novo coronavírus evidenciou um processo que já evoluía na última metade da década de 2010. Com o isolamento social, boa parte das atividades passou a ser feita à distância. Das reuniões de trabalho às compras, muitos serviços antes realizados presencialmente ganharam versões on-line. Mais do que seu pro-

vimento por meio da Internet, a conexão de ofertantes e contratantes de produtos e serviços ganhou nas plataformas seu espaço privilegiado.

Da refeição entregue pelo Rappi à compra feita pelo Mercado Livre diante de lojas fechadas, das aulas em casa no Google Education às reuniões de trabalho do Microsoft Teams, de lojas de apps viabilizando programas de meditação ou exercícios aos "encontros virtuais" no Zoom, do consumo crescente de vídeos no YouTube ao acesso incessante de informações sobre a pandemia no WhatsApp, Google ou Facebook, as plataformas não apenas consolidaram seu protagonismo no ecossistema digital como se espraiaram para cada vez mais nichos da vida em sociedade.

O que têm em comum serviços tão distintos? Todos podem ser classificados, como apontado anteriormente, como plataformas digitais. O conceito ganhou força nos últimos anos, mas ainda é objeto de intensos debates. Até mesmo os termos empregados variam, como intermediários de Internet (MaCKINNON, HICKOK, BAR & LIM, 2014), intermediários digitais (JIN, 2015), *matchmakers* (EVANS e SCH-MALENSEE, 2016), plataformas tecnológicas (GAWER, 2014), plataformas de agregação e intercomunicações (ALLEN E FLORES, 2013), plataformas on-line (HELBERG, PIERSON & POELL, 2018; ABÍLIO, 2017), plataformas sociodigitais (DANTAS, 2017), plataformas (GILLESPIE, 2010) ou, o que adotamos na presente pesquisa, plataformas digitais (CAMMAERTS E MANSELL, 2020; ANDERSSON SCHWARTZ, 2017; HERSCOVICI, 2019; VALENTE, 2019).

A mesma diversidade de olhares ocorre também nas definições. Gillespie (2017) classifica as plataformas como "sites e serviços que hospedam expressão pública, armazenam-na e utilizam-na na nuvem, organizam acesso a ela por meio de busca e recomendação e a instalam em dispositivos móveis (s/p)". Helberg, Pierson & Poell (2018, p. 01), em um sentido mais abrangente, entendem as plataformas como "arquiteturas sociotécnicas que permitem e

dirigem interações e comunicações entre usuários por meio da coleta, processamento e circulação de dados de usuários".

De acordo com Valente (2019), as plataformas digitais podem ser compreendidas como agentes ou sistemas tecnológicos nos quais ocorrem atividades sobre uma base tecnológica, comandados por proprietários, mas das quais participam outros agentes (produtores, intermediários, usuários), cujos controle e gestão jogam papel organizador chave e nos quais operam lógicas econômicas, práticas culturais e normas diversas (regulatórias e internas), a partir das quais esses sistemas promovem uma mediação ativa na realização dessas interações e transações.

Esses agentes, bastante diversos, são marcados por uma série de traços comuns. O primeiro deles é a capacidade de conectar diversos "lados". Por meio do Facebook ou do Twitter, anunciantes alcançam seus "públicos-alvo". Pelo WhatsApp ou Telegram, famílias, grupos de amigos e colegas de trabalho interagem como mensagens instantâneas. Pelo Google ou DuckduckGo, pessoas buscam por informações e encontram diversas opções de sites na Internet que tratam daquela temática. Pela Play Store ou a Apple Store, usuários encontram apps para as mais diversas finalidades.

Essa conexão é facilitada por uma dinâmica classificada como "efeitos de rede". Por ela, uma rede se tornará mais atrativa quanto maior o número de integrantes. Pessoas criam perfis em redes sociais por seus amigos ou colegas de trabalho estarem nelas. Mas, ao mesmo tempo, o efeito de rede pode ser negativo. Jovens podem deixar uma plataforma por seus pais estarem nela e a utilizarem para vigiar suas vidas.

Outra característica é a intensa coleta de dados. Em um cenário em que essa prática é cada vez mais comum na chamada "economia digital", as informações pessoais assumem a condição de insumo fundamental da oferta de bens ou serviços. No caso das plataformas, as atividades, transações e interações de seus usuários

são registradas e catalogadas, com o intuito de criação de perfis que serão utilizadas para mapeamento de demandas, definição das recomendações e modulação de comportamentos e facilitação da conexão com quem deseja atingir este público, seja para uma venda direta ou para promover um produto por mensagens publicitárias. As plataformas coletam, inclusive, dados para além de sua base de usuários, como ocorre com o Facebook e com o Google ao monitorar internautas por meio de diversos mecanismos, como cookies e outras formas de troca de dados em aplicativos.

Essas atividades se beneficiam da operação on-line desses atores. A disseminação de dispositivos móveis e a difusão do acesso à Internet (em que pese 30% da população brasileira ainda não estar conectada e outra parcela possuir uma conexão precária) criaram o ambiente propício para o avanço das plataformas. Com isso, o aumento do tempo on-line e a mobilidade permitiram que muitos dos serviços se dessem em tempo real, podendo, assim, assumir a condição de substitutos de atividades antes presenciais.

Por fim, essas plataformas estão ancoradas sobre fortes bases tecnológicas. O sucesso do Google passa também pelo sistema de "escaneamento" de um contingente enorme de páginas de Internet e pelo seu algoritmo de ranqueamento. O YouTube conseguiu se tornar a quase única plataforma de massa gratuita de vídeos sob uma pesada infraestrutura de servidores para hospedar os conteúdos publicados, o mesmo válido para Facebook e Twitter em seus colossais fluxos de textos, imagens e sons trocados a cada dia.

As plataformas são agentes econômicos, mesmo que nem sempre tenham finalidade lucrativa. Assim, seus modelos de negócio impactam diretamente suas lógicas de funcionamento. O mesmo para sua arquitetura tecnológica, que é influenciada por seus interesses (de negócios, políticos e sociais) e também molda a atuação dos usuários no seu interior. A tecnologia é vista, assim, como processo, forjada pelas demandas gerais e específicas da sociedade e dos seus controladores e que, da mesma forma, gera consequências

para as práticas sociais, em um fenômeno que chamamos aqui de "regulação tecnológica".

O limite de caracteres para mensagens no Twitter determina os modos de diálogo nessa rede social. A moderação de conteúdo no Facebook define o que é e o que não é visto, bem como os critérios de sucesso na rede. O foco em imagens no Instagram estimula uma projeção de imagens públicas padronizadas, mais do que abre espaço para o acolhimento das diversidades e complexidades da vida contemporânea.

Mais do que apenas o disciplinamento pela tecnologia, as plataformas constituem suas "leis" internas por meio de seus termos de uso e de diversas políticas específicas, como diretrizes das comunidades, regras de privacidade, normas sobre direitos autorais e outros regulamentos semelhantes. O acesso à plataforma e ao seu interior é condicionado à "aceitação" imposta dessas normas, muitas vezes opacas, pouco compreensíveis e em evolução dinâmica com pouco acompanhamento por parte dos usuários.

Por meio de suas "leis" tecnológicas e de práticas, as plataformas se constituem como muito mais do que intermediários, como alguns autores as chamam. Não há neutralidade em um ambiente povoado por limites e parâmetros de comportamento. Quando há a mediação das interações e dos diálogos no interior de tais espaços, ela é fortemente ativa. Ao definirem as regras do jogo, as plataformas podem não determinar diretamente o resultado, mas moldam a forma como os agentes atuam para a consecução dos seus objetivos.

Esse caráter ativo é empregado como forma não apenas de reforço de sua condição de liderança em determinados segmentos como no espraiamento de seus tentáculos para outras áreas. Explorando dados coletados, base tecnológica e amplas bases de usuários, as plataformas conseguem se mover para negócios distintos dos seus originais (a fabricação de carros autônomos no caso do Google, aparelhos de videogames no caso da Microsoft, equi-

pamentos de realidade virtual no caso do Facebook, agregadores de notícias no caso da Apple e séries e filmes, no caso da Amazon).

Em pesquisa anterior, o Intervozes apontou a emergência do fenômeno dos monopólios digitais. "O termo não é utilizado em sentido estrito, designando uma determinada estrutura de mercado, mas com conotação qualitativa, apontando um fenômeno protagonizado por estes agentes. Estes são conglomerados com presença fortemente dominante em uma área mas com atuações para além dela. São digitais pois seus negócios são fortemente centrados nas TICs e neste suporte. Entre as características estão: (1) Forte domínio de um nicho de mercado; (2) Grande número de clientes, sejam eles pagos ou não; (3) Operação em escala global; (4) Espraiamento para outros segmentos para além do nicho original; (5) Atividades intensivas em dados; (6) Controle de um ecossistema de agentes que desenvolvem serviços e bens mediados pelas suas plataformas e atividades; (7) Estratégias de aquisição ou controle acionário de possíveis concorrentes ou agentes do mercado" (VALENTE, PITA, 2018).

Esse conjunto de aspectos e as estratégias adotadas potencializaram a dominância desses agentes no ecossistema on-line. Isso conferiu o que alguns autores chamam de "poder de plataforma". Khan (2018) emprega esse conceito para discutir o crescimento das atividades controladas pelas grandes plataformas (como Google, Facebook, Amazon, Microsoft e Apple), o que apresenta ameaças reais a diversos aspectos, "do sufocamento da inovação de startups e de prejudicar a inovação à manipulação do fluxo de informação e permissão de interferência externa em nossas eleições"[2] (KHAN, 2018).

Para o que nos interessa mais detidamente nesta pesquisa, vale ressaltar que a concentração de mercado, a reunião de milhares de pessoas nas plataformas de redes sociais, a intensa coleta de dados que permite conhecimento, segmentação e tentativas de modulação de comportamentos dos usuários das redes são alguns dos

elementos que têm favorecido o desenvolvimento de estratégias de desinformação, conforme apresentado no livro *Desinformação: crise política e saídas democráticas para as fake news*, do Intervozes (2020). Quanto à ausência de neutralidade, vale mencionar como exemplo o crescimento de canais de extrema direita com base na projeção deles por meio de sistemas de recomendação do YouTube[9]. É por isso que tanto legislações quanto as propostas de órgãos de direitos humanos se voltam também – e no caso destes agentes com centralidade – à abordagem das plataformas.

1.3. PANORAMA DO CONTEXTO REGULATÓRIO SOBRE DESINFORMAÇÃO

Na presente seção, será apresentado um mapeamento sintético da regulação da desinformação em diversos países, como forma de ilustrar as demandas legais e alternativas discutidas no Brasil e em outros países. Considerando que as plataformas em análise são empresas estadunidenses, mas de alcance global, a aprovação de normas legais em países muitas vezes tem impacto sobre o funcionamento destas. As abordagens regulatórias em geral variam conforme o peso das obrigações e da atuação de órgãos estatais.

Como discutido em Valente (2019), a autorregulação é aquela cuja crença é de que os entes regulados por si só garantirão o que se espera do desempenho de suas atividades. Esta pode ser tomada como, também, uma visão regulatória à medida que entes estatais se abstenham da responsabilidade de dar algum tipo de resposta, seja pela ausência de leis específicas ou da fiscalização de um órgão. No tocante à desinformação, as abordagens regulatórias de diversas nações permanecem centradas na autorregulação, seja pela adoção de medidas sem caráter vinculante, seja pela

9. https://theintercept.com/2019/08/28/ranking-youtube-extrema-direita/

existência de propostas no Parlamento ou em outra instância estatal ainda não aprovadas.

É o caso dos Estados Unidos. Desde que o escândalo da Cambridge Analytica revelou a forte presença de notícias falsas nas eleições presidenciais de 2016, iniciativas emergiram, mas sem a concretização na forma de regras. Uma delas foi a investigação aberta pelo Congresso, cujo foco não mirou a desinformação em si, mas a influência russa no pleito concluído com a eleição de Donald Trump incluindo outros recursos, como conteúdos impulsionados. Representantes de grandes plataformas foram ouvidos. O Facebook admitiu a ação de agentes russos no impulsionamento de conteúdos durante as eleições. A investigação foi concluída sem o indiciamento de integrantes da campanha Trump. O Facebook terminou multado por violação de privacidade no escândalo da Cambridge Analytica, e não por eventual contribuição na disseminação de mensagens falsas. No Senado, foi apresentado o "Projeto de lei dos anúncios honestos" (*Honest Ads Bill*) prevendo que plataformas com mais de 50 milhões de usuários no país guardassem informações sobre os anúncios, como responsáveis pelo custeio e público alcançado. Embora trate de publicações pagas, caso aprovado, pode contribuir para coibir a prática de desinformação ao facilitar a identificação das fontes.

No Canadá, o governo federal divulgou em 2019 financiamento público no valor de $ 7 milhões para iniciativas de literacia midiática visando combater a disseminação de conteúdos enganosos e interferência externa no processo eleitoral do mesmo ano (AIELO, 2019)[10]. O projeto envolve a criação de um "Protocolo de Incidentes Eleitorais Críticos"[11], supervisionado por um grupo

10. AIELO, R. Feds unveil plan to tackle fake news, interference in 2019 election. CTV News. Disponível em https://www.ctvnews.ca/politics/feds-unveil-plan-to-tackle-fake-news-interference-in-2019-election-1.4274273

11. https://www.canada.ca/en/democratic-institutions/services/protecting-democracy/critical-election-incident-public-protocol.html.

de cinco altos oficiais do governo, incluindo gestores das áreas de Justiça e Relações Internacionais. Eles terão a prerrogativa de decidir quais casos configuram comportamento on-line preocupante e fazer alertas neste sentido. Em maio do mesmo ano, o governo anunciou uma "carta digital", um plano envolvendo diversos aspectos, entre eles atacar o problema da desinformação e do discurso de ódio.

Outros governos apostaram em campanhas de educação mediática para qualificar o consumo de mensagens para identificar conteúdo enganoso. Na Nigéria, uma campanha foi lançada em 2018 visando alertar os cidadãos sobre os riscos ao país e ao regime político, multiétnico e multirreligioso, deste fenômeno e ressaltar a importância da atuação dos cidadãos para evitar o compartilhamento deste tipo de mensagem. Os esforços envolveram a corporação pública nacional de mídia e acordos com plataformas onde a prática ocorre, como o Facebook.

A Dinamarca lançou em 2018 um plano de ação com diretrizes e medidas para atacar o problema (Danish Ministry of Foreign Afairs, 2018). A iniciativa foi composta de 11 ações, iniciadas com a montagem de uma força-tarefa intergovernamental, que deverá avançar em medidas concretas. O Ministério de Relações Internacionais foi incumbido de desenvolver métodos de monitoramento de desinformação voltadas à Dinamarca, bem como o treinamento de servidores públicos sobre como lidar com esses conteúdos. As agências de inteligência do país foram designadas para verificar eventuais campanhas de influência, assim como identificar vulnerabilidades e ameaças aos processos eleitorais[12].

12. O Executivo prometeu apoio aos partidos disputando vagas no Parlamento contra os riscos de influências externas. E também indicou a disposição de dialogar com meios de comunicação e plataformas digitais com vistas a discutir formas de evitar a influência de desinformação. O plano dinamarquês exemplifica o caráter flexível das estratégias regulatórias, começando dentro de uma lógica de autorregulação, mas podendo evoluir para instrumentos de corregulação ou de regulação de comando e controle.

O exemplo de perspectiva de corregulação mais notório é o da Comissão Europeia. Diante da preocupação com o problema um grupo de experts foi reunido, que elaborou um relatório com um conjunto de recomendações (DE BUNING et al., 2018). Os pesquisadores produziram recomendações estruturais, como a ampliação da transparência acerca da forma como a informação é produzida e difundida, o fomento a um ambiente plural e diverso de mídia no bloco, o fortalecimento do jornalismo de qualidade e da credibilidade das notícias e a promoção de iniciativas de educação midiática. O grupo ressaltou a relevância de intensificar as iniciativas de checagem de fatos e garantir o conhecimento destas pelos utilizadores.

Para além das bases gerais do ambiente informacional on-line, cuja constituição demanda mudanças de longo prazo, o núcleo modelo foi estruturado sob uma lógica corregulatória ancorada em um código de práticas a ser adotado pelas plataformas digitais, com a formulação de planos de ação acerca de como tais diretrizes seriam implementadas. As companhias assumiriam o compromisso de respeitar o código e fornecer informes periódicos acerca da sua implementação[13]. O documento traz uma série de medidas às plataformas, como: 1) Ajustar as políticas de publicidade, identificando incentivos monetários à difusão de desinformação e avaliando formas de diminuir esses mecanismo; 2) Distinguir claramente o conteúdo, patrocinado e não pago, publicado pelos usuários; 3) Assegurar transparência e fiscalização pública sobre seu funcionamento e a forma de gestão dos conteúdo dos usuários; 4) Permitir o acesso a dados sobre sua operação e a mensagens difundidas para agências de checagem de fatos e pesquisadores; 5) Disponibilizar aos usuários configurações para permitir que estes definam a forma de organização do fluxo de informações

13. Os relatórios das plataformas e parciais por mês estão disponíveis no site da Comissão Europeia: https://ec.europa.eu/digital-single-market/en/tackling-online-disinformation. Acesso: 15 de set. 2019.

em suas linhas do tempo; 6) Melhorar a visibilidade de notícias precisas e confiáveis e facilitar o acesso dos usuários a essas reportagens; 7) Publicar vinculado a notícias de grande disseminação sugestões de artigos relacionados com outras visões sobre um determinado fato; 8) Ofertar ferramentas para que usuários criem links que permita saber se há contestação de checadores de fatos sobre o que recebe no seu feed; 9) Criar mecanismos para evitar abusos por usuários nas ferramentas de identificação de notícias denunciadas como falsas; 10) Disponibilizar dados relevantes sobre seus serviços para pesquisadores independentes.

Em um conjunto de países, autoridades estatais adotaram medidas diretas de comando e controle contra a prática de desinformação. Parte dos países tipificou (ou já havia incluído em sua legislação) a proibição da criação e difusão de conteúdos falsos, com correspondente responsabilização. A Malásia foi pioneira na aprovação de lei em abril de 2018 criminalizando "notícias, informações ou dados e reportagens no todo ou em parte falsos" (HAFIDZAH, 2019). A Lei estabeleceu pena de até seis anos de prisão. O ato gerou polêmicas intensas no país, com parlamentares defendendo a revogação da matéria. O novo governo prometeu aprovar uma nova Lei revogando o efeito da anterior, o que foi repelido pelo Senado.

Em Cingapura, foi aprovada uma lei em maio de 2019 criminalizando a disseminação de "falsos enunciados de fato". A lei permite a qualquer ministro determinar correções a mensagens definidas como falsas (RUSSEL, 2019). A norma prevê multa e até cinco anos de prisão, podendo ser dobrada em caso de publicação por conta não autêntica ou bot. O texto também abre possibilidade de julgamento de casos envolvendo a relação de Cingapura com outros países.

Na Europa, as iniciativas legislativas partiram de duas das principais nações do continente: França e Alemanha. Na primeira, a Lei de Imprensa já estipulava multas a quem reproduzisse e

propagasse notícias falsas. Mas em novembro de 2018 foi aprovado regulamento específico, voltado a coibir o comportamento nas eleições, proibindo "fake news" definidas como "alegações ou imputações inexatas ou notícias que falsamente reportam fatos com o intuito de mudar a sinceridade de um voto". A abrangência da norma restringe-se aos períodos eleitorais. Juízes ganharam a prerrogativa de agir para responder a conteúdos falsos, determinando sua retirada, por exemplo.

Na Alemanha, a Lei conhecida como NetzDG estabeleceu a obrigação das plataformas excluírem "conteúdos ilegais" previamente previstos no arcabouço legal do país. Assim, a NetzDG não tipificou a prática de desinformação, embora elenque entre as práticas ilegais um tipo diretamente próximo ("falsificação") e outras que podem aparecer relacionadas ("insulto e difamação", "falsificação de dados com o intuito de fornecer prova", "difamação de religiões ou associações ideológicas"). Para além de determinar a fiscalização e a remoção às plataformas, a lei prevê mecanismos de transparência sobre essa fiscalização e as denúncias recebidas.

Em outros países, tipos penais não específicos sobre informações falsas também passaram a ser usados como forma de combater a prática. Na Arábia Saudita, o regime autocrático passou a ameaçar a punição de quem disseminasse material falso a partir da legislação de crimes cibernéticos, segundo a qual é crime quebrar a ordem pública, valores religiosos, a moral pública e a privacidade (NOVAK, 2018). A pena prevista chega até cinco anos. A prática, assim, amplificou o exercício de poder pelo regime, marcado por um domínio calcado em base religiosa. A atitude ocorreu em meio a suspeitas de que o regime saudita estaria envolvido com o assassinato do colunista do Washington Post Jamal Khashoggi.

No Brasil, como mencionado no início deste estudo, o Código Eleitoral já previa a punição de quem publicasse informação "sabidamente inverídica" durante as eleições. Em 2019, o Congresso Nacional aprovou atualização da legislação eleitoral inserindo

novo trecho, criando o crime de "denunciação caluniosa com finalidade eleitoral" estipulando pena de até oito anos para quem imputa a alguém fato "falsamente atribuído" sabendo de sua inocência. O presidente da República, Jair Bolsonaro, havia vetado o dispositivo, mas o Congresso derrubou a decisão e restaurou a previsão.

Com o advento da pandemia do novo coronavírus e a profusão de mensagens falsas sobre o tema, diversos países passaram a adotar medidas para combater a desinformação[14]. Na Argentina, o governo implementou o que chamou de "patrulhas cibernéticas", monitorando redes sociais para verificar a disseminação de mensagens falsas. Essa iniciativa resultou em processos contra usuários utilizando um outro tipo penal, de "intimidação pública", com pena de até seis anos de prisão. A Bósnia e Herzegovina aprovou decreto que proíbe a difusão de mensagens falsas para causar "pânico e desordem" durante o estado de emergência, com sanção de multa. Em Botsuana, a legislação voltada à pandemia proíbe a difusão sobre Covid-19 diferente daquelas da autoridade nacional de saúde e da Organização Mundial de Saúde, cuja violação pode ensejar multa ou prisão de até cinco anos. A Etiópia aprovou lei sobre o estado de emergência que, entre outras ações, veda a distribuição de informação que possa causar "alarme ou confusão". Na Hungria, a lei que suspendeu as eleições e permitiu o governo do presidente por decreto também previu a punição de quem "distorça" ou divulgue notícias falsas. Nas Filipinas, a lei de ampliação dos poderes do governo também criminalizou a difusão de informação falsa sobre a Covid-19, prática sujeita a multa e a prisão de até dois meses.

14. Um levantamento completo é feito pelo projeto de mapeamento da organização ICNL, disponível em: https://www.icnl.org/covid19tracker/?location=&issue=4&date=&type=.

1.4. ABORDAGENS SOB A PERSPECTIVA DA DEFESA DOS DIREITOS HUMANOS

A relevância adquirida pelo problema da desinformação, sua influência em processos eleitorais e a adoção de medidas variadas por parte de agentes públicos e privados, com impactos para a liberdade de expressão, têm suscitado questionamentos e formulações sobre o tema por parte de relatores especiais que fazem parte dos Procedimentos Especiais do Conselho de Direitos Humanos da Organização das Nações Unidas (ONU) e órgãos de direitos humanos regionais e nacionais, como na União Europeia e no Brasil, respectivamente.

Em março de 2017, relatores especiais designados por organismos internacionais (ONU, Organização dos Estados Americanos - OEA, Organização para Segurança e Cooperação na Europa - OSCE e Comissão Africana dos Direitos Humanos e dos Povos - CADHP) formularam conjuntamente a *Declaração conjunta sobre a Liberdade de Expressão e Notícias Falsas, Desinformação e Propaganda*, primeiro documento desses agentes que tratou explicitamente do fenômeno e sugeriu recomendações a Estados, empresas de tecnologia e meios de comunicação, com vistas, sobretudo, à garantia da liberdade de expressão.

Em junho de 2018, durante Assembleia Geral das Nações Unidas, foi apresentado o *Informe do Relator Especial sobre a promoção e proteção do direito à liberdade de opinião e de expressão*. Assinado por David Kaye, o informe aborda a questão da regulação do conteúdo gerado por usuários da rede mundial de computadores e, como parte disso, trata também da questão da desinformação, ao abordar especialmente preocupações no âmbito de normas sobre conteúdos.

Ainda no plano internacional, a preocupação com as eleições motivou a apresentação, em abril de 2020, da *Declaração Conjunta sobre Liberdade de Expressão e Eleições na Era Digital*, assinada

por relatores especiais da ONU, da OSCE e da OEA. Nela, os especialistas dizem estar "alarmados" com o uso indevido de redes sociais por parte de atores estatais e privados para "subverter" processos eleitorais e, por isso, recomendam dezenas de ações para diferentes agentes.

Na União Europeia, ainda em março de 2015, o Conselho Europeu convidou a Alta Representante a elaborar um plano de ações para enfrentar campanhas de desinformação que seriam organizadas pela Rússia, que estava envolvida em intensa disputa com a Ucrânia. Em junho de 2017, o Parlamento Europeu solicitou à Comissão Europeia que analisasse a situação em profundidade e que verificasse a possibilidade de intervenção legislativa para limitar a divulgação e difusão de conteúdos falsos. A comissão criou, então, um grupo de peritos para a aconselhar sobre a matéria.

O grupo de especialistas apresentou o documento *Uma abordagem multidimensional à desinformação – Relatório do Grupo de Alto Nível independente sobre notícias falsas e desinformação on-line*, em março de 2018. A partir dele e de contribuições da população obtidas por meio de consulta pública com quase três mil respostas e de um inquérito Eurobarómetro que abrangeu os 28 Estados-Membros, foi apresentado, em abril do mesmo ano, o documento *Combater a desinformação on-line: uma estratégia europeia*. O texto foi remetido ao Parlamento Europeu, ao Conselho, ao Comitê Econômico e Social Europeu e ao Comitê das Regiões.

No Brasil, em junho do mesmo ano o Conselho Nacional dos Direitos Humanos (CNDH) publicou a *Recomendação Nº 4 de 2018*, com medidas "de combate às *fake news* (notícias falsas) e garantia do direito à liberdade de expressão". A resolução dialogava com o cenário de debate de medidas para o pleito de 2018, que estava em curso especialmente no âmbito do Tribunal Superior Eleitoral (TSE) e do Congresso Nacional, onde tramitava mais de uma dezena de projetos de lei sobre o tema.

Ainda que as sugestões desses documentos não sejam vinculantes, já que não há ainda convenções ou tratados firmados especificamente sobre o tema, tratam-se de colaborações importantes para a compreensão do problema e das formas de enfrentá-lo. Ademais, são referências que podem orientar políticas públicas e boas práticas dos agentes envolvidos. É nesse sentido que esta seção aborda também recomendações e declarações contextualizadas acima, quais sejam[15]:

1. *Declaração conjunta sobre a Liberdade de Expressão e Notícias Falsas, Desinformação e Propaganda* (ONU, OEA, OSCE e CADHP)

2. *Uma abordagem multidimensional à desinformação – Relatório do Grupo de Alto Nível independente sobre notícias falsas e desinformação on-line* (Grupo de Alto Nível independente sobre notícias falsas e desinformação on-line - HLEG, na sigla em inglês)

3. *Combater a desinformação em linha: uma estratégia europeia* (Comissão Europeia)

4. *Informe do Relator Especial sobre a promoção e proteção do direito à liberdade de opinião e de expressão* (ONU)

5. *Recomendação N° 4 de 2018* (CNDH)

6. *Declaração Conjunta sobre Liberdade de Expressão e Eleições na Era Digital* (ONU, OSCE e OEA)

Para uma observação sistemática dos documentos, foram consideradas as seguinte questões, que buscam apreender tanto a abordagem geral conferida à desinformação quanto posicionamentos diante de ações específicas das plataformas: i) Conceito utilizado e enquadramento do problema; ii) Abordagem regulatória e/ou agente(s) ao(s) qual(is) as propostas são dirigidas; iii) Medidas recomendadas e iv) Avaliação da possível moderação de conteúdos por parte das plataformas.

15. Ao longo do texto, serão utilizados os números dessa lista para fazer referência aos documentos que indicam, de modo a evitar a repetição exaustiva dos títulos completos.

1.4.1 CONCEITO UTILIZADO E ENQUADRAMENTO DO PROBLEMA

Uma questão fundamental a ser enfrentada no combate à desinformação é a definição do fenômeno que, conforme já explicitado nesta pesquisa, não é consensual. Dos documentos analisados, sobressai a utilização do termo desinformação, ainda que seja mencionada também a expressão "fake news".

O enquadramento prioritário como desinformação já aparece na declaração conjunta de 2017, na qual se afirma que "as proibições gerais à divulgação de informações com base em conceitos imprecisos e ambíguos, incluídos 'notícias falsas' ('fake news') ou 'informação não objetiva', são incompatíveis com os padrões internacionais sobre restrições à liberdade de expressão [...] e deveriam ser revogadas". No mesmo sentido, a recomendação do CNDH, ainda que não apresente um conceito, justifica a opção pelo termo desinformação "por ser mais apropriado para dar conta da complexidade do tema e de suas múltiplas expressões". Cita ainda que "categorizar todas as formas de desinformação sob a expressão fake news ou 'notícias falsas', sem diferenciar as suas especificidades, dificulta o processo de elaboração de estratégias para combater esse fenômeno" (CNDH, p. 01).

Apenas os documentos da União Europeia (2 e 3 da lista acima) detalham o conceito de forma precisa. No relatório dos especialistas, consta que a desinformação "inclui todas as formas de falsidade, informações inexatas ou enganosas concebidas, apresentadas e promovidas para causar intencionalmente danos públicos ou para fins lucrativos. Não abrange questões decorrentes da criação e divulgação on-line de conteúdos ilegais (nomeadamente difamação, discurso de ódio, incitamento à violência), que estão sujeitas a soluções regulamentares ao abrigo da legislação comunitária ou nacional. Também não abrange outras formas de distorção deliberada mas não enganosa de factos como a sátira e a paródia" (HLEG, 2018, p. 05).

Em outro ponto, detalha que "o risco de danos inclui ameaças aos processos e valores políticos democráticos, que podem visar especificamente uma variedade de setores, tais como saúde, ciência, educação, finanças e muito mais. É impulsionada pela produção e promoção da desinformação para fins econômicos ou para objetivos políticos ou ideológicos, mas pode ser exacerbada pela forma como diferentes públicos e comunidades recebem, ativam e amplificam a desinformação" (Ibidem, p. 10).

O termo notícias falsas é considerado inadequado para captar a complexidade do problema da desinformação, que envolve um conteúdo que não é real ou completamente "falso", mas informações fabricadas misturadas com fatos, bem como práticas que vão muito além de tudo o que se assemelhe a "notícias", incluindo algumas formas de contas automáticas, redes de seguidores falsos, vídeos fabricados ou manipulados, publicidade direcionada etc. Os especialistas mencionam, ainda, que o termo "notícias falsas" foi apropriado e usado enganosamente por atores poderosos para descartar a cobertura de mídia desagradável a eles.

Na esteira de tal elaboração, no documento da Comissão Europeia tem-se que "a desinformação é entendida como informação comprovadamente falsa ou enganadora que é criada, apresentada e divulgada para obter vantagens econômicas ou para enganar deliberadamente o público, e que é suscetível de causar um prejuízo público. O prejuízo público abrange ameaças aos processos políticos democráticos e aos processos de elaboração de políticas, bem como a bens públicos, tais como a proteção da saúde dos cidadãos da UE [União Europeia], o ambiente ou a segurança. A desinformação não abrange erros na comunicação de informações, sátiras, paródias ou notícias e comentários claramente identificados como partidários" (CE, p. 04).

Em todos os documentos o problema é enquadrado de forma ampla, tendo em vista sua relação com o ecossistema midiático e com a própria educação das e dos cidadãos, ainda que busquem

especificar suas características particulares para oferecer remédios direcionados. Nesse sentido, há uma frequente associação da questão à temática da liberdade de expressão de forma geral, mencionada em todos os textos analisados. Merecem destaque as formulações da Comissão Europeia e do CNDH, as quais, além de enquadrar a questão a partir da ótica da liberdade de expressão, mencionam também a proteção da privacidade e dos dados pessoais.

1.4.2 ABORDAGEM REGULATÓRIA E/OU AGENTE(S) AO(S) QUAL(IS) AS PROPOSTAS SÃO DIRIGIDAS

Nem todos os documentos citam expressamente uma abordagem regulatória ou mesmo a necessidade de uma regulação[16]. Para perceber o tipo de regulação defendido, consideramos também para quais agentes as recomendações são dirigidas. Se voltadas apenas para as plataformas digitais ou outros agentes privados, como agências de checagem, a abordagem é enquadrada nos marcos da autorregulação. Caso sejam recomendadas ações multissetoriais ou mencionada cooperação entre agentes públicos e privados, aproxima-se da corregulação. Caso a ação seja recomendada apenas para o Estado, será entendida aqui como regulação estatal.

A primeira declaração conjunta (1) dirige possíveis respostas a Estados, intermediários, jornalistas e meios de comunicação, além de fazer menção à sociedade civil. De acordo com o texto, "todos os atores interessados – incluindo intermediários, meios de comunicação, a sociedade civil e acadêmicos – deveriam receber apoio para formular iniciativas participativas e transparentes que favoreçam uma maior compreensão do impacto da desinformação

16. Por regulação entendemos não apenas a fixação de regras por meio do Estado, mas variadas formas de coordenação de atividades sociais, que se dá inclusive quando o Estado se abstém de tomar iniciativas, deixando-as sob responsabilidade de agentes privados, por isso são consideradas propostas de regulação estatal, corregulação e autorregulação.

e da propaganda na democracia, na liberdade de expressão, no jornalismo e no espaço cívico, bem como respostas adequadas a esses fenômenos" (ONU; OEA; OSCE; CADHP, p. 05), motivo pelo qual é possível inferir que se trata de uma abordagem pautada na corregulação.

No mesmo sentido, no informe à ONU (4), há seções específicas intituladas "Obrigações dos Estados" e "Responsabilidades das empresas", o que já indica a partilha de responsabilidade entre ambos os agentes, no sentido da corregulação para a promoção e proteção do direito à liberdade de opinião e de expressão. O documento brasileiro (5) dirige recomendações ao Congresso Nacional, ao Tribunal Superior Eleitoral (TSE), portanto ao Judiciário, aos Executivos Federal e Estadual e às plataformas privadas (nomeadamente Facebook, Twitter e Google). Na declaração conjunta sobre eleições (6), as recomendações são dirigidas aos Estados e aos "atores não estatais", entre os quais meios de comunicação tradicionais como emissoras de rádio e TV, os intermediários da Internet e especificamente as plataformas digitais. Propõe ainda ações a "outras partes interessadas", destacadamente partidos, políticos e candidatos.

Já o documento dos especialistas da União Europeia diz que: "de uma perspectiva política geral, o HLEG sugere, como primeiro passo, uma abordagem de autorregulação baseada num processo de envolvimento claramente definido das várias partes interessadas [multi-stakeholder], enquadrado num roteiro vinculativo para a implementação e centrado num conjunto de ações a curto e médio prazo. Esse processo deverá decorrer em paralelo com as intervenções propostas para reforçar a educação para a mídia e a diversidade e sustentabilidade do ecossistema da informação digital, ações que, pela sua própria natureza, têm um horizonte temporal mais longo"[17] (HLEG, p. 35), as quais são dirigidas especialmente

17. Tradução própria do original em inglês: "From a general policy perspective, the HLEG suggests, as a first step, a self-regulatory approach based on a clearly defined multi-stakeholder engagement process, framed within a binding roadmap for implemen-

às autoridades públicas.

No caso, é possível aferir uma preferência pela autorregulação, ainda que sejam feitos apelos para a cooperação entre os agentes e para o engajamento dos Estados e das instituições da União Europeia. A justificativa para isso é que o grupo entende que respostas regulamentares podem se tornar rapidamente inadequadas para resolver um problema considerado multifacetado e cujas características podem sofrer alterações. Por isso, como veremos adiante, tem destaque a ideia de um código de condutas firmado pelos agentes envolvidos (plataformas, organizações de mídia, jornalistas, verificadores de fatos, criadores de conteúdo independentes e a indústria da publicidade).

Para além dessa questão, é perceptível influência da perspectiva da não intervenção ou de uma atividade negativa por parte da autoridade estatal, isto é, de defesa do afastamento do Estado das relações individuais e sociais. Ao tratar do Estado, a Comissão Europeia (3) considera que "a principal obrigação dos intervenientes estatais no que se refere à liberdade de expressão e à liberdade dos meios de comunicação social é abster-se de interferir e censurar e garantir um ambiente favorável a um debate público inclusivo e pluralista" (CE, p. 01) e que "a autorregulação pode contribuir para estes esforços [de combater a desinformação], desde que seja implementada e acompanhada de modo eficaz" (Ibidem, p. 09).

Partindo de tal avaliação, formula propostas voltadas especialmente às plataformas digitais, especialmente no que tange à implementação de medidas de proteção para seus utilizadores. A indústria de publicidade também é referida. Por considerar que "soluções eficazes a longo prazo exigem a sensibilização do público, um índice mais elevado de literacia mediática, um amplo

tation, and focused on a set of short and medium-term actions. This should go in parallel to the proposed interventions to strengthen edia and information literacy and the diversity and sustainability of the digital information ecosystem, actions that by their own very nature have a longer time horizon".

envolvimento das partes interessadas e a cooperação entre autoridades públicas, plataformas digitais, anunciantes, sinalizadores de confiança, jornalistas e grupos de comunicação social" (CE, p. 07), dirige-se também a esses possíveis agentes.

Tendo em vista a busca por soluções compartilhadas e que sejam capazes de acompanhar as transformações tecnológicas e na própria configuração da desinformação, uma questão associada à regulação estatal é frequentemente mencionada nos textos: a fixação de leis sobre o tema. Em geral, a proposta é vista com parcimônia ou mesmo desaconselhada nos documentos analisados.

Em 2017, os relatores (1) escreveram que "as normas de direito civil relativas ao estabelecimento de responsabilidades ulteriores por declarações falsas e difamatórias unicamente serão legítimas se se concede aos demandados uma oportunidade plena de demonstrar a verdade dessas declarações e estes não realizam tal demonstração e si, além disso, os demandados também puderem afirmar outras defesas, como comentário razoável ('*fair comment*')". Adicionam que "se deverá considerar a necessidade de proteger as pessoas da imposição de responsabilidade legal pelo simples fato de haver redistribuído ou promovido, através de intermediários, conteúdos que não sejam de sua autoria e que elas não tenham modificado" (ONU; OEA; OSCE; CADHP, p. 03).

Na União Europeia, o grupo de altos estudos postulou que as melhores respostas são provavelmente aquelas impulsionadas por colaborações das partes envolvidas, minimizando as intervenções regulamentares legais, pois avalia que "a regulamentação da desinformação do governo ou da UE pode ser um instrumento contundente e arriscado"[18] (HLEG, p. 19). No caso de respostas regulamentares, cita que possíveis leis devem garantir o devido processo e ter em vista princípios como responsabilidade e proporcionalidade. A Comissão Europeia (3) não trata explicitamente

18. Tradução própria do original em inglês: "regulation of disinformation can be a blunt and risky instrument".

do estabelecimento de leis, nem mesmo quando aborda a responsabilização on-line.

O informe apresentado à ONU (4) parte da compreensão de que possíveis restrições à liberdade de expressão devem estar fixadas por leis que sigam condições de legalidade, necessidade e legitimidade. Mas a possibilidade não significa estímulo, pois reconhece que riscos podem ser derivados da intervenção dos Estados e também da operação privada. "Os Estados exigem periodicamente que as empresas que restrinjam o conteúdo manifestamente ilegal, como a representação de abusos sexuais de crianças, as ameaças diretas e verossímeis de causar um dano e a incitação à violência, tomando como certo que com isso satisfazem as condições de legalidade e necessidade. Alguns Estados vão muito mais além e recorrem à censura e à aplicação da lei para configurar o entorno regulamentar da atividade on-line. Leis restritivas com uma redação de caráter muito geral sobre o 'extremismo', a blasfêmia, a difamação, o discurso 'ofensivo', as 'notícias falsas' e a 'propaganda' costumam servir de pretexto para exigir que as empresas removam expressões legítimas. Outras leis podem interferir na privacidade on-line de maneira a desencorajar os usuários de exercer o direito à liberdade de opinião e expressão" (ONU, p. 06-07), alerta.

No caso brasileiro, o documento do CNDH (5) aponta que leis devem seguir os padrões de direitos humanos e a proteção de dados pessoais. Mais uma vez, possibilidade não significa necessidade. Na opinião do órgão colegiado, o combate a informações falsas, caluniosas e difamadoras já encontra um arcabouço regulatório e institucional adequado na tipificação dos crimes contra a honra, em recursos como o direito de resposta e em mecanismos de retirada de conteúdos, tais como os previstos no Marco Civil da Internet. Entende ainda que criminalizar o compartilhamento de conteúdos pelo público geral configura medida desproporcional e que a condenação à prisão "não pode ser vista como uma medida eficaz para enfrentar esse problema, uma vez que, na maior parte

das vezes, o cidadão comum sequer tem informações ou instrumentos para verificar a veracidade de um conteúdo que circula pela internet" (CNDH, p. 02).

O documento sobre eleições (6) na era digital também aborda a resposta legislativa com parcimônia. "Os Estados deveriam avaliar a possibilidade de respaldar medidas positivas para abordar o problema da desinformação em linha, como promoção de instâncias independentes para verificação de informações e campanhas de educação pública, e ao mesmo tempo, evitar adotar normas para criminalizar a desinformação" (ONU; OEA; OSCE; CADHP, p. 03). O texto recomenda que os Estados "não devem estabelecer leis gerais nem ambíguas sobre desinformação, como proibições a respeito da difusão de 'falsidade' ou 'informação não objetiva'. Por outro lado, avalia que deveriam adotar leis claras e proporcionais "que proíbam a difusão de declarações que persigam o propósito específico de obstaculizar o direito das pessoas a votar; por exemplo, difundindo deliberadamente informação incorreta sobre onde e quando votar" (ONU; OSCE; OEA, p. 03).

O remédio legislativo, portanto, ainda que possível para enfrentar violações, não é estimulado pelos órgãos de direitos humanos. Quando tratam dessa possibilidade, recomendam também proporcionalidade e mecanismos como devido processo, pois avaliam que podem estar associados riscos quanto à garantia da liberdade de expressão e à imposição de censura, seja privada ou estatal. Interessante notar que apenas o texto da Comissão Europeia não trata explicitamente dessa questão, o que denota sua relevância.

Ainda quanto ao papel do Estado, há a expectativa de que atue para estabelecer condições para o exercício da liberdade de expressão e para garantir pluralidade e diversidade nos meios de comunicação, bem como condições, sobretudo vinculadas à educação, para que a sociedade possa receber e interpretar os conteúdos. Por fim, chama atenção o fato de praticamente não haver menção ao Judiciário e a forças policiais, o que indica que as abordagens

baseadas na perspectiva dos direitos humanos não caminham no sentido da persecução penal.

1.4.3 MEDIDAS RECOMENDADAS

Dada a abordagem multidimensional conferida ao problema da desinformação, são inúmeras as propostas apresentadas nos textos em análise, por isso optamos por apresentar neste tópico os eixos orientadores das medidas recomendadas para combater a desinformação, com menções a ações específicas que exemplificam as abordagens propostas.

Os documentos propõem medidas que possam materializar o equilíbrio entre a garantia do direito à liberdade de expressão e o estabelecimento de restrições a ele quando necessário para combater a desinformação, por isso os textos 1, 4 e 6 fazem menções aos princípios da necessidade e da proporcionalidade. O primeiro significa que os atos devem ser menos restritivos aos direitos individuais. O segundo, que deve haver uma proporção adequada entre os meios utilizados e os fins desejados.

Em relação ao Estado, na declaração conjunta (1), são desaconselhadas medidas que podem impor restrições à liberdade de expressão em geral, tanto pelo Estado quanto pelas plataformas (bloqueios de sites inteiros, responsabilidade de agentes privados pelos conteúdos, imposição de sistemas de filtro), ainda que existam exceções com vistas ao equilíbrio referido acima (como proibição da apologia ao ódio). Também é dito o que os Estados não são aconselhados a fazer (como fixar leis baseadas em conceitos imprecisos, fomentar ou difundir desinformação), além de mencionar que "os atores estatais deveriam procurar difundir informação confiável e fidedigna, incluídos sobre temas de interesse público, como economia, saúde pública, segurança e de meio ambiente" (p. 04). Há, na sequência, diversas propostas reunidas sob o sub-

título "entorno propício à liberdade de expressão", onde são referidas medidas proativas do Estado em relação aos sistemas de comunicações em geral (assegurar diversidade dos meios, garantir sustentabilidade de meios públicos e independentes, promover alfabetização midiática, fomentar igualdade e não discriminação).

Quanto aos intermediários, as propostas são todas no sentido da transparência, de garantia do devido processo e do empoderamento digital dos usuários. Entre elas, estão: adotar políticas claras e preestabelecidas que regulem possíveis medidas de restrição de conteúdos de terceiros, as quais não deveriam exceder o legalmente necessário; assegurar que seus usuários possam consultar facilmente e compreender as políticas e práticas, até mesmo automatizadas, e incluí-las nas condições de serviço; apoiar a investigação e o desenvolvimento de soluções tecnológicas; cooperar com iniciativas que ofereçam serviços de verificação de dados aos usuários e revisar seus modelos de publicidade para garantir que não tenham um impacto adverso na diversidade de opiniões e ideias.

Aos meios de comunicação e jornalistas, os especialistas e órgãos de direitos humanos também sugerem ações propositivas para ampliação do conhecimento crítico sobre desinformação (oferecer uma cobertura crítica da desinformação e a propaganda como parte de seus serviços de notícias, sobretudo em períodos eleitorais, e apoiar sistemas efetivos de autorregulação que incluam padrões para propiciar a veracidade das notícias, entre outras coisas, contemplando o direito de retificação e/ou réplica em caso de fatos incorretos nos meios).

O documento (2) do grupo de peritos é o mais detalhado em termos de proposições, que se dividem tanto pela periodicidade (curto e médio prazos) quanto por eixos, quais sejam: transparência (garantir transparência das fontes de financiamento, aumentar a transparência das fontes de notícias on-line e dos processos jornalísticos, ampliar transparência e eficiência dos processos de checagem); literacia midiática e de informação (ações que

promovam uma reavaliação e ajuste das políticas educacionais, apoio à mídia e programas de alfabetização para cidadãos de todas as idades, que no Brasil são tratadas como "edução para a mídia"); empoderamento de usuários e jornalistas (desenvolver e disponibilizar ferramentas que agreguem sinais de qualidade, como rótulos de identidade, indicadores de transparência de origem, rótulos de conteúdos verificados e, ao mesmo tempo, aumentem o controle dos usuários sobre o conteúdo que eles pesquisam ou são recomendados, facilitando a exibição da maior variedade possível de notícias relevantes e informações aos consumidores); diversidade e sustentabilidade do ecossistema dos meios de comunicação social (para que o público seja crítico, engajado e bem informado, sugere estratégia de "diluir" a desinformação com maior transparência e visibilidade de notícias confiáveis, mas tal ação deveria ser combinada com outras projetadas para preservar a diversidade e a sustentabilidade econômica de longo prazo da mídia); processos e avaliação (garantia de processo estruturado de implementação e avaliação das ações, estabelecimento de código de práticas a serem adotadas em toda Europa, participação de várias partes interessadas na elaboração das medidas).

Dada a relevância do código proposto para estabelecer regras concretas de conduta em função do papel que as plataformas, a mídia de notícias e as organizações de verificação de fatos devem desempenhar para proteger um ambiente propício à liberdade de expressão e, ao mesmo tempo, promover a transparência e inteligibilidade de diferentes tipos de canais de informação digital, praticamente todas as propostas foram acolhidas pela Comissão Europeia e constam no documento aqui em análise, conforme detalhado abaixo.

A Comissão Europeia estabelece como princípios: transparência, diversidade, credibilidade e adoção de soluções inclusivas. Conforme explicitado no ponto anterior, o documento parte de uma visão negativa da intervenção estatal, por isso aos Estados

são propostas sobretudo medidas de média e longo prazos, com foco na educação para ampliação da resiliência das e dos cidadãos. No mesmo sentido, a comissão incentiva os Estados-membro a disponibilizarem recursos para esse tipo de política e, devido à dimensão transfronteiriça da desinformação, avalia que a UE tem um papel a desempenhar no apoio à divulgação das boas práticas entre os Estados-Membros. Quanto a estes, no caso das eleições para o Parlamento Europeu de 2019, diz que incentivou as autoridades nacionais competentes a identificar as melhores práticas em matéria de identificação, redução e gestão dos riscos que os ataques informáticos e as campanhas de desinformação comportam para o processo eleitoral. À própria comissão, que tem caráter de autoridade pública, são definidas tarefas de desenvolvimento tecnológico para a investigação de comportamentos maliciosos.

O texto menciona ainda que a comissão convocará um fórum multilateral sobre a desinformação, "com vista a proporcionar um quadro para uma cooperação eficaz entre as partes interessadas, incluindo as plataformas digitais, a indústria de publicidade e os principais anunciantes, os meios de comunicação social e representantes da sociedade civil, e obter compromissos a favor da coordenação e intensificação dos esforços de luta contra a desinformação".

Seguindo os especialistas, o documento foca nas ações destinadas aos agentes privados. Diz o texto: "é necessário promover mudanças adequadas no comportamento das plataformas, uma maior responsabilização no ecossistema de informações, o reforço das capacidades de verificação de fatos e do conhecimento coletivo sobre desinformação, bem como a utilização de novas tecnologias a fim de melhorar a forma como as informações são produzidas e divulgadas em linha" (CE, p. 08). No tópico "Plataformas digitais que ajam de forma rápida e eficaz para protegerem os utilizadores da desinformação", propõe para as plataformas digitais e para indústria publicitária um rol de objetivos relacionados à transparência, circulação de conteúdos e verificação. Entre eles:

1) Melhorar significativamente o controle da colocação de anúncios, nomeadamente para reduzir as receitas dos agentes desinformadores e restringir as opções de direcionamento de propaganda política;

2) Garantir a transparência em relação a conteúdos patrocinados, especialmente propaganda política e publicidade temática. Estas medidas devem ser complementadas por repositórios com informação exaustiva sobre os conteúdos patrocinados, que revelem a identidade real do patrocinador, os montantes despendidos e os critérios de direcionamento utilizados. Devem ser instituídos mecanismos semelhantes para que os utilizadores compreendam a razão pela qual foram alvo de um determinado anúncio;

3) Intensificar e demonstrar a eficácia dos esforços para bloquear contas falsas;

4) Facilitar a avaliação de conteúdos por parte dos utilizadores, mediante indicadores da fiabilidade das fontes de conteúdos, baseados em critérios objetivos e aprovados por associações de órgãos de comunicação social, em consonância com os princípios e processos jornalísticos, a transparência no que respeita à propriedade dos meios de comunicação social e a verificação da identidade;

5) Diminuir a visibilidade de desinformação, melhorando a facilidade de localização de conteúdos fiáveis;

6) Estabelecer sistemas de marcação e regras claras para os robôs digitais e assegurar que a sua atividade não pode ser confundida com interações humanas;

7) Capacitar os utilizadores com ferramentas que permitam uma experiência em linha [on-line] personalizada e interativa, a fim de facilitar a descoberta de conteúdos e o acesso a diferentes fontes de notícias representando diferentes pontos de vista; fornecer-lhes instrumentos de acesso fácil para a denúncia de casos de desinformação;

8) Garantir que os serviços digitais incluam salvaguardas contra a desinformação; por exemplo, incluindo informações porme-

norizadas sobre o comportamento de algoritmos que estabelecem a prioridade de apresentação de conteúdos, bem como o desenvolvimento de métodos de ensaio;

9) Fornecer às organizações de verificação de fatos de confiança e aos meios acadêmicos o acesso aos dados da plataforma (nomeadamente através de interfaces de programação de aplicações), respeitando a privacidade dos utilizadores, os segredos comerciais e a propriedade intelectual. Tal permitir-lhes-á melhorar a compreensão do funcionamento de algoritmos conexos e a análise e acompanhamento das dinâmicas de desinformação e do seu impacto na sociedade (CE, p. 08-09).

Além disso, há proposições interessantes voltadas especificamente à verificação de fatos, dispostas no tópico "reforço da verificação de fatos, do conhecimento coletivo e das capacidades de acompanhamento da desinformação". São elas: 1) acompanhar continuamente a dimensão, as técnicas e as ferramentas de desinformação, bem como a sua natureza exata e o seu potencial impacto; 2) identificar e mapear mecanismos da desinformação que contribuem para a ampliação digital; 3) contribuir para o desenvolvimento de indicadores justos, objetivos e fiáveis da transparência das fontes; 4) partilhar conhecimentos com os órgãos de comunicação social, as plataformas e as autoridades públicas a fim de reforçar a sensibilização do público para a desinformação (CE, p. 10).

A Comissão Europeia avança ainda ao estabelecer como resposta um eixo de "Promoção da responsabilização em linha". Opina que a identificação da fonte de desinformação, assegurando a sua rastreabilidade ao longo do processo de divulgação, é essencial para a responsabilização, bem como para reforçar a confiança nos fornecedores de informação identificáveis e para incentivar um comportamento em linha mais responsável. Nesse tópico, à própria comissão, que tem caráter de autoridade pública, são definidas tarefas, particularmente para a investigação de comportamentos

maliciosos on-line, como implementação da versão 6 do protocolo Internet (IPv6) e realização de esforços para melhorar o funcionamento, a disponibilidade e a exatidão das informações constantes do sistema de nomes de domínio (DNS) e do IP WHOIS, um protocolo que informa endereço, e-mail, telefone, servidores DNS, data de criação, entre outros.

Por fim, sugere ainda o "Aproveitamento das novas tecnologias", como inteligência artificial, algoritmos cognitivos, tecnologias que permitam aos meios de comunicação oferecer experiências on-line personalizáveis e interativas e tecnologias como a cifragem progressiva (blockchain) para ajudar os cidadãos a descobrir conteúdos e a identificar desinformação, bem como para ajudar a preservar a integridade do conteúdo, a validar a fiabilidade das informações e/ou das suas fontes, a permitir a transparência e a rastreabilidade.

No caso do informe de David Kaye à ONU (4), como tem como foco a promoção e a proteção da liberdade de opinião e expressão, são feitas propostas que podem ser consideradas como medidas correlatas porque têm o potencial de causar impactos sobre a questão específica da desinformação. Assim como os textos analisados até aqui, as medidas têm caráter amplo, propositivo, estando ancoradas em eixos como transparência e promoção da diversidade. Nesse sentido, aos Estados são propostas ações como a publicação de relatórios detalhados de transparência sobre todas as solicitações relacionadas a conteúdo feitas a intermediários e buscar "genuinamente" a participação do público em todas as considerações normativas.

Para as empresas, sugere: "adotar abordagens à transparência radicalmente diferente em todas as etapas de suas operações, desde o desenvolvimento de padrões até a compilação e aplicação de 'jurisprudência' que enquadre a interpretação de padrões privados. A transparência requer mais interação com organizações de direitos digitais e outros setores relevantes da sociedade civil e recusa

em celebrar acordos secretos com os Estados sobre regras relativas ao conteúdo e sua aplicação. Tendo em conta o seu impacto na esfera pública, as empresas devem abrir-se à prestação de contas" (ONU, p. 08).

O documento do CNDH (5) também é garantista em suas recomendações ao Congresso Nacional (aprovação do que viria a se tornar a Lei Geral de Proteção de Dados Pessoais e de projetos legislativos que respeitem padrões de direitos humanos), ao TSE (adoção de parâmetros de direitos humanos em suas recomendações) e ao Executivo Federal e Estadual (propostas de políticas de alfabetização midiática e campanhas informativas). Às plataformas, recomenda: "a adoção de políticas que garantam transparência sobre o seu funcionamento e as regras das suas comunidades e que ampliem o controle dos usuários sobre os conteúdos que publicam e acessam, incidindo sobre o chamado efeito bolha e a estrutura de monetização que estimula a criação e difusão das chamadas 'notícias falsas'. Além do respeito às normas infralegais e internacionais em matéria de direitos humanos e a adequação das políticas das plataformas às especificidades sociais e culturais dos países onde gerem seus modelos de negócios" (CNDH, p. 04).

A transparência é também um eixo central na declaração sobre eleições (6), que inclusive cita que não deveria ser permitida publicidade política orientada em função de dados pessoais, salvo que as pessoas tenham dado seu consentimento. Transparência sobre gastos de campanha e acordos entre os agentes também é requerida. Quanto à desinformação especificamente, dispõe que os Estados deveriam avaliar a possibilidade de respaldar medidas positivas para abordar o problema da desinformação na rede, como a promoção de instâncias independentes para a verificação de informação e campanhas de educação pública.

Para as plataformas, além da adoção de padrões de direitos humanos, as sugestões também caminham no sentido da transparência, inclusive das práticas publicitárias, e na promoção de infor-

mações. As principais propostas para os intermediários da Internet e a mídia digital são, resumidamente: i) implementar os Princípios Orientadores da ONU Negócios e Direitos Humanos e tomar medidas com a devida diligência para garantir que seus produtos, políticas e práticas não afetam os direitos humanos, incluindo as áreas de coleta privada e de microdados e roteamento de mensagens; ii) fazer esforços suficientes para adotar medidas que permitam aos usuários acessar várias ideias e perspectivas políticas. Em particular, eles devem garantir que as ferramentas automáticas, como algoritmos de classificação, não impeçam acesso a conteúdo relacionado a eleições e disponibilidade de diferentes pontos de vista para os usuários; iii) avaliar se seus produtos, políticas ou práticas de publicidade política limitam arbitrariamente a capacidade de candidatos ou partidos para divulgar suas mensagens; iv) abordar desinformação ou manipulação intencional, bem como spam relacionados às eleições. Isso inclui promover instâncias de verificação independentes e implementar outras medidas, como manter arquivos sobre publicidade política contratada, moderação apropriada de conteúdos e alertas públicos; v) adotar medidas de transparência em relação ao uso e impacto que as ferramentas automatizadas que as plataformas usam podem ter na prática, incluindo até que ponto essas ferramentas afetam a coleta de dados, publicidade direcionada e divulgação, classificação e/ou remoção de conteúdo, especialmente aqueles relacionados a eleições (ONU, OSCE e OEA, p. 04).

Há, em geral, uma abordagem ampla do fenômeno, considerado complexo e multifacetado, propositiva, que aponta como eixos recorrentes das soluções: promover pluralidade e diversidade de informações, estimular a adoção de transparência sobre políticas das plataformas e sobre as práticas publicitárias, garantir maior controle sobre a circulação de conteúdos e fomentar a literacia midiática ou, termos mais utilizado no Brasil, a educação para a mídia. Muitas propostas extrapolam o ambiente digital e se voltam

também à mídia tradicional ou à sociedade em geral, como aquelas que tratam de campanhas informativas. A maior parte dessas medidas é direcionada às plataformas digitais e outros agentes, como verificadores de fatos e empresas midiáticas tradicionais. Ainda que celebrada nos textos, também há ponderações quanto à checagem de fatos, com alertas para que não se constitua um "monopólio da verdade", expressão utilizada pelo grupo de altos especialistas, e para que os verificadores sigam padrões internacionais e sejam diversos. Percebe-se que as propostas voltadas ao Estado estão, sobretudo, centradas na criação de um entorno ambiente propício à liberdade de expressão.

1.4.4 AVALIAÇÃO DA POSSÍVEL MODERAÇÃO DE CONTEÚDOS POR PARTE DAS PLATAFORMAS

Na apresentação dos motivos que justificam o documento de 2017 (1), os relatores afirmam estar "consternados" por "algumas medidas tomadas por intermediários para limitar a consulta ou a difusão de conteúdos digitais, inclusive através de processos automatizados, como sistemas de eliminação de conteúdos baseados em algoritmo ou no reconhecimento digital, que não são transparentes, descumprem os padrões mínimos de devido processo e/ou limitam de maneira indevida o acesso a conteúdos ou sua difusão" (ONU; OEA; OSCE; CADHP, p. 02). Na esteira dessa compreensão, entre os princípios gerais apresentados, consta que: "os intermediários não deveriam ser legalmente responsáveis em nenhum caso por conteúdos de terceiros relacionados com esses serviços, a menos que intervenham especificamente nesses conteúdos ou se neguem a acatar uma ordem ditada em consonância com garantias de devido processo por um órgão de supervisão independente, imparcial e autorizado (como um tribunal) que ordene a remoção de tal conteúdo, e tenham capacidade técnica para fazê-lo" (Ibidem, p. 03).

Menciona ainda que, caso os intermediários pretendam tomar medidas para restringir conteúdos de terceiros (como a eliminação ou a moderação de conteúdos) que excedam o exigido legalmente, deveriam adotar políticas claras e preestabelecidas que regulem tais medidas, que deveriam estar baseadas em critérios objetivamente justificáveis e não em fins ideológicos ou políticos, e, se possível, deveriam ser adotadas após manter consultas com seus usuários (Ibidem, p. 04). O devido processo é posteriormente detalhado no texto, que diz que isto inclui a notificação oportuna aos usuários quando os conteúdos que tenham criado, carregado ou hospedado possam estar sujeitos a uma ação e dê ao usuário a oportunidade de questioná-la. Anota ainda que o exposto vale a processos automatizados.

No texto do grupo de especialistas (2), o tema não é abordado especificamente, mas durante a avaliação das medidas implementadas pelos diferentes agentes apresenta a seguinte avaliação, que pode ser relacionada à moderação de conteúdos: "as más práticas tendem a apresentar riscos para a liberdade de expressão e outros direitos fundamentais. Entre elas estão a censura e vigilância on-line e outras respostas erradas que podem ter um efeito adverso substancial e que podem ser utilizadas por vendedores de desinformação numa narrativa 'nós contra eles', que pode deslegitimar as respostas contra a desinformação e ser contraproducente a curto e longo prazo. Atenção deve também ser dada à falta de transparência e à privatização da censura por delegação a organismos/entidades específicos ou empresas privadas"[19] (HLEG, p. 14), alerta.

19. Tradução própria do original em inglês: "bad practices tend to present risks to freedom of expression and other fundamental rights. They include censorship and online surveillance and other misguided responses that can backfire substantially, and that can be used by purveyors of disinformation in an "us vs. them" narrative that can de-legitimize responses against disinformation and be counter-productive in the short and long run. Attention should also be paid to lack of transparency and to the privatization of censorship by delegation to specific bodies/entities or private companies".

A mesma posição aparece na elaboração da Comissão Europeia (3), que não trata diretamente do tema, mas inclui, no rol de objetivos a serem alcançados na definição do código de boas práticas: "facilitar a avaliação de conteúdos por parte dos utilizadores, mediante indicadores da fiabilidade das fontes de conteúdos, baseados em critérios objetivos e aprovados por associações de órgãos de comunicação social, em consonância com os princípios e processos jornalísticos, a transparência no que respeita à propriedade dos meios de comunicação social e a verificação da identidade" (CE, p. 09). Logo em seguida, alerta que as ações devem respeitar rigorosamente a liberdade de expressão e incluir salvaguardas que impeçam a sua utilização abusiva, por exemplo, a censura de discursos críticos, dissidentes, satíricos ou chocantes.

Mais detalhado é o informe da relatoria da ONU, que dedica uma seção inteira ao tema da moderação de conteúdos por empresas. É justamente no tópico "Esferas de preocupação no âmbito das normas sobre o conteúdo" que há a menção mais detalhada à desinformação. Pontua, nesse sentido, que as empresas estão enfrentando uma pressão cada vez maior para evitar a propagação da desinformação, mas que no entanto, como as medidas vigorosas, como a remoção ou o bloqueio do sítio web, podem resultar em graves interferências com a liberdade de expressão, as empresas devem agir com cuidado ao desenvolverem políticas relacionadas com a desinformação.

O texto menciona que as empresas têm adotado várias respostas, incluindo acordos com terceiros responsáveis pela verificação, controle reforçado do cumprimento das políticas de publicidade, maior vigilância de contas suspeitas, além de mudanças na edição de conteúdos e de algoritmos que dão lugar à classificação das buscas e formação de usuários sobre a detecção de informações falsas. Avalia que algumas dessas medidas, em especial as centradas na restrição do conteúdo das notícias, podem pôr em perigo

fontes de notícias independentes e alternativas ou conteúdos satíricos (ONU, p. 13).

As medidas apresentadas são antecedidas pelos seguintes termos, que situam o enquadramento delas: padrões de direitos humanos; legalidade; necessidade e proporcionalidade; e não discriminação. A maior parte das recomendações trata da transparência das empresas. Sugere, por exemplo, que "as condições de serviço devem abandonar a abordagem discricionária ancorada nas necessidades genéricas e autocentradas da 'comunidade'. Em vez disso, sugere que as empresas devem assumir compromissos políticos de alto nível para manter plataformas onde os usuários possam desenvolver suas opiniões, se expressar livremente e acessar informações de todos os tipos de maneira consistente com os padrões de direitos humanos".

A fim de evitar que as empresas sejam pressionadas a interferir na circulação de conteúdos, sugere que os Estados evitem "delegar para empresas a responsabilidade como avaliadoras de conteúdo, com o que se põe o juízo corporativo sobre os valores dos direitos humanos em detrimento dos usuários". É importante notar que o documento também desaconselha os Estados a adotarem modelos de regulação em que sejam os organismos governamentais, e não as autoridades judiciais, os árbitros do que seja uma expressão legítima.

O CNDH (5) também é explícito ao avaliar que "não cabe às plataformas privadas, de maneira solitária, avaliar quais conteúdos tipificam fake news e que seus padrões de funcionamento e dos seus algoritmos de distribuição e remoção dos conteúdos devem primar pela transparência e estar em consonância com as normas legais nacionais e internacionais sobre liberdade de expressão, dados pessoais e usos da internet". Alerta que "as leis que estabelecem prazos curtos para que as plataformas privadas, como o Google ou o Facebook, removam conteúdos identificados pelas próprias empresas como fake news sem que haja a necessidade

de uma decisão judicial, sob pena de pesadas sanções econômicas, representam um grande risco para a liberdade de expressão" (CNDH, p. 02-03).

No mesmo sentido, o documento sobre eleições (6) explicita que "os intermediários de internet não deveriam se considerar responsáveis pela desinformação, a informação errônea ou a informação manipulada, que tenha sido difundida de forma deliberada em suas plataformas, salvo que intervenham especificamente nesse conteúdo ou que descumpram uma ordem legalmente vinculante de retirar esse conteúdo" (ONU; OSCE; OEA, p. 03).

Ao pesquisar como os documentos abordam a questão da moderação de conteúdos por parte das plataformas digitais, vemos que em nenhum caso há recomendação de que as plataformas sejam responsáveis por eles. Também não foi verificada proposta que leve à interrupção da circulação, como bloqueios de sites, suspensão do funcionamento de aplicações. Esse tipo de medida é referida como extrema. Como tal, só poderia ser justificada pela necessidade e adotada se ancorada em lei. A moderação de conteúdos específicos por parte das plataformas também é abordada com cautela, sendo desestimuladas práticas de suspensão unilateral e recomendada adoção de devido processo. No mesmo sentido, não são recomendados sistemas de filtragem de conteúdos que não sejam controlados pelo usuário final, a quem os documentos, em geral, buscam empoderar.

1.5. ENFRENTANDO A DESINFORMAÇÃO: CONCEITO E PRÁTICAS DEFENDIDAS PELO INTERVOZES

Desde o momento em que o fenômeno da desinformação associada ao ambiente digital veio à tona no debate público brasileiro, o Intervozes tem participado das discussões e produzido conteúdos sobre o tema, a exemplo da cartilha *Desinformação: ameaça ao*

direito à comunicação muito além das fake news e do livro *Desinformação: crise política e saídas democráticas para as fake news*. Além disso, a organização atua em diversas frentes, seja por meio da promoção de debates junto a diferentes agentes, produção de textos de opinião e outros materiais de comunicação para sensibilização da sociedade ou da incidência junto, sobretudo, ao Congresso Nacional, contribuindo na discussão dos projetos de lei já apresentados sobre fake news.

Aqui, objetivamos apenas apresentar, ainda que de forma sumária, a forma com que o Intervozes tem compreendido a questão e as propostas que têm lançado para o debate público. No livro citado antes, mencionamos que, assim como outros pesquisadores, grupos da sociedade civil e instituições, evitamos utilizar a expressão fake news, pois entendemos que ela não é capaz de explicar o problema para além de sua aparência. O Intervozes tem "optado por adotar o conceito de desinformação, com o qual se busca ressaltar a intencionalidade na produção e na propagação de informações falsas, equivocadas ou descontextualizadas para provocar uma crise comunicacional e, assim, obter ganhos econômicos e/ou políticos" (MARTINS, 2020, p. 12-13).

A compreensão que nos orienta é a de que a desinformação não é um fenômeno recente nem ocasional. Assim como aconteceu em vários momentos históricos por meio do controle dos meios de comunicação, com destaque para a televisão, a desinformação integra processos de disputas de sentido e pelo poder, estando, portanto, em geral associada a interesses de determinados agentes. O que ocorre no momento atual deve ser pensado tendo em vista tanto a profunda crise política que atinge diferentes sociedades em um contexto de crise sistêmica, inclusive no Brasil, bem como a utilização da Internet em sua configuração progressivamente marcada pela presença de monopólios digitais.

A desinformação é, a nosso ver, um subproduto do modelo de negócio estabelecido pelas plataformas digitais, um modelo que

se baseia na concentração e na exploração indevida de dados pessoais de usuários/as, por isso é fundamental que as plataformas sejam instadas a desenvolver medidas, mas também que não façam isso de forma isolada, pois, como espaços públicos cada vez mais relevantes e impactantes, devem ser objeto de regulação e acompanhamento de suas práticas de forma multissetorial e transparente.

Das tecnologias da informação e da comunicação, de forma mais ampla, advêm particularidades do fenômeno na contemporaneidade, como maiores velocidade e alcance na difusão das mensagens e maior opacidade em todo o processo de comunicação, mas é sempre necessário registrar que, do nosso ponto de vista, os meios tradicionais também produziram e seguem produzindo desinformação. Além disso, o afastamento da população do debate sobre a comunicação, que favoreceu a concentração de propriedade e diversas violações de regras e direitos por parte dos grupos tradicionais, também deixa marcas como a falta de conhecimento sobre os processos operacionais e de editoração, bem como de dificuldades para parte da população desenvolver uma leitura crítica da mídia, seja das tradicionais ou das novas.

Partindo dessa leitura, entendemos que as saídas para a desinformação e para a construção de um fluxo de informações saudável para a cidadania e democracia são complexas, devem envolver diversos agentes e ser planejadas tendo em vista objetivos de curto, médio e longo prazos. Nenhuma proposta isolada será capaz de resolver um problema multidimensional e que tangencia questões sensíveis, como a discussão sobre a verdade e a garantia de direitos fundamentais como à liberdade de expressão e à proteção de dados pessoais. É preciso que sejam apontadas saídas que transformem as estruturas que favorecem a ocorrência do problema, além de medidas que enfrentem a produção da desinformação em escala industrial.

Abaixo, listamos as dez propostas apontas como centrais pelo Intervozes para combater a desinformação:

1. Punir o uso artificial e industrial de plataformas de mensageria instantânea

Por trás de estratégias de campanhas de desinformação no WhatsApp, estão empresas que realizam métodos de automação e envio massivo. Disparar mensagens em massa a partir de sistemas automatizados é uma prática proibida pelo Tribunal Superior Eleitoral (TSE), mas é preciso avançar na identificação e punição de tais empresas, além de ser necessário discutir as próprias estruturas das plataformas.

O WhatsApp, empresa do conglomerado Facebook, durante o período eleitoral, anunciou o banimento de centenas de milhares de contas do aplicativo. Os perfis são suspeitos de compartilhar notícias falsas e foram rastreados com ajuda de um filtro que identifica spam automaticamente. Segundo a empresa, a lista inclui números de agências que comercializam disparo de mensagem em massa, mas tais empresas, até o momento, não foram devidamente investigadas, o que mostra também a necessidade de haver uma continuidade e uma resposta baseada em investigações.

2. Direito de resposta rápido e proporcional

Para combater a desinformação, também se faz necessário buscar rápidos instrumentos de reparação do dano, considerando o alcance e a velocidade da distribuição da mensagem que motivou a garantia de direito de resposta, a fim de que este possa chegar a um público amplo e potencialmente atingido. Ao conceder direito de resposta, o Judiciário deve levar em consideração o alcance e a velocidade da publicação do conteúdo que motivou o pedido de reparação e propor medidas que viabilizem repercussão ampla. A disseminação das mensagens de desmentido ou esclarecimento deve chegar aos usuários atingidos pelos conteúdos enganosos na

mesma velocidade ou até mesmo em rapidez maior, em casos específicos (como na reta final de eleições).

3. Proteger os dados pessoais e fiscalizar seu uso

Boa parte da operação de distribuição de desinformação gerada intencionalmente e de forma massiva para produzir dano está assentada no uso de dados pessoais. A Lei Geral de Proteção de Dados (Lei 13.709/2018) foi aprovada justamente para garantir os direitos dos titulares de dados, criar bases legais para estes tratamentos e fiscalizar o uso deste importante ativo na era da informação. Se a LGPD for respeitada, a efetividade das práticas de desinformação poderá ser reduzida e, assim, desestimulada. Uma das medidas já recomendadas pela ONG Safernet é a vedação à utilização de características do público-alvo para direcionamento de publicidade por meio dos impulsionamentos, como aquelas relacionadas a atributos sensíveis como origem racial ou étnica, convicções religiosas, filiação a sindicatos ou organizações de caráter religioso, dados referentes à saúde ou à vida sexual.

É preciso, porém, que a Autoridade Nacional de Proteção de Dados (ANPD) – prevista em lei, mas ainda não criada no momento em que finalizamos esta pesquisa – tenha autonomia política e independência financeira para realizar o trabalho de educação, fiscalização e sanção da lei. Vale também atualizar a legislação eleitoral para adequá-la à LGPD, colocando o cidadão no controle do uso que partidos e campanhas eleitorais fazem hoje de seus dados.

4. Vedar a publicidade de candidatos e campanhas on-line

Recentemente, o Twitter anunciou que não mais aceitará publicidade política em sua plataforma. Não apenas a publicidade paga

por candidatos será coibida, mas a publicidade política como um todo. Isso porque o Twitter entende que os anúncios políticos da Internet apresentam desafios totalmente novos ao discurso cívico: otimização baseada em aprendizado de máquina de mensagens e segmentação múltipla, informações enganosas não verificadas e as deep fakes, montagens imperceptíveis. Tudo com velocidade crescente, sofisticação e escala esmagadora. Se o Twitter pode fazer, por que as demais plataformas não seguem este caminho?

Da mesma forma que na era da radiodifusão se definiu que as inserções de campanha seriam regidas por regras de forma a dar visibilidade proporcional ao tamanho das coligações, e não apenas ao poder econômico das candidaturas, também na Internet não pode vigorar a regra de que quem paga mais sai ganhando. O fim da publicidade paga de candidatos na Internet também impactará de maneira positiva na distribuição de desinformação paga on-line.

Em 2018, a organização SaferNet apresentou 14 propostas para o Tribunal Superior Eleitoral (TSE) sobre essa questão, entre elas a vedação aos chamados *hidden posts* ou *dark posts*, postagens pagas direcionadas a um público específico que o resto da população não consegue ver; e vedação do pagamento de anúncios e impulsionamento de conteúdo político em moeda estrangeira, conforme documento entregue, porém não assimilado pelo Tribunal.

Acrescentamos a necessidade das plataformas oferecerem ao TSE, durante o período eleitoral, visibilidade acerca de publicações de grande alcance, mesmo as não impulsionadas, que citem candidatos, coligações e partidos, de forma que o órgão proativamente analise a necessidade de tomar medidas, como oferecer direito de resposta ou sugerir restrições ao alcance de determinados conteúdos.

5. Acesso integral às aplicações de Internet

A maioria dos brasileiros hoje acessa a Internet por banda larga móvel, no celular, e não dispõe de pacotes de dados perene. Impera, assim, no Brasil, o uso limitado da Internet a poucas aplicações, por meio dos chamados acordos de *zero rating* (ou tarifa zero) que as operadoras de telefonia firmam com determinadas plataformas. Neste modelo, a navegação acaba restrita a tais aplicações e a checagem de informações, assim como a leitura de uma matéria em sua integralidade no site de origem, por exemplo, são comprometidas, fortalecendo o fenômeno da desinformação com base em imagens e títulos, que não raras vezes se utilizam de estratégias chamativas em busca de cliques e não apresentam os fatos em sua complexidade.

Assim, conter a desinformação requer a ampliação do acesso à conexão de qualidade e a preços módicos e à Internet integral. Uma das medidas importantes neste sentido é proibir o bloqueio da navegação após o fim da franquia de dados, garantindo às operadoras do Serviço Móvel Pessoal e do Serviço de Comunicação Multimídia a liberdade de reduzir a velocidade de navegação, desde que não inviabilize o direito de acesso.

6. Transparência e autonomia diante de algoritmos

É fundamental considerar o efeito dos novos modelos de distribuição de conteúdo na construção da esfera pública. Nesse sentido, é relevante o impacto do uso de algoritmos na filtragem e priorização de conteúdo, na busca por notícias e informações em plataformas. O viés na escolha de informações sempre existiu, mas a possibilidade de circunscrição não consciente do cidadão em determinado espectro de informações políticas que se tem hoje é algo completamente diferente e gera o efeito bolha, um dos

resultados negativos não previsto e não intencional deste modelo de negócios.

Para reverter danos, é importante assegurar a possibilidade de escolha do conteúdo visualizado pelo usuário (em feeds de notícias, resultados de busca, serviços de acesso a notícias e similares) e tornar transparentes os critérios utilizados pelos algoritmos para a sua ordenação e/ou direcionamento, se possível explicitando os efeitos para o usuário. As grandes plataformas devem fornecer mecanismos de filtragem personalizados de forma transparente, explícita, revogável/editável e sob o controle do usuário, para que eles sejam os que finalmente decidam o conteúdo que desejam priorizar e como fazê-lo (por exemplo, podem preferir ver conteúdos postados por ordem cronológica e não a partir da definição de seu perfil) e nas ações de priorização de conteúdos on-line acessíveis ao usuário (feeds, resultados de busca e outros). Acordos comerciais que interferem na organização dos conteúdos devem ser informados de forma visível junto aos conteúdos.

É fundamental, ainda, que se avance em medidas positivas de promoção de diversidade e pluralidade nas plataformas, permitindo que os algoritmos sejam utilizados para ampliação do fluxo comunicacional, situação inversa ao que temos atualmente, em que alguns conteúdos não têm alcance desejado na Internet por causa da priorização algorítmica.

7. Quebra dos monopólios e oligopólios digitais

Como nos meios tradicionais, o cenário de concentração nos meios digitais impede a diversidade e pluralidade no debate público, fragilizando as sociedades democráticas. Nesse sentido, considerando a especificidade da comunicação por sua natureza simbólica (cultural/social/política), é preciso aliar ações de promoção da concorrência e políticas antitruste às teorias e práticas

de proteção e promoção da liberdade de expressão, pluralismo e diversidade.

Por este entendimento, é preciso que reguladores e autoridades responsáveis pela aplicação da concorrência e das políticas públicas para a comunicação desenvolvam análises e ações ex ante, como prognósticos sobre o impacto da concentração econômica sobre a diversidade e pluralidade. Mais tarde, com os mercados já definidos, é muito difícil corrigir assimetrias que acabam se tornando estruturais. Assim, devem ser efetivadas restrições à propriedade horizontal, vertical e cruzada para agentes com poder de mercado significativo. No caso das grandes plataformas, é importante haver desmembramento dos vários negócios de um mesmo agente.

8. Incentivos ao jornalismo responsável, plural e diverso

A diversidade de vozes e posições é um fator essencial para a saúde da democracia. Assim, é preciso avançar em regras para impedir o monopólio e oligopólio nos meios de comunicação, como requer a Constituição, e elaborar mecanismos de incentivo à produção de conteúdo responsável que visem à promoção do interesse público. Ainda, é fundamental que, para a aplicação dos recursos públicos em veículos de comunicação, sejam estabelecidas regras objetivas e amplamente divulgadas, que considerem não apenas a audiência, mas a diversidade, bem como o compromisso com o interesse público, em consonância com o artigo 221 da Constituição. Ainda que estas sejam diretrizes para a comunicação que utiliza de espectro público, devem ser guias também para o investimento de recursos públicos no setor midiático em geral.

9. Educação midiática

É preciso criar, desde já, instrumentos para fomentar a análise crítica sobre os meios de comunicação tradicionais e novos junto ao conjunto da população, em larga escala. A escola é certamente um local privilegiado de formação, mas não deve ser o único. Vale ressaltar que qualquer formação para a mídia precisa ir além da análise dos conteúdos e abordar também a estrutura econômica dos meios no Brasil e no mundo, bem como a regulação prevista no país e em outras jurisprudências. Apenas assim será possível formar cidadãos críticos quanto ao fluxo de informações, considerando conteúdos e também interesses políticos e econômicos das empresas de comunicação, não necessariamente observáveis na análise individualizada das produções de texto, vídeo etc.

Concretamente, é preciso desenvolver, a partir da participação multissetorial, a proposição de um processo pedagógico amplo que envolva os meios de comunicação e a escola; inserir a educação para mídia no currículo escolar, partindo da premissa da escola como exercício reflexivo do que acontece na vida das pessoas; bem como apoiar, por meio de políticas públicas, ações de educação midiática desenvolvidas por organizações da sociedade civil nos diversos territórios.

10. Transparência na comunicação dos servidores

Pela ausência de procedimentos que garantam o armazenamento das comunicações dos servidores públicos e sua disponibilização periódica para controle social, vale a discussão acerca da necessidade de regulação dos canais oficiais de comunicação à disposição dos servidores, bem como dos processos para guarda, armazenamento e posterior disponibilização pública destes registros públicos para acesso pela sociedade brasileira.

CAPÍTULO 2

MEDIDAS ADOTADAS PELAS PLATAFORMAS DIGITAIS

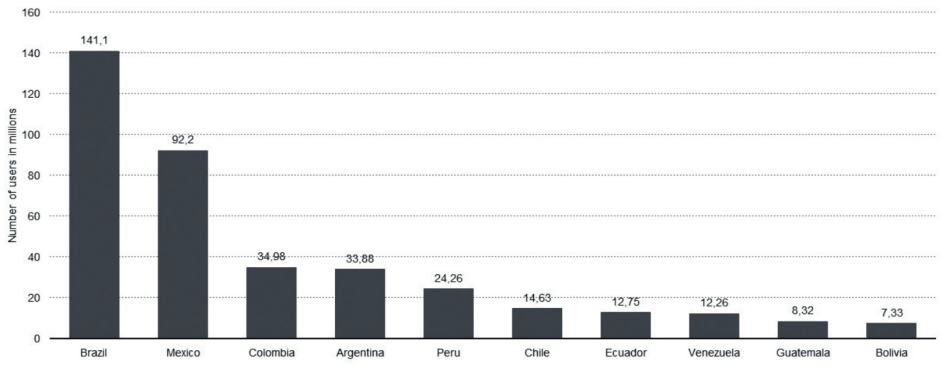
Neste capítulo, serão apresentadas as medidas adotadas entre 2018 e 2020 que tratam diretamente de desinformação ou que podem impactar a circulação de conteúdos desse tipo nas redes. Além de dados sobre fundação e penetração, particularmente no Brasil, as ações de cada rede serão organizadas a partir das categorias que construímos para a análise e que foram detalhadas no primeiro capítulo, quais sejam: 1. Abordagem do fenômeno; 2. Moderação de conteúdo; 3. Promoção de informações e transparência; e 4. Medidas correlatas.

2.1. FACEBOOK

Segundo ranking da consultoria Statista, em abril de 2020 o Facebook comandava os quatro principais serviços de rede social no mundo: Facebook app (2,498 bilhões de usuários), WhatsApp (2 bilhões), Facebook Messenger (1,3 bilhão) e Instagram (1 bilhão). Chegam próximas redes sociais chinesas, como WeChat e QQ, mas

ainda com uma penetração restrita ao país e a migrantes deste vivendo em outras nações do globo.

A empresa é dominante neste segmento no mundo e no Brasil, com 141 milhões de usuários (STATISTA, 2020), conforme mostra a tabela abaixo:



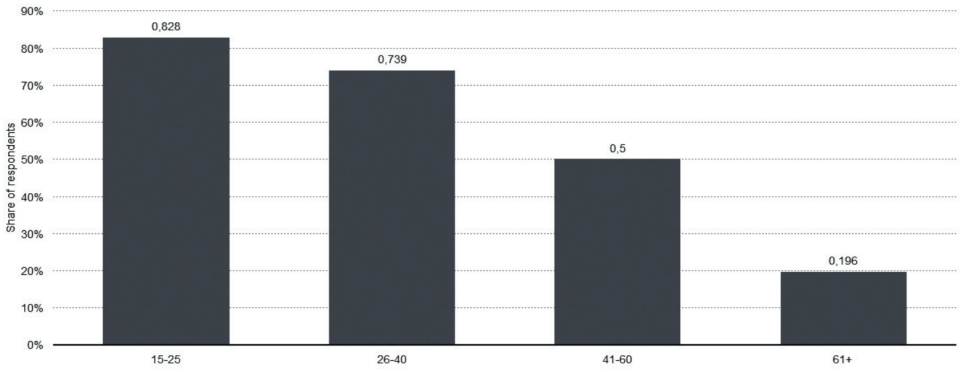
Fonte: Statista, 2020

O Facebook utiliza suas aplicações de forma combinada para reforçar seu poder de mercado. Em pesquisa anterior, o Intervozes mostrou o predomínio do Facebook na camada de aplicações e conteúdos da Internet, liderando os rankings de apps mais baixados, de número de downloads e estando bem posicionado entre os sites mais acessados (VALENTE, PITA, 2018).

Ainda de acordo com a pesquisa *Social Media Usage in Latin America*, da consultoria Statista (2020), o Brasil era o 4º mercado da empresa no mundo (atrás da Indonésia, Estados Unidos e Índia) e o 1º da América Latina. Entre os usuários, a proporção por gêneros é semelhante.

Fonte: Statista, 2020

Já a distribuição por idade é mais diversificada, com maior presença de jovens entre 15 e 25 anos, e variando depois dos mais jovens aos mais velhos.



Fonte: Statista, 2020

2.1.1 ABORDAGEM DO FENÔMENO

A política da empresa sobre desinformação não possui um documento específico, e é composta por um conjunto de normas, procedimentos e iniciativas adotadas nos últimos anos. O Facebook trabalha com algumas estratégias contra desinformação, destaca-

damente: 1) identificar conteúdos falsos; 2) promover restrições a este conteúdo, como redução da circulação e fornecimento de alertas; 3) aplicar sanções a produtores recorrentes de desinformação; 4) oferecer informações adicionais; e 5) remover contas por outras violações, mas que impactam indiretamente os produtores de desinformação. As políticas são apresentadas de forma resumida nos Padrões da Comunidade[20].

Em termos de estrutura institucional dedicada ao tema, o Facebook anunciou, em 2019, a criação de um conselho de supervisão (*oversight board*). O corpo será formado por 40 pessoas com mandato de três anos. Tanto o Facebook quanto usuários poderão remeter casos considerados complexos para avaliação do conselho. O FB se comprometeu a encaminhar as decisões do conselho em até sete dias. O conselho também pode fazer recomendações sobre políticas da plataforma, que serão revisadas pela equipe desta. Até abril de 2020, o projeto ainda não havia sido implementado. Abaixo, será detalhada a cronologia das ações referidas até aqui e de outras medidas adotadas pela plataforma nos últimos dois anos.

Em setembro, o diretor de governança e assuntos globais, Brent Harris, publicou texto[21] onde apresentava a estrutura e detalhes do projeto de criação de um conselho de supervisão, cuja existência foi indicada pelo diretor executivo, Mark Zuckerberg, em novembro de 2018. O principal documento trazendo essas informações (*the charter*) estabeleceu o modus operandi do ente, que passaria a ser responsável pela revisão de casos envolvendo conteúdos a partir de demanda tanto de usuários quanto do próprio FB.

A equipe do conselho faz a análise de casos remetidos. O conselho deve, então, escolher determinados processos para um exame mais aprofundado. Conforme o Facebook, caberá ao colegia-

20. Informação disponível em: https://www.facebook.com/communitystandards/false_news.
21. HARRIS, Brent. Preparing the Way Forward for Facebook's Oversight Board. Publicada em: 28 jan. 2020. Disponível em: <https://about.fb.com/news/2020/01/facebooks-oversight-board/>. Acesso: 14 dez. 2020.

do fixar os critérios para a seleção desses casos, considerando sua complexidade e relevância global. Quando houver uma escolha, um grupo de conselheiros será indicado para a averiguação. Serão encaminhadas informações da pessoa que submeteu o recurso e do Facebook. Após a conclusão da análise, esta será publicada no site do órgão. Em outubro de 2020, o conselho começou a receber casos e, em dezembro do mesmo ano, o órgão anunciou os episódios selecionados para avaliação.

Em entrevista à equipe da pesquisa, o Facebook informou que o contingente trabalhando com análise de conteúdo foi triplicado nos últimos anos, alcançando 35.000 pessoas tanto para a rede social quanto para o Instagram.

A elaboração das políticas para enfrentar a desinformação é realizada pela equipe de "Políticas Globais de Conteúdo". De acordo com a empresa, essa definição passa por diálogos com especialistas externos, como acadêmicos e jornalistas. A cada "conjunto de semanas" (a companhia em sua entrevista não explicitou a periodicidade) são conduzidas reuniões de avaliação a partir de pesquisas e novas dados sobre essas práticas.

2.1.2 MODERAÇÃO DE CONTEÚDO

Para identificar os conteúdos, há três métodos. O primeiro é pelo recurso de denúncia juntamente a um conteúdo, que não é tomado diretamente, mas enquadrado como um "sinal" de potencial de revisão da mensagem. O segundo envolve a análise dos sistemas automatizados da plataforma, que utilizam tecnologias de aprendizagem de máquina (*machine learning*) para analisar as publicações. Mesmo que o autor da publicação não tenha criado o conteúdo, mas compartilhado algo proveniente de terceiros, ele é considerado responsável por aquela publicação. Um usuário pode classificar determinada publicação como falsa.

O terceiro método é a verificação feita por meio de parcerias com agências de checagem de fatos em diversos países, as quais são instadas e mesmo financiadas pelo Facebook para proceder com verificações ou podem realizá-las por conta própria e comunicar à plataforma. Uma das exigências para que essas checagens embasem ações dela é que sejam certificadas pela Rede Internacional de Verificadores de Fatos (IFCN, na sigla em inglês).

Em entrevista à equipe da pesquisa, a empresa afirmou que não possui uma definição própria de desinformação (*misinformation*). Mas o exame por checadores classifica o conteúdo em nove categorias:

1. **Falso:** uma ou mais das afirmações principais do conteúdo são factualmente imprecisas. Em geral, corresponde a classificações do tipo 'falso' ou 'na maior parte, falso' nos sites dos verificadores de fatos.
2. **Parcialmente falso:** uma ou mais das afirmações do conteúdo são uma mistura de fatos precisos e imprecisos ou a afirmação principal é enganosa ou está incompleta.
3. **Verdadeiro:** uma ou mais das afirmações principais do conteúdo são factualmente precisas. Em geral, corresponde a classificações do tipo 'verdadeiro' ou 'na maior parte, verdadeiro' nos sites dos verificadores de fatos.
4. **Título falso:** uma ou mais das afirmações principais do conteúdo do corpo do artigo são verdadeiras, mas a afirmação principal no título é factualmente imprecisa.
5. **Não qualificado:** o conteúdo inclui uma afirmação que não é comprovável, que era verdadeira quando foi escrita ou que vem de uma Página ou de um site cujo objetivo principal é expressar a opinião ou a agenda de uma figura política.
6. **Sátira:** o conteúdo é publicado por uma Página ou por um domínio que faz publicações de sátira conhecidas ou que uma pessoa sensata entenderia que o conteúdo é ironia

ou humor com uma mensagem social. Ainda assim, seria melhor se o conteúdo apresentasse contexto adicional.

7. **Opinião:** o conteúdo defende ideias e tira conclusões com base na interpretação de fatos e dados, informando ao público o que o autor ou colaborador pensa sobre determinado evento ou tema. As opiniões podem incluir citações ou fatos relatados, mas enfatizam o pensamento, as preferências pessoais e as conclusões do autor. Elas também podem incluir editoriais, apoios ou conteúdo com a palavra 'opinião' no título, escritos por um colunista conhecido por compartilhar opiniões ou compartilhados de um site ou de uma Página cujo objetivo principal é expressar as opiniões ou agendas de figuras públicas, think tanks, ONGs e empresas.
8. **Gerador de pegadinhas:** sites que permitem que os usuários criem as próprias notícias com pegadinhas para compartilhar em sites de redes sociais.
9. **Sem classificação:** esse é o status padrão antes de um conteúdo ser verificado ou se a URL não estiver funcionando. Deixar o conteúdo nesse status (ou retorná-lo para essa classificação depois que outra já tiver sido dada) indica que não devemos tomar providências com base na classificação.

No caso de conteúdos marcados como falso por verificadores, são apresentados alertas em relação a essa publicação. Esses incluem tanto um rótulo específico quanto um questionamento quando o usuário atua para compartilhar o post. Além disso, usuários recebem uma notificação caso um conteúdo repassado seja classificado como falso posteriormente. Quando um conteúdo é rotulado como desinformativo, se ele foi publicado em um grupo, o administrador também é alertado[22]. Também são oferecidos

22. https://about.fb.com/news/2020/09/keeping-facebook-groups-safe/.

outros conteúdos de checadores de fatos com esclarecimentos e contrapontos, recurso chamado de "artigos relacionados". Estes visam oferecer o fato em relação ao conteúdo ou uma outra perspectiva quando este não é faticamente falso, mas se enquadra em outras categorias intermediárias.

Em abril de 2019, a empresa discutiu em seu site o que chamou de "nova fase no combate às notícias falsas"[23]. As ações realizadas até então foram classificadas como importantes, mas havia o problema de cobrir a grande escala de publicações, que somavam, então, cerca de 1 bilhão por dia. Foi iniciada, assim, a discussão sobre uma iniciativa de elaboração juntamente a experts externos de formas de dar resposta à demanda de análise desse montante de informação. Uma das medidas é considerar não apenas as denúncias, mas a possibilidade de pessoas indicarem fontes jornalísticas que contradigam as alegadas mensagens enganosas.

Em texto no blog oficial, a companhia reconheceu[24] os limites da checagem. Em primeiro lugar, não há agências em todos os países e, quando há, elas não necessariamente conseguem dar conta de todas as demandas de conteúdos. "Para fazer progresso real, temos que manter melhorando nossos sistemas de aprendizagem de máquina e tentar outras táticas que possam funcionar melhor no mundo", declarou a gerente de produto Tessa Lyons. Além disso, há diferenças entre os países tanto em parâmetros de jornalismo quanto em liberdade de imprensa. O tratamento de opinião e sátira também colocam desafios, pois uma linha tênue separa esse tipo de conteúdo da desinformação.

23. SILVERMAN, Henry. The next phase in fighting misinformation. Publicado em: 10 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/tackling-more-false-news-more-quickly/.

24. LYONS, Tessa. Hard Questions: How Is Facebook's Fact-Checking Program Working?. Publicado em: 14 jun. 2018. Disponível em: https://about.fb.com/news/2018/06/hard-questions-fact-checking/.

Em novembro de 2019, a companhia divulgou entre suas medidas para as eleições gerais do Reino Unido uma mudança no design dos rótulos de conteúdos verificados. O intuito foi deixá-lo mais visível no caso de classificações como falsos ou parcialmente falsos[25]. Nos Estados Unidos, para as eleições de 2020 os rótulos foram alterados neste sentido, colocando um filtro cinza que diferencia a imagem das demais.

Discursos originais e anúncios de políticos não passam, entretanto, por processos de checagem, ficando longe dessa avaliação. O único filtro é a recusa de inclusão em um anúncio político de um conteúdo que "já foi refutado no Facebook". "Isso é diferente de uma afirmação ou declaração feita por um político. Se uma afirmação for feita diretamente por um político na página dele, em um anúncio ou no site dele, ela será considerada um discurso direto. Por isso, ela não estará qualificada para nosso programa de verificação de fatos terceirizada, mesmo que a essência da afirmação tenha sido refutada em outro lugar". A companhia pede aos verificadores que considerem "políticos em todos os níveis", o que inclui "representantes eleitos, seus indicados, partidos políticos, seus líderes", além de "ex-candidatos e ex-representantes", abarcando organizações de apoio como "super PACs".

Para não classificar os discursos e anúncios políticos, a empresa argumenta possíveis impactos na liberdade de expressão e no debate público. Os discursos de políticos são tratados como exceção e como válidos (*newsworthy*). "Não acreditamos que é um papel apropriado para nós arbitrar debates políticos e prevenir discursos de políticos de atingir sua audiência e ser sujeito ao debate e escrutínio públicos. É por isso que o Facebook excepciona políticos do programa de checagem de terceiros", explicou o vice-presidente de

25. Facebook. How Facebook Has Prepared for the 2019 UK General Election. Publicado em: 7 nov. 2019. Disponível em: https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/.

comunicações e assuntos globais, Nick Clegg[26], em post publicado no site da companhia no dia 24 de setembro de 2019.

O enquadramento como "*newsworthy*" implica a avaliação se o interesse público em determinado conteúdo supera o risco de dano contido nele. Nessa análise, acrescentou Clegg, são considerados aspectos como: se há eleições no país, a natureza do discurso, o objeto dele e a estrutura da nação, incluindo se ela possui uma mídia livre. As exceções trazem aspectos relevantes, uma vez que acabam contribuindo para a visibilidade das opiniões dos líderes políticos para a população. Contudo, é preciso encontrar um equilíbrio para que não se torne uma naturalização do discurso desinformativo. Por isso, um caminho pode ser incluir os alertas sobre as verificações e inserir limitações, como a impossibilidade de compartilhamento ou interação.

No segundo semestre de 2020, a companhia anunciou que passaria a identificar os conteúdos de exceção enquanto tais. Mas em junho de 2020 a rede social excluiu um anúncio do presidente Donald Trump justificando que este trazia símbolos nazistas[27].

De acordo com a empresa, anúncios sobre "questões sociais, eleições ou política" são "feitos em nome ou sobre um candidato para cargo público, uma figura política, um partido político ou militantes atuando para influir no resultado de eleições para cargos públicos; ou sobre qualquer eleição, referendo ou iniciativa de consulta em cédula, incluindo incentivo à votação ou campanhas eleitorais; ou sobre assuntos sociais em qualquer lugar onde o anúncio está sendo veiculado; ou regulado como anúncio político". Há definições diferentes dos "temas sociais" a depender

26. CLEGG, Nick. Facebook, elections and political speach. Publicada em: 24 set. 2019. Disponível: <https://about.fb.com/news/2019/09/elections-and-political-speech/>. Acesso: 14 dez. 2020.

27. O'SULLIVAN, D. Facebook says it took down Trump ads because they used Nazi symbol. CNN Business. Publicada em: 19 jun. 2020. Disponível em: https://edition.cnn.com/2020/06/18/tech/facebook-trump-ads-triangle-takedown/index.html.

do país. Na página institucional sobre o assunto, não há uma listagem própria para o Brasil[28].

Desde 2018, a equipe do Facebook analisa também imagens e vídeos. Esse material é decomposto em três categorias: manipulada, tirada de contexto e com textos ou áudios falsos[29]. No caso dos vídeos, entram aí os chamados "*deep fakes*". São consideradas nessa categoria "mídias manipuladas" (*manipulated media*), quando são editadas ou "sintetizadas" de formas imperceptíveis para uma pessoa ou que possam levar o espectador a ser enganado sobre o sujeito do vídeo ter dito algo que não disse, bem como formas de uso de inteligência artificial que mescle ou altere a imagem buscando fazê-la parecer autêntica[30].

A identificação é o requisito para ações sobre esses conteúdos. Não há remoção diretamente por uma mensagem ser falsa, mas há áreas limítrofes, como a exclusão de conteúdos por risco de danos graves e violência, informações falsas sobre vacinas e mensagens enganosas sobre processos eleitorais e que visem obstaculizar o ato do voto. A verificação de uma publicação como falsa provoca a perda de visibilidade e tem a circulação reduzida na plataforma. Essa diminuição era em 2018 de, em média, 80%. "Uma vez que um conteúdo é classificado como falso pelos verificadores de fatos, reduzimos sua distribuição e exibimos marcações de aviso com mais contexto. Com base em uma verificação de fatos, podemos iniciar métodos de detecção para identificar conteúdos similares", explicou o vice-presidente de Integridade, Guy Rosen, em post no blog da empresa no dia 16 de abril de 2020. Segundo a empresa,

28. Informação disponível em: https://www.facebook.com/business/help/214754279118974?id=288762101909005&helpref=search&sr=11&query=political%20figure.

29. LYONS, Tessa. Seeing the truth. Publicada em: 19 set. 2018. Disponível em: https://about.fb.com/news/2018/09/inside-feed-tessa-lyons-photos-videos/.

30. Em entrevista à equipe da pesquisa, o Facebook afirmou que vem buscando melhorar a detecção de mídias manipuladas por meio de projetos como o "Deepfake Detection Challenge" com bolsas para averiguação desses conteúdos. Outra iniciativa foi a construção de uma parceria com a Reuters para qualificar redações na análise dessas mídias.

quando as pessoas veem uma marcação, em 95% das vezes elas não consultam o conteúdo que deu origem ao link. Além disso, os usuários são avisados de que um determinado conteúdo foi checado quando interagiu ou tenta interagir com ele[31].

Em agosto de 2020, foram divulgados os padrões de referência para recomendações da plataforma, como contas para estabelecer amizade, páginas para curtir ou grupos para integrar[32]. O texto na central de ajuda da plataforma elenca cinco categorias que "talvez não estejam qualificadas para recomendações", entre elas conteúdos falsos ou enganosos. Nas diretrizes para recomendações de contas, o informe coloca que a plataforma "tenta não recomendar contas" que difundam os conteúdos das cinco categorias, entre eles os falsos e enganosos. Também foi listada como uma condição para "tentar não recomendar contas" a disseminação de desinformação sobre vacinas e que publicaram informação falsa conforme agências de checagem parceiras.

Em outubro de 2020, o Facebook reduziu o alcance de uma notícia publicada pelo *New York Post* com acusações contra o candidato democrata Joe Biden[33] antes da verificação por uma agência. A medida foi objeto de críticas. Cristina Tardáguila, da Rede Internacional de Checadores de Fatos (IFCN, na sigla em inglês), publicou artigo questionando a conduta da plataforma. "Ao que parece, o Facebook e o Twitter decidiram assumir o posto de árbitros da verdade, posição que rechaçaram por muito tempo. São ingênuos aqueles que creem que isso não é perigoso ou que não têm conexão com as eleições dos Estados Unidos"[34].

31. JIN, Kang-Jing. Keeping People Safe and Informed About the Coronavirus. Facebook. Publicada em: 22 abr. 2020. Disponível em: https://about.fb.com/news/2020/04/coronavirus/.

32. https://www.facebook.com/help/1257205004624246.

33. MORRIS E FONROUGE, 2020. https://nypost.com/2020/10/14/hunter-biden-emails-show-leveraging-connections-with-dad-to-boost-burisma-pay/.

34. TARDAGUILA, Cristina. Sin metodología o transparencia, Facebook y Twitter se vuelven 'árbitros de la verdad'. Univision Noticias. Publicada 15 out. 2020. Disponível

Já no tocante aos anúncios, a Política de Anúncios estabelece a proibição da veiculação de mensagens publicitárias "que contenham reivindicações refutadas por verificadores de fatos independentes ou, em determinadas circunstâncias, por organizações com uma especialidade específica". Os anunciantes que veicularem mensagens falsas repetidamente podem sofrer sanções como "ter a capacidade de anunciar no Facebook restringida"[35].

Páginas que difundirem esses conteúdos de forma recorrente são objeto da redução de circulação de seus conteúdos. Entre as sanções para esses agentes está a proibição de utilizar o recurso dos "artigos instantâneos", um sistema de publicação de notícias de carregamento mais rápido e com um sistema de inserção de anúncios próprio para veículos jornalísticos.

Além disso, essas páginas poderão ter sanções em suas atividades comerciais na plataforma, como o impedimento de rodar anúncios e de impulsionar conteúdo. Elas também serão proibidas de se registrar como página de notícias. Se as páginas interromperem o comportamento, as sanções deixam de ter efeito e voltam a ter acesso aos recursos integralmente, bem como à circulação das publicações. Os conteúdos classificados como falsos também não poderão ser inseridos em anúncios.

O Facebook também empregou a estratégia de redução de circulação para temáticas específicas. Em março de 2019, a empresa divulgou medidas adotadas para combater a desinformação relacionada a vacinas na plataforma[36]. Em outubro de 2020, um texto do chefe de saúde, Kang-Xing Jin, e do diretor de gestão de produto, Rob Leathern, anunciou que a companhia deixaria de auto-

em: https://www.univision.com/noticias/opinion/sin-metodologia-o-transparencia-facebook-y-twitter-se-vuelven-arbitros-de-la-verdad. Acesso em 8 dezembro 2020.

35. Informação disponível em: https://www.facebook.com/policies/ads/prohibited_content/misinformation.

36. BICKERT, Monica. Combatting vaccine misinformation. Publicado em: 7 mar. 2019. Disponível em: <https://about.fb.com/news/2019/03/combatting-vaccine-misinformation/>. Acesso: 14 dez. 2020.

rizar anúncios com conteúdos que desencorajassem as pessoas a tomar vacinas. Grupos e páginas que disseminam esses conteúdos passaram a ter reduzido seu ranqueamento e foram excluídos da recomendação dos resultados de busca. Anúncios com mensagens enganosas sobre vacinas passaram a ser proibidos.

Quando um usuário tiver sua publicação marcada como falsa por um checador de fatos, o Facebook não oferece um mecanismo de solução de controvérsia. O responsável pelo conteúdo deve buscar diretamente a agência. Se esta aceitar a ponderação e retirar a classificação ao conteúdo, as sanções serão retiradas.

A plataforma passou a adotar a remoção em dois casos. O primeiro é em relação a conteúdos desinformativos que possam causar violência iminente ou dano físico (FACEBOOK, 2019). O segundo envolve mensagens que possam enganar os usuários sobre votações, como em relação a datas, formas, locais, exigências para a participação em eleições e ameaças violentas relacionadas a essa prática. Diante da complexidade da remoção a partir da dificuldade da análise de conteúdo, a possibilidade de dano iminente ou de mais riscos e potencial de dano (especialmente em função da capacidade de disseminação, velocidade e alcance), a retirada pode ser admitida, combinada com regras adequadas de devido processo para que eventuais abusos possam ser coibidos. Caso haja uma reparação pela exclusão da publicação, o dano dessa suspensão temporária é menor do que os problemas potenciais da divulgação da mensagem.

Em janeiro de 2020, por ocasião da declaração da situação de emergência envolvendo o novo coronavírus como uma pandemia, o Facebook anunciou providências relacionadas ao tema. A plataforma passou a remover conteúdos com teorias da conspiração ou afirmações falsas de acordo com as autoridades de saúde que possam gerar danos a pessoas. Isso abrange publicações com tratamentos ou métodos de prevenção falsos, como a ingestão de água sanitária. "Nós estamos fazendo isso como uma extensão de nossas políticas

existentes de remoção de conteúdo que poderia causar danos físicos", explicou o chefe de saúde do Facebook, Kang-Xing Jin[37].

Em fevereiro do mesmo ano, o Facebook anunciou a proibição de anúncios relacionados ao coronavírus, como de produtos apontados por seus comerciantes como envolvidos de alguma forma em tratamentos ou com a cura da Covid-19. Semanas depois, a empresa informou que passaria a banir a propaganda de máscaras médicas utilizadas como equipamento de proteção individual para profissionais de saúde. Em março, novos itens foram adicionados à lista de proibições, como kits de teste, lenços desinfetantes e desinfetantes para mãos (como álcool em gel). Em agosto, a empresa voltou a permitir o anúncio de produtos de higienização das mãos e de lenços de desinfecção de superfícies.

Com a adoção das medidas de distanciamento social e teletrabalho, a força de trabalho das plataformas digitais também foi afetada. No dia 19 de março de 2020, o Facebook anunciou que enviou revisores de conteúdos para casa, com possibilidade de impacto neste procedimento, o que afetaria diretamente a remoção de limitação de circulação de publicações na plataforma. Com menos revisores trabalhando, essa consequência implicou um maior poder dos sistemas automatizados de análise e remoção de conteúdo. Isso traria, afirma texto da companhia deste dia, mais demora no processo e chances maiores de erros. Outra decorrência foi a abolição do procedimento de recurso quando um conteúdo é excluído e o usuário responsável contesta a decisão. As reclamações passaram a servir apenas para avaliação interna, mas a segunda análise deixou de ser feita.

Em abril de 2002, o Facebook anunciou US$ 1 milhão para projetos de checagem de fatos, abarcando agências em diversos países, como Grécia, México, Espanha, Colômbia, Turquia, Índia

37. JIN, Kang-Jing. Keeping People Safe and Informed About the Coronavirus. Facebook. Publicada em: 22 abr. 2020. Disponível em: https://about.fb.com/news/2020/04/coronavirus/. Acesso: 14 dez. 2020.

e Brasil. O aporte foi uma parceria com a Rede Internacional de Checadores de Fatos (IFCN, na sigla em inglês). Entre as ações que os recursos financiarão estão a tradução para novos idiomas, produção de conteúdos multimídia com as checagens, trabalho com especialistas para cobertura baseada em ciência.

Em julho de 2020, a companhia informou a derrubada do grupo "Desmascarando a América" (*Unmasking America*), voltado ao questionamento e ativismo contra o uso de máscaras de proteção facial como método de prevenção contra a Covid-19. A medida foi tomada, segundo o site especializado *The Verge*, após um questionamento deste acerca do grupo[38].

2.1.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Como parte da estratégia sobre desinformação, a plataforma oferece mais informações em relação aos conteúdos verificados como falsos. Em uma publicação, é possível clicar em um ícone "i" para buscar mais informações sobre a fonte. Esses dados são ofertados com base nos artigos do Wikipedia acerca do veículo, além de mais notícias dele e identificação de amigos que compartilharam o link da mensagem que o indivíduo está visualizando. O Facebook introduziu o recurso de mais informações sobre a fonte da mensagem publicada como forma de atestar a credibilidade do responsável por ela. A ferramenta foi lançada nos Estados Unidos. Em maio de 2020, foi estendida para o Reino Unido. Em setembro, para Canadá, Austrália, Irlanda, Nova Zelândia, Brasil, Colômbia e México. Em outubro foi expandida para outros países da Europa (como França e Espanha) e em dezembro para todo o mundo[39].

38. LYONS, K. Facebook suspends anti-mask group for spreading COVID-19 misinformation. The Verge. Publicada em: 20 jul. 2020. Disponível em: https://www.theverge.com/2020/7/20/21331060/anti-mask-groups-facebook-misinformation.

39. HUGHES, Taylor. SMITH, Jeff. LEAVITT, Alex. Helping People Better Assess the

Em julho de 2020, no contexto do novo coronavírus e anunciado como parte das medidas adotadas diante da profusão de desinformação no seu interior sobre o tema, a plataforma começou a disponibilizar nos topos dos feeds dos usuários lembretes com recomendações sobre a pandemia, como a importância de utilizar máscaras faciais para evitar a disseminação do vírus e a contaminação por este.

Como parte de suas reações ao crescimento da desinformação no seu interior, o Facebook passou a financiar iniciativas de educação para a mídia. A plataforma criou uma "Biblioteca de alfabetização digital" voltada a educadores com conteúdos, projetos e sugestão de planos de ensino sobre temas diversos, como segurança, envolvimento da comunidade, privacidade, comportamento positivo, exploração da intimidade, produção de conteúdo e aprendizado de informações[40].

No Brasil, uma delas é o projeto "Vaza, Falsiane" (https://vazafalsiane.com/), promovida por organizações da sociedade civil. Ele envolveu oferta de cursos sobre o tema, incluindo diversos conteúdos, como a definição de notícias falsas, suas características, como são criadas e difundidas, o mercado da desinformação e como combatê-las. Além do curso on-line, que teve início em 2018, o projeto abrangeu também a realização de palestras sobre os assuntos abordados.

Outra iniciativa para o público brasileiro foi a criação do robô (bot) chamado Fátima, criado pela empresa em parceria com a agência de checagem Aos Fatos, que funciona no aplicativo de mensagem FB messenger. O robô fornece dicas de checagem e é possível também incluir um link. Caso ele já tenha sido verificado, o bot responde com a avaliação realizada pela equipe da agência.

Stories They See in News Feed with the Context Button. Publicada em 3 abr. 2018. Disponível em: https://about.fb.com/news/2018/04/news-feed-fyi-more-context/.

40. A biblioteca está disponível em: https://www.facebook.com/safety/educators.

Na Europa, em 2019 eram apoiadas organizações diversas. No âmbito do bloco, houve apoio ao Fórum Jovem Europeu (European Youth Forum) para promover boas práticas entre este segmento da população. Também foram direcionados auxílios a entidades em países específicos, como Zeitverlag e Digibits na Alemanha. Na Polônia, a empresa conduziu palestras em 2019 sobre identificação de mensagens enganosas para alunos de jornalismo nas principais universidades do país[41].

Nos Estados Unidos, também foram direcionados recursos para apoiar entidades. Para as eleições presidenciais de 2020, a empresa anunciou US$ 2 milhões em investimentos para organizações que trabalham com ações de formação de pessoas para ter uma relação mais crítica com os conteúdos que circulam na plataforma.

No âmbito da promoção de informações, o FB implantou um centro de projetos para apoio a veículos jornalísticos (Facebook Journalism Project). Embora ele lide com aspectos gerais relacionados a esse tipo de conteúdo (incluindo monetização de notícias na plataforma), parte das iniciativas visa apoiar experiências jornalísticas com o argumento da ampliação da circulação de notícias profissionais na plataforma como reação à desinformação. Entre as ações estão atividades de capacitação de profissionais e veículos[42], bolsas e outros apoios financeiros a veículos e a organizações que atuam com notícias locais[43].

A empresa criou materiais com dicas para identificar notícias falsas. Em um destes[44], estão elencados como itens a serem consi-

41. FACEBOOK. Q&A on our preparation for EU Elections. Publicado em: 26 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/eu-elections-press-call/.

42. Em dezembro de 2019, o FB lançou em parceria com o Instituto Reuters um curso para identificar mídias manipuladas. Mais informações em: https://www.facebook.com/journalismproject/reuters-course-manipulated-media.

43. Em 2020, o FB anunciou as organizações vencedoras do apoio nos Estados Unidos, que receberam US$ 25 mil. Mais informações em: https://www.facebook.com/journalismproject/programs/community-network/grant-winners-january-2020.

44. O material foi disponibilizado em: https://www.facebook.com/journalismproject/

derados nessa análise o cuidado com as manchetes, especialmente se são apelativos, observar o link e para onde ele leva, investigar a fonte, examinar se há alguma manipulação nas imagens ou vídeos, checar as datas, recorrer a outras fontes sobre o mesmo assunto e perceber intenções na disseminação da mensagem. A companhia informou em entrevista à equipe da pesquisa que a ferramenta educacional foi posicionada no início do *News Feed* durante alguns dias em 14 países. No caso brasileiro, a firma acrescentou que publicou uma campanha sobre o cuidado com notícias falsas em veículos impressos no país em 2018.

Em junho de 2020, a empresa promoveu em parceria com organizações da sociedade uma iniciativa de formação denominada "Três passos para eliminar as notícias falsas". Foram divulgados materiais e mensagens trazendo três perguntas com o objetivo de auxiliar pessoas a não repassar publicações enganosas. Eram elas: 1) de onde vem a notícia?; 2) o que está faltando (incentivando a não observar somente o título)?; e 3) como você se sente (para estimular o cuidado com a manipulação)?[45].

Em abril 2018, o Facebook promoveu uma iniciativa para permitir que centros de pesquisa pudessem desenvolver investigações sobre o impacto das redes sociais em eleições e no debate democrático[46]. Em fevereiro de 2020, a empresa anunciou que aumentou os dados disponibilizados para 60 pesquisadores em 17 laboratórios de pesquisa em 30 universidades de todo o mundo. Neste universo estão 38 milhões de links com informação agregada para ajudar os acadêmicos a saber quem interagiu com esses conteúdos.

Como parte das medidas relacionadas à pandemia do novo coronavírus anunciadas em 2020, foram empregadas providências

programs/third-party-fact-checking/tips-to-spot-false-news.

45. GYEKYE, 2020. https://about.fb.com/news/2020/10/supporting-elections-across-africa/.

46. SCHRAGE, Elliot. Ginsberg, David. Facebook Launches New Initiative to Help Scholars Assess Social Media's Impact on Elections. Publicado em 9 abr. 2018. Disponível em: https://about.fb.com/news/2018/04/new-elections-initiative/.

de oferta de mais informações. Quando pessoas faziam buscas por termos associados ao vírus, eram apresentados como resultados informações educacionais de fontes de autoridades de saúde. A empresa forneceu créditos de publicidade a autoridades de saúde e organizações atuando no combate à pandemia em regiões afetadas. Em fevereiro de 2020, a medida anunciada de apresentação de informações de fontes oficiais sobre o vírus e direcionamento ao site da OMS foi estendida para a plataforma em todo o planeta.

Em março de 2020, o Facebook anunciou a expansão das notificações de novidades para mais autoridades locais. Inicialmente adotada para a OMS e depois para órgãos nacionais, como ministérios da saúde de cada país, a providência visou ramificar os informes para autoridades estaduais e municipais. No dia 16 daquele mês, as principais empresas de tecnologia dos Estados Unidos emitiram um curto comunicado conjunto acerca de suas iniciativas em relação à pandemia[47]. "Estamos juntos nos esforços de resposta à Covid-19. Estamos ajudando milhões de pessoas a ficarem conectadas enquanto também combatendo a fraude e desinformação sobre o vírus, elevando o conteúdo de autoridades e compartilhando atualizações críticas de forma coordenada com órgãos governamentais de saúde pelo mundo. Convidamos outras companhias a se juntarem a nós enquanto trabalhamos para manter nossas comunidades saudáveis e seguras", indicou a nota.

No mesmo mês, o Facebook lançou o Centro de Informações sobre Coronavírus. Este foi colocado disponível no alto do *News Feed* e suas atualizações são destacadas aos usuários com notificações. Nele foram publicadas atualizações de órgãos oficiais de saúde, como a OMS, e artigos sobre medidas recomendadas, como o distanciamento social. Ao seguir o centro, a pessoa recebe as

47. MICROSOFT. A joint industry statement on COVID-19 from Microsoft, Facebook, Google, LinkedIn, Reddit, Twitter and YouTube. Publicação de Facebook. Publicada em 17 mar. 2020. Disponível em: https://m.facebook.com/Microsoft/photos/a.195193573720/10157344325923721/?type=3&source=48&__tn__=EH-R

mensagens diretamente no *feed*. Nos Estados Unidos, para além do Centro de Informações, os usuários podiam ser conectados com grupos que estão atuando para combater a pandemia e oferecer ajuda a outros cidadãos de suas comunidades. Na Europa, o centro foi implementado em cidades com alta taxa de incidência, como Itália, Espanha, França, Reino Unido e Alemanha.

Em abril de 2020, o Facebook publicou uma atualização das medidas adotadas para combater a desinformação na plataforma relacionada à pandemia do novo coronavírus. Entre elas estavam: 1) direcionamento a mais de 2 bilhões de pessoas a informações da OMS e outras autoridades de saúde no Facebook e Instagram. Destas, 350 milhões clicaram para obter mais informações; 2) trabalho com mais de 60 agências de checagem em mais de 50 idiomas em todo o mundo. Entre março e abril foram adicionados oito novos parceiros e atuação em 12 países; 3) avisos associados a conteúdos verificados; foram 40 milhões de posts em relação a quatro mil conteúdos em março; 4) remoção de "centenas de milhares" de conteúdos de desinformação "que poderiam ter levado a danos físicos no mundo real". Exemplo: afirmação de que beber água sanitária seria eficiente na eliminação do vírus do corpo e que o distanciamento social não auxilia no combate à pandemia.

Além disso, foram anunciadas novas ações: exibição de mensagens para quem interagiu com conteúdos "falsos nocivos" sobre a Covid-19 que foram removidos. As publicações direcionam à página da OMS dedicada a desmentir mitos sobre a pandemia; adição da seção "conheça os fatos" na Central de Informações sobre a Covid-19, onde são disponibilizadas checagens de fatos realizadas por parceiros. Os artigos são escolhidos pelo time editorial da empresa.

Em entrevista à equipe de pesquisa, a companhia afirmou que emitiu 40 mil alertas sobre conteúdos falsos sobre Covid-19, relacionados a aproximadamente 4 mil artigos de checadores externos. Em abril de 2020 foram marcados 50 milhões de conteúdos sobre

a pandemia, relacionados a cerca de 7.500 artigos dos agentes responsáveis pela verificação de conteúdo. Quando pessoas veem o rótulo, em geral 95% delas não visualiza a mensagem original.

Também em entrevista à equipe de pesquisa, o Facebook afirmou não ter definido ainda quais medidas adotadas ao longo da pandemia serão mantidas. "A pandemia continua evoluindo. E à medida que isso ocorre, vamos olhando para a plataforma, avaliando tendências sobre conteúdo e dialogando com especialistas, como os da Organização Mundial de Saúde, autoridades nacionais de saúde e atores no espectro dos usuários de nossos serviços e vamos prover políticas adicionais quando apropriado para manter os membros da comunidade seguros durante a crise".

2.1.4 MEDIDAS CORRELATAS

A empresa adota estratégias sobre conteúdos que, embora não mirem diretamente a desinformação, têm consequência sobre ela. "Alguns conteúdos ou comportamentos [proibidos] são geralmente relacionados com desinformação (como spam, discurso de ódio, comportamento inautêntico, sub-representação e contas se passando por outras), então quando age sobre essas violações o Facebook está substancialmente contribuindo para combater a difusão de desinformação" (FACEBOOK, 2020).

A primeira é o discurso de ódio, elencado como um dos tipos de conteúdos passíveis de remoção segundo as diretrizes da comunidade. Em abril 2018, estas diretrizes foram publicadas pela primeira vez. A empresa passou a permitir o direito de apelação a pessoas objeto de sanções, como a remoção de publicações[48]. O procedimento foi desenvolvido com o usuário solicitando revisão

48. BICKERT, Monica. Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process. Publicado em: 24 abr. 2018. Disponível em: https://about.fb.com/news/2018/04/comprehensive-community-standards/

adicional para contestar a primeira análise realizada. Esta avaliação é realizada por uma pessoa em 24h. Em caso de concordância com o recurso, as sanções são revertidas. Em algumas situações, a companhia desenvolveu uma ação específica focada. Foi o caso Myanmar em 2018, onde a desinformação espalhada na plataforma alimentou disputas violentas entre grupos étnicos. Entre estas ações proativas estão a análise e remoção de conteúdos de discurso de ódio.

Em março de 2019, o atentado a pessoas em uma mesquita na Nova Zelândia no dia 15 e a circulação de vídeos mostrando o episódio e com conteúdos sobre o caso gerou diversos questionamentos acerca do Facebook e suas plataformas, entre elas o Instagram, como espaço de difusão de discurso de ódio. O ocorrido gerou novas pressões sobre a companhia e suas medidas para impedir este tipo de prática. Em resposta, a diretora operacional da companhia, Sheryl Sandberg, afirmou que as contas dos responsáveis pela transmissão foram derrubadas nas duas redes sociais e que os sistemas automatizados foram utilizados para impedir a publicação de vídeos relacionados[49]. Foram anunciadas restrições para o uso das transmissões ao vivo (*lives*) no Facebook, como a consideração de violações prévias às diretrizes da comunidade. Outra medida anunciada foi a "melhoria" dos sistemas automatizados para identificar vídeos violentos, mesmo quando editados, e removê-los. Ela informou também novos esforços para remover discurso de ódio das plataformas, inclusive banindo grupos naquele país que disseminam este tipo de conteúdo.

Em março de 2020, o Facebook anunciou o banimento de ações de estímulo e defesa do nacionalismo e separatismo branco. Esses grupos e suas mensagens na plataforma foram objeto desta ação pela compreensão de que operam por meio da promoção do

49. SANDBERG, Sheryl. By working together we can win against hate. Publicado em: 29 mar. 2019. Disponível em: <https://about.instagram.com/blog/announcements/sheryl-sandbergs-statement-on-new-zealand-shooting/>. Acesso: 14 dez. 2020.

discurso de ódio, proibido pelas diretrizes da comunidade. Buscas relacionadas à supremacia branca, como exaltação a nazismo, passaram a receber indicações de materiais de combate ao ódio. Em julho de 2020, a companhia anunciou a derrubada de três páginas de supremacistas brancos a partir de uma investigação da emissora Al Jazeera que apontou 120 páginas com essa finalidade que acumulavam 800 likes. As outras 117 foram deixadas no ar[50].

Em agosto de 2020, foi anunciado que não seriam mais autorizadas contas, páginas e conteúdos relacionados ao movimento que ficou conhecido como QAnon, que difunde teorias da conspiração. Em outubro, a decisão foi expandida para a exclusão total do movimento[51] e a companhia declarou que passaria a proibir conteúdos de negação do holocausto. A iniciativa foi apresentada como parte da política de combate ao discurso de ódio, mas que também tem efeito concreto sobre uma narrativa desinformativa acerca deste importante episódio histórico e da negação por grupos supremacistas de genocídio promovido por nazistas.

Além de discurso de ódio, outra categoria proibida nos Padrões da Comunidade que possui relação com desinformação é "fraude e dolo", segundo a qual é vedado "conteúdo destinado a enganar pessoas com o objetivo de obter vantagem injusta ou privar de dinheiro, propriedade ou direito legal"[52]. Fica impedida, por exemplo, a publicação de oferta, solicitação e comercialização de avaliações de usuários falsos, documentos falsos ou manipulados (como cupons ou prescrições médicas). Também não são autorizadas

50. MAACK, M.M. Facebook removes 3 white supremacist band pages — leaves (at least) 117 others. TNW. Publicada em: 10 jul. 2020. Disponível em: https://thenextweb.com/facebook/2020/07/10/facebook-removes-3-white-supremacist-band-pages-leaves-at-least-117-others/https:/www.aljazeera.com/news/2020/07/exclusive-facebook-extensively-spread-neo-nazi-music-200710075507831.html/.

51. COLLINS, Ben. ZADROZNY, Brandy. Facebook ban Qanon across its platforms. NBC News. Publicada em: 6 out. 2020. Disponível em:< https://www.nbcnews.com/news/amp/ncna1242339>. Acesso em 8 dezembro 2020.

52. Disponível em: https://www.facebook.com/communitystandards/fraud_deception.

práticas como criação de empresas falsas, esquemas ponzi ou investimentos com promessas de altas taxas de retorno.

A segunda estratégia envolve a derrubada de contas falsas ou daquilo que a empresa chama de "comportamento não autêntico coordenado". O termo é tomado como quando um "grupo de pessoas ou páginas atuam conjuntamente para enganar outras sobre quem são ou o que estão fazendo"[53]. Atores que adotam essa prática recorrentemente podem ser considerados "personas non gratas" e serem removidos da plataforma permanentemente. Alguns tipos de comportamentos são objeto de atenção pela equipe responsável. É o caso das interferências externas em debates ou processos eleitorais ou ataques de governos sobre indivíduos do seu país. Já o comportamento não autêntico envolve um rol de atitudes, como usar várias contas, empregar recursos da plataforma para enganar pessoas sobre questões como a identidade, objetivo de um público ou comunidade ou fonte de um conteúdo.

Em texto publicado em maio de 2020, o vice-presidente de análise do Facebook, Alex Schultz, explicou que a derrubada de contas falsas foca em perfis que possam gerar dano aos usuários[54]. Um dos métodos utilizados é dificultar o processo de abertura, buscando verificar a autenticidade neste momento. São empregadas táticas de análise como identificar se há criação de contas em massa em uma determinada localidade, ou bloquear um endereço IP. Para aquelas contas criadas, os sistemas automatizados mapeiam "sinais" de condutas "maliciosas" verificadas em outras contas falsas.

Em julho de 2018, o Facebook anunciou a derrubada de 196 páginas e 87 contas no Brasil por violação das políticas de auten-

53. GLEISCHER, Nathaniel. Coordinates inauthentic behaviour explained. Publicado em: 6 dez. 2018. Disponível em: https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/.

54. SCHULTZ, Alex. How Does Facebook Measure Fake Accounts?. Publicado em 23 mai. 2019. Disponível em: <https://about.fb.com/news/2019/05/fake-accounts/>. Acesso: 14 dez. 2020.

ticidade. A empresa explicou que essas páginas faziam parte de uma rede que operava para disseminar desinformação. Ao longo do ano, a empresa anunciou novas retiradas de páginas, como algumas ligadas à agência russa acusada de intervenção nas eleições de 2016. Em março de 2019, novas redes de contas foram derrubadas, no Reino Unido e na Romênia[55].

Essas medidas são defendidas por seu impacto indireto em práticas de desinformação embora tenham como foco contas ou outros tipos de conteúdo que violam as diretrizes da comunidade. "Embora notícias falsas não violem nossos Padrões da Comunidade, normalmente viola nossas políticas em outras categorias, como spam, discurso de ódio e contas falsas, o que nós removemos", explicou a gerente de produto Tessa Lyons em texto publicado no blog da companhia em maio de 2018[56].

Além disso, o Facebook reduz a circulação de contas identificadas como difusoras de conteúdo de baixa qualidade. Isso ocorre por diversas medidas, como a diminuição dos incentivos financeiros para esquemas de *spammers* que difundem esses conteúdos para gerar cliques e auferir renda da visita a sites. Assim, a plataforma reduz a circulação de conteúdos "caça-clique", de links geralmente distribuído por *spammers* e do que chama de "sites de baixa qualidade" (definidos como aqueles com pouco conteúdo substantivo e com "anúncios maliciosos, chocantes e disruptivos").

Em 2018, veículos do Reino Unido e dos Estados Unidos noticiaram que a consultoria política Cambridge Analytica obteve dados pelo Facebook e os utilizou para manipular eleições e processos políticos nos EUA, Nigéria e Reino Unido, entre outros.

55. GLEICHER, Nathaniel. Removing Coordinated Inauthentic Behavior From the UK and Romania. Publicado em: 7 mar. 2019. Disponível em: https://about.fb.com/news/2019/03/removing-cib-uk-and-romania/.

56. LYONS, Tessa. Hard Questions: What's Facebook's Strategy for Stopping False News?. Publicado em: 23 mai. 2018. Disponível em: https://about.fb.com/news/2018/05/hard-questions-false-news/

O escândalo colocou a plataforma como alvo de críticas e de investigações por autoridades regulatórias de diversos países, dos Estados Unidos ao Brasil. Em post no dia 21 de março daquele ano, o diretor executivo da empresa, Mark Zuckerberg, apresentou medidas, entre as quais a exclusão das contas da Cambridge Analytica, uma auditoria nos apps rodando na plataforma e novos mecanismos para que os usuários pudessem controlar melhor as autorizações a aplicações de terceiros. Embora essa decisão não tenha sido vinculada diretamente à desinformação, ao restringir o acesso à coleta de dados por apps de terceiros o Facebook objetivou, ao menos em seu discurso, evitar novos episódios de promoção de conteúdos falsos ou manipulação de eleições como o da Cambridge Analytica. Em agosto de 2018, a empresa anunciou que a partir da auditoria de apps mais de 400 foram suspensos.

Por ocasião da investigação do Congresso dos EUA sobre interferência de agentes externos nas eleições de 2016, em maio a empresa divulgou as ações adotadas para impedir que atores maliciosos usassem desinformação visando minar o processo democrático. Foi criado um arquivo de anúncios com informações sobre pagamento e alcance, disponibilizado globalmente em junho daquele ano. Dos anunciantes passou a ser exigida confirmação de residência para promover publicidade política nos EUA. Os critérios de segmentação foram revistos, retirando parte daqueles utilizados pela agência russa identificada como o principal agente tendo atuado no pleito de 2016. Foi aumentada a capacidade dos sistemas tecnológicos para bloquear e remover contas falsas. Foram derrubadas 70 contas no Facebook e 65 no Instagram, bem como 163 páginas, da agência russa IRA[57].

Na linha que chama de "proteção à integridade de eleições", a empresa promove outras ações, entre elas estão medidas de transparência de anúncios políticos. Em 2018 foram implementados

57. FACEBOOK. Russian ads released by congress. Publicado em: 10 mai. 2018. Disponível em: https://about.fb.com/news/2018/05/russian-ads-released-by-congress/

recursos como arquivos de anúncios (identificando responsáveis e alcance da propaganda), mais informações sobre a página e rótulos dessas peças quando aparecem no *News Feed*. Houve medidas para diversos países, como México e Nigéria[58].

Parte destas estratégias foi adotada no Brasil. Em uma conferência divulgada pelo site do Facebook em 19 de setembro de 2018, a diretora de políticas públicas do Facebook para o país, Monica Rosina, apresentou as ações da plataforma voltadas às eleições de 2018 no Brasil. Segundo ela, a equipe estava atuando para combater o comportamento abusivo e removendo atores maus. Ela citou como exemplo a derrubada de 196 páginas e 87 contas por comportamento não autêntico. Em agosto de 2018, uma nova ação retirou 74 grupos, 57 contas e cinco páginas. Esses atores foram identificados pelo laboratório forense do Conselho Atlântico. Outra estratégia é a verificação de conteúdos por agências de checagem com as correspondentes sanções adotadas pela plataforma. Foram implementadas também iniciativas como bots para dar dicas sobre identificação de notícias falsas e ações educativas e cursos sobre o tema.

A segunda frente de atuação, continuou Rosina, esteve relacionada à transparência em anúncios políticos, que foram permitidos pela primeira vez para aquele pleito. Foram adotados mecanismos como o arquivo de anúncios, a rotulagem dessas peças e informações sobre essas mensagens (como os financiadores delas). Em nova publicação no dia 22 de outubro, o Facebook relatou a retirada de 68 páginas e 43 contas vinculadas à empresa Raposo Fernandes Associados (RFA), que criaram páginas utilizando contas

58. Em texto no blog da empresa publicado em janeiro, a diretora de política global, Katie Harbath, anunciou medidas para integridade das eleições em 2019. Na Nigéria, foi proibido o pagamento de anúncios políticos por agentes de fora do país. Anunciantes passariam a precisar ser aprovados para poder rodar um anúncio. A biblioteca de anúncios disponibilizaria informação sobre quem pagou e quem foi atingido (com as características demográficas, como idade, gênero e localização).

falsas e contas múltiplas usando os mesmos nomes. Essas páginas difundiam conteúdos que visavam captar usuários para sites que foram classificadas como "fazendas de anúncios".

Em junho de 2020, o Facebook anunciou novas medidas para as eleições municipais deste ano no Brasil. Os anunciantes passaram a ter de passar por um processo mais rígido de identificação, com comprovação de identidade, residência no Brasil e fornecimento de e-mail, telefone e site. Foi reforçada a política de marcação das mensagens publicitárias como "propaganda eleitoral". Além disso também foram atualizados mecanismos de inclusão nas campanhas de informações sobre os anunciantes e responsáveis pela campanha. Os anúncios sobre temas sociais de políticos não identificados são passíveis de não aprovação.

Nos Estados Unidos, nas eleições parlamentares de 2019 (*midterms elections*), uma das ações foi o combate à desinformação sobre o ato de votar, uma vez que este não é obrigatório no país. De acordo com a firma, foram identificados 45 mil conteúdos associados ao desincentivo à participação nas eleições. Deste total, 90% foram proativamente mapeados pela empresa e não por usuários ou agências de checagem[59]. Para as eleições de 2020[60], a firma anunciou diversas providências. Os rótulos de publicações checadas foram tornados "mais claros". No plano da segurança, as contas de candidatos ganharam proteções extras para evitar invasões, como autenticação em dois fatores e um monitoramento para evitar acessos indevidos de locais estranhos. Nos anúncios, os responsáveis pela difusão deste tipo de mensagem tiveram de fornecer mais informações e serem verificados após este procedimento. Um rastreador dos gastos dos candidatos a presidente foi

59. FACEBOOK. Q&A on our preparation for EU Elections. Publicado em: 26 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/eu-elections-press-call/.
60. ROSEN, Guy. Harbath, Katie. Gleicher, Nathaniel. Leathern, Rob. Helping to protect the 2020 US elections. Publicada em: 21 out. 2019. Disponível em: https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/.

criado. Foram banidas mensagens pagas que preguem contra o ato de votar. Em junho, foi implementada uma opção aos estadunidenses de não receber anúncios políticos[61].

Nas páginas, passou a funcionar a identificação dos responsáveis e a confirmação pelo FB, a exemplo de contas no Twitter. Também serão exibidas informações sobre o local do responsável e se foi fundida com outras páginas. Essas ferramentas valerão apenas para as páginas com muitos seguidores. As páginas de veículos e instituições estatais passarão a ser submetidas a um nível maior de transparência. Serão incluídos, por exemplo, rótulos marcando as páginas de veículos controlados em todo ou em parte por esses entes. Essas identificações serão aplicadas também no Instagram.

Nas eleições presidenciais dos Estados Unidos de 2020, o vice-presidente de Integridade da companhia, Guy Rosen, afirmou que, entre março e setembro, a plataforma removeu 120 mil conteúdos do Facebook e do Instagram, mostrou alertas em 150 milhões de publicações, rejeitou 2,2 milhões de solicitações de anúncios e derrubou 30 redes envolvidas em comportamento inautêntico coordenado, além de ter excluído 6,5 bilhões de contas falsas[62].

Em seu texto no site oficial da empresa, Rosen declarou que foram adotadas novas medidas para o pleito estadunidense. Um sistema de monitoramento de conteúdos virais, independentemente da violação de condutas já vedadas, foi implementado. Foram acrescentadas às hipóteses de remoção de conteúdos mensagens de intimidação ou constrangimento de integrantes da Justiça Eleitoral ou sobre votantes.

61. HATMAKER, T. Facebook adds options for US users to turn off political ads, launches voting info hub. Tech Crunch. Publicada em: 17. Jun. 2020. Dissponível em: https://techcrunch.com/2020/06/16/facebook-turn-off-political-ads-voting-info-center/.

62. ROSEN, Guy. Preparing for election Day. Publicada em 7 out. 2020. Disponível em: < https://about.fb.com/news/2020/10/preparing-for-election-day/>. Acesso em 8 dezembro 2020.

Em setembro de 2020, o Facebook anunciou que iria suspender os anúncios uma semana antes das eleições nos Estados Unidos. Segundo o diretor presidente da companhia, Mark Zuckerberg, a decisão foi tomada porque este prazo exíguo antes das votações dificultam que os conteúdos de anúncios possam ser verificados e contestados. No mesmo anúncio, o líder da empresa informou que o Facebook Messenger limitará o encaminhamento de mensagens a cinco contatos, seguindo mudança adotada pelo WhatsApp.

Nas eleições de 2020 em Myanmar, o Facebook promoveu um seminário com 13 veículos de mídia para apresentar conteúdos que auxiliam "boas práticas" na atuação no pleito. Um programa de TV foi veiculado a cada mês sobre educação para a mídia, chamado "Tea Talks".

Na Europa, a plataforma empregou estratégias tanto para as eleições do bloco quanto específicas em alguns países. No primeiro caso, um exemplo foi o pleito do Parlamento Europeu de maio de 2019. Foram firmadas parcerias com 21 agências de checagem em 14 idiomas. Entre elas um consórcio denominado FactCheckEU. No bloco, são 28 países e 24 idiomas falados. Para lidar com essa diversidade, quando não possui agências de determinados idiomas a empresa depende de seus sistemas de tradução.

Para evitar interferências externas, os anunciantes passaram a ter que obter autorização e fornecer informações de identificação. Assim como em outras ocasiões, foram implementados mecanismos de transparência de anúncios, como a biblioteca. Estas medidas foram adotadas não somente para anúncios de candidatos, mas relacionados a temáticas objeto de grande debate[63].

Ferramentas de identificação do indivíduo e local de residência também foram definidas para administradores de páginas com muitos seguidores. Outra frente de ação foi o combate à conteúdos associados ao que a plataforma chama de "supressão da votação",

63. GEISEL, Anika. Protecting the european parliament elections. Publicada em: 28 jan. 2019. Disponível em: https://about.fb.com/news/2019/01/european-parliament-elections/.

como publicações com datas e locais errados ou incentivando a não participação no pleito. Também foi instalado um centro de operações em Dublin, na Irlanda. Segundo o Facebook, essa estrutura permite uma coordenação em tempo real com as equipes da sede da empresa, nos EUA[64].

Nas eleições gerais de 2019 no Reino Unido, foram aplicadas as regras para anúncios (como rotulagem destes, criação das bibliotecas de anúncio, identificação dos responsáveis pelos anúncios e se moram no país, mais informações sobre o porquê de estar recebendo o anúncio e a alternativa de bloquear anunciantes). Foi instalado um "centro de operações" com trabalhadores de diferentes equipes da empresa. Uma rede de 137 contas do Facebook e do Instagram foi removida em março por enquadramento em comportamento inautêntico coordenado[65].

No primeiro semestre de 2020, segundo o Facebook, ferramentas de transparência de anúncios estavam presentes em 194 países. Também neste período, eram exigidas autorizações específicas em 62 países, incluindo o Brasil[66]. Entre essas as de identificação deste tipo de conteúdo e dos anunciantes, bem como recursos como a exibição de rótulos que apresentem os conteúdos como anúncios aos usuários.

2.2. INSTAGRAM

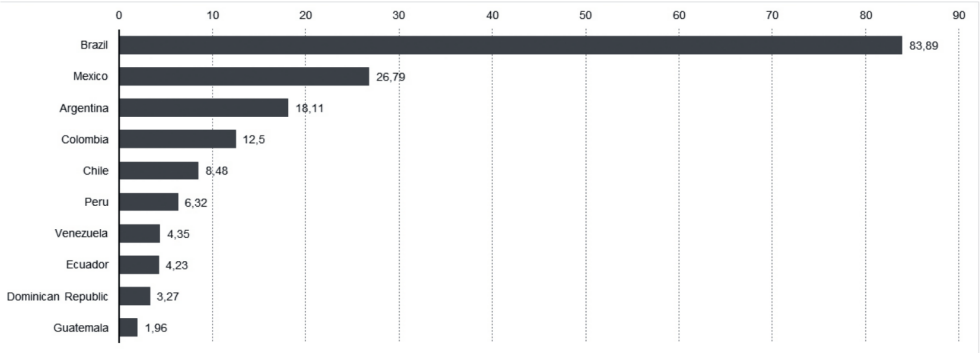
Com o Facebook tendo estabilizado sua expansão, o Instagram assumiu a condição de rede social da última década. Criado em

64. FACEBOOK. Q&A on our preparation for EU Elections. Publicado em: 26 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/eu-elections-press-call/.
65. FACEBOOK. How Facebook Has Prepared for the 2019 UK General Election. Publicado em: 7 nov. 2019. Disponível em: https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/.
66. A listagem dos países pode ser encontrada aqui: https://www.facebook.com/business/help/2150157295276323.

2010, seu potencial foi logo percebido pelo Facebook, que adquiriu a startup por US$ 1 bilhão em 2012. Em 2014, já eram 200 milhões de usuários, número que cresceu para 400 milhões em 2015, 600 milhões em 2016, 800 milhões em 2017 e chegou a 1 bilhão em 2018, como mostra Felix Richter da consultoria Statista. Em 2020, o Instagram ocupa o posto de quarta maior rede social do planeta, ficando atrás apenas de FB Messenger, WhatsApp e Facebook.

Segundo o relatório *O estado do mundo móvel 2020*, em que a consultoria App Annie analisa dados relativos a 2019, o Instagram é o quarto app com mais usuários ativos mensalmente no Brasil, perdendo apenas para FB Messenger (terceiro), Facebook (segundo) e WhatsApp (primeiro). No ranking por downloads, o Instagram também fica na quarta colocação. O relatório *Digital in 2020*, da WeAreSocial e Hootsuit, lista o Instagram como a terceira rede social mais popular, utilizada por 79% dos usuários de Internet entre 16 e 64 anos.

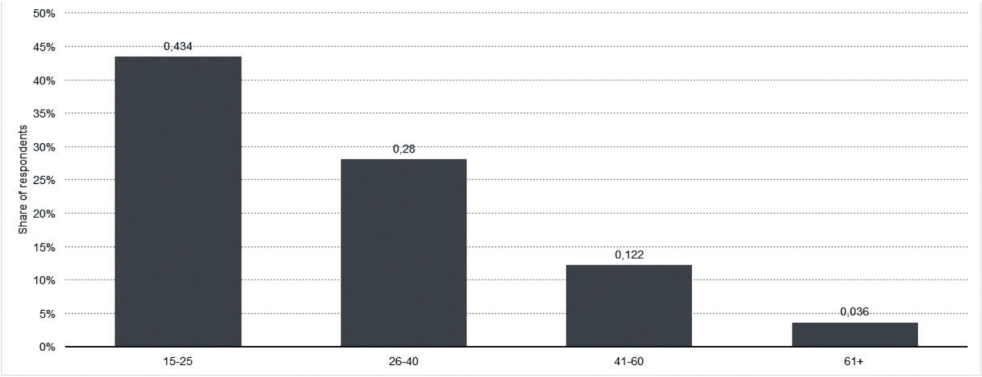
Conforme a pesquisa da consultoria Statista Social Media Usage in Latin América (STATISTA, 2020), o Brasil é o 3º mercado mundial do Instagram e o 1º na América Latina.



Fonte: Statista (2020)

No recorte por gênero, as proporções são parecidas, enquanto no por idade fica clara a prevalência dos mais jovens no público da rede social.

Fonte: Statista (2020)



Fonte: Statista (2020)

2.2.1 ABORDAGEM DO FENÔMENO

Como rede social de propriedade do Facebook, no Instagram foi implantada abordagem semelhante à da plataforma de mesmo nome da companhia, discutida no item anterior. Assim, as medidas foram erigidas sobre o modelo da identificação a partir de parcerias com agências de checagem de fatos, rotulagem das publicações classificadas como tal, adoção de medidas de moderação para os conteúdos considerados desinformativos e oferta de informações adicionais ou "contexto", como a empresa chama.

As políticas da rede social são espelhadas naquelas do serviço original de seu controlador. “Em geral, tanto o Facebook quanto o Instagram compartilham as mesmas políticas, significando que se um conteúdo é considerado violador no Facebook também o é no Instagram”, explicou a companhia em entrevista à equipe da pesquisa.

Embora o modelo seja o mesmo, as regras e diretrizes foram implementadas depois no Instagram. Se no próprio Facebook já vem havendo uma dinâmica de iniciar em determinados países e depois expandir a mudança para o restante das nações onde a plataforma está presente, no caso do Instagram as providências foram adotadas sempre após, com sua generalização também ocorrendo depois de já ter sido incorporada no Facebook no restante do mundo.

No Instagram, há também medidas correlatas com rebatimento no tratamento da desinformação no interior da rede social. Embora as normas internas do Instagram não mencionem explicitamente conteúdos falsos, os Termos de Uso (revisados pela última vez no dia 19 de abril de 2018) determinam que o usuário não pode “fazer algo ilícito, enganoso, fraudulento ou com finalidade ilegal ou não autorizada”.

Em geral, as estratégias relacionadas a conteúdos estiveram mais focadas no combate ao assédio e a outras práticas agressivas e de ódio entre os usuários. Mas as diretrizes da comunidade da aplicação também elencam um conjunto de conteúdos e comportamentos mais ou menos diretamente relacionados à desinformação. Também têm impacto sobre a difusão de notícias falsas as políticas de bloqueio e remoção de contas, assim como no Facebook.

Como aplicação controlada pelo Facebook, o Instagram não possui estrutura institucional e processos independentes ou específicos. Assim, valem para a rede social as instâncias, fluxos e procedimentos adotados pelo Facebook, incluindo as equipes e

processos de revisão de conteúdo e o Conselho de Supervisão criado recentemente (ver item 2.1).

2.2.2 MODERAÇÃO DE CONTEÚDO

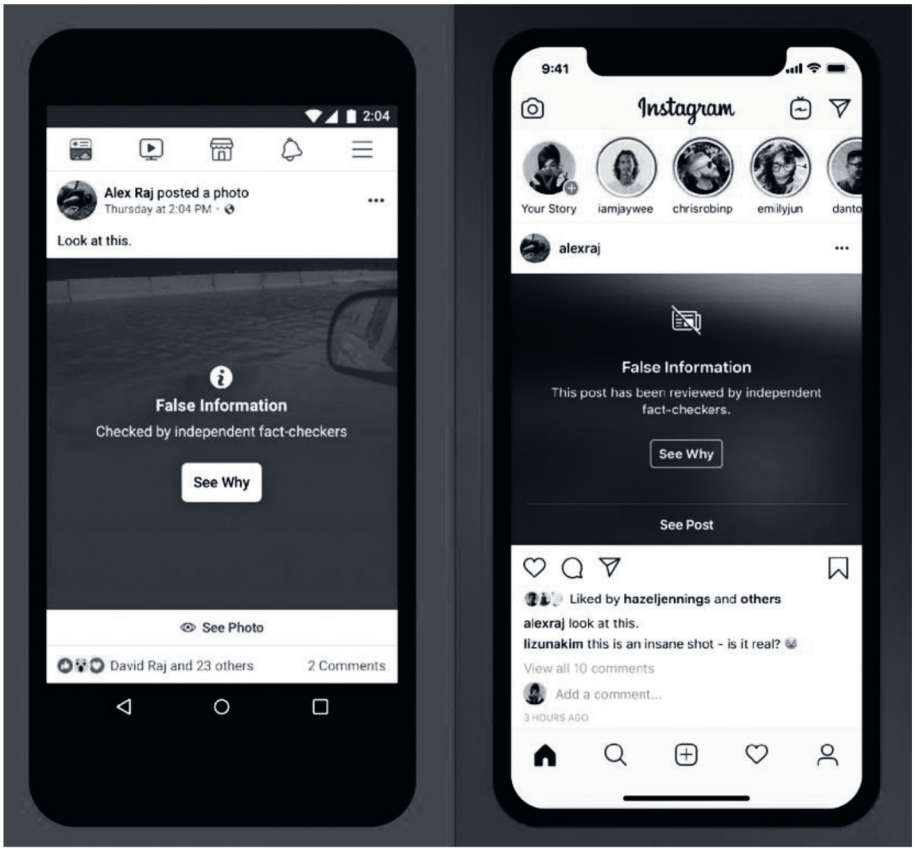
Como ocorre no Facebook, há um processo de análise de conteúdos a partir de agências de verificação de fatos, que classificam uma publicação em diferentes categorias. Os critérios e procedimentos são os mesmos empregados pelo Facebook. Em cada publicação, há a opção de o usuário apontá-la como "informação falsa", também tomada como um "sinal" considerado na análise geral procedida sobre determinado post para avaliar as medidas correspondentes. Este recurso foi implementado no início de 2019. Em fevereiro, por ocasião do Dia da Internet Segura, a companhia divulgou os recursos disponíveis para evitar diversas práticas consideradas nocivas. Entre esses estão a possibilidade de denúncia de comentários, o bloqueio de contas, controle dos comentários em *posts*.

A caracterização do caráter falso ou parcialmente falso de um post é feito por agências de checagem. No início de 2020, eram 45 em todo o mundo, conforme site oficial da companhia. Os critérios são os mesmos adotados para o Facebook. Em maio de 2019, o Instagram começou a trabalhar com agências de checagem de fatos para análise dos conteúdos de acordo com os parâmetros firmados no Facebook em 2018. A iniciativa foi ativada inicialmente nos Estados Unidos. Em 19 de dezembro, a empresa anunciou que o trabalho com agências de checagem de fatos iniciado em maio foi expandido para todo o mundo.

Os conteúdos são marcados quando são verificados como desinformativos (ver item seguinte), condição a partir da qual são promovidas as ações de moderação. Há diferenciação na execução dessas medidas. O Instagram também "reduz a circulação" dos

posts considerados falsos ou parcialmente falsos pelos verificadores. Isso significa uma menor visibilidade no feed e nos stories. Essa providência abarca também a retirada do objeto da marcação das páginas de palavras-chave (hashtags) e da página "explorar". As hashtags ganharam prevalência no Instagram, tornando-se inclusive uma referência a qual pode ser "seguida" pelo usuário, não somente acessada momentaneamente.

Os usuários com práticas reiteradas de divulgação de mensagens enganosas verificadas são objeto de uma redução maior da circulação. Quem tem um post rotulado como falso pode apelar da classificação. Uma ferramenta no perfil do usuário permite que ele envie um e-mail contestando a avaliação e buscando explicar porque considera que o conteúdo não é falso. Contudo, não é possível recorrer de publicações excluídas para os verificadores de fatos.



Dentro da política do Facebook de promoção do que chama de integridade das eleições, iniciativas foram promovidas também para o Instagram. Nos Estados Unidos, para as eleições de 2020, os rótulos foram alterados nesse sentido, colocando um filtro cinza que diferencia a imagem das demais quando o conteúdo é verificado como falso.

No Instagram, passaram a ser exibidos avisos antes de um usuário compartilhar uma informação apontando que ela foi verificada e qual foi a classificação pelas agências checadoras de fatos. A medida veio ao encontro das preocupações anunciadas não apenas de identificação para o consumo da informação pelo usuário, mas no reforço de imprimir alertas no momento do compartilhamento. Para as eleições do Reino Unido de 2019, em medidas conjuntas para FB e Instagram foram implementadas novas ferramentas de transparência para anúncios e criado um centro de operações. Uma rede de 137 contas do Facebook e do Instagram foi removida em março por enquadramento em comportamento inautêntico coordenado[67].

Em março de 2019, como parte das ações do Facebook contra desinformação sobre vacinas, passaram a ser retirados das páginas de hashtags e da "explorar" posts com informações enganosas sobre o método de tratamento. Em janeiro de 2020, o Facebook anunciou que a companhia passaria a retirar conteúdos relacionados a falsos tratamentos ou métodos de prevenção em relação à Covid-19. No caso do Instagram, também foram incluídos entre os esforços a exclusão de hashtags que pudessem promover desinformação nesse sentido[68]. Quando usuários buscam por hashtags

67. FACEBOOK. How Facebook Has Prepared for the 2019 UK General Election. Publicado em: 7 nov. 2019. Disponível em: https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/.

68. JIN, Kang-Jing. Keeping People Safe and Informed About the Coronavirus. Facebook. Publicada em: 22 abr. 2020. Disponível em: https://about.fb.com/news/2020/04/coronavirus/

associadas ao vírus serão apresentados como resultados de busca informações de autoridades de saúde. Assim como no Facebook, autoridades e organizações atuando para conter a pandemia receberam créditos para promover campanhas publicitárias.

Em março do mesmo ano, o Instagram fez um balanço das ações adotadas até aquele momento em relação à desinformação e à pandemia[69]. Foram banidos anúncios enganosos ou de produtos relacionados à Covid-19 que cumprissem determinados critérios, como "a intenção de criar urgência, garantir cura ou impedir as pessoas de contraí-la". Também foram proibidas propagandas de determinados produtos, como máscaras faciais.

A plataforma passou a remover contas associadas à Covid-19 das recomendações de contas e conteúdos associados ao tema da página "explorar", à exceção daqueles de autoridade oficiais de saúde. Além disso, conteúdos verificados como falso por checadores de fatos parceiros da empresa passaram a ter sua classificação rebaixada para o feed e nos stories. "Se as publicações forem classificadas como falsas por verificadores de fatos terceirizados, nós as removeremos das páginas de hashtag e da página Explorar", acrescentou a companhia[70]. Assim como na diretriz adotada pelo Facebook, teorias da conspiração ou falsas alegações registradas como danosas pelas autoridades de saúde passaram a ser removidas da plataforma.

2.2.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Como suas políticas são as adotadas pela empresa controladora,

69. INSTAGRAM. Fornecer informação, segurança e apoio às pessoas no Instagram. Publicado em: 24 mar. 2020. Disponível em: https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram

70. INSTAGRAM. Fornecer informação, segurança e apoio às pessoas no Instagram. Publicado em: 24 mar. 2020. Disponível em: https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram.

o Facebook, e espelham aquelas implementadas na rede social de mesmo nome, os projetos são conduzidos por esta. Assim, as políticas de promoção de informações e transparência devem ser consideradas entre as desenvolvidas pela companhia (ver item 2.3).

Assim como no Facebook, o conteúdo é "marcado" ou "rotulado" como forma de chamar a atenção dos usuários para a checagem sobre ele. Segundo a rede social, "para que as pessoas possam tomar decisões mais embasadas sobre o que ler, compartilhar e no que confiar"[71]. Os rótulos aparecem no feed, nos stories, no perfil e nas mensagens diretas.

O rótulo é fixado também no que a empresa chama de "outras ocorrências do conteúdo". Para isso, utiliza sistemas informatizados que mapeiam as informações circulando e buscam identificar onde o post apareceu novamente por meio de uma técnica que o Instagram chama de "correspondência de imagens". Essa avaliação é realizada também entre Facebook e Instagram, com um rótulo sendo aplicado em uma rede social se o conteúdo foi verificado como falso em outra.

No caso do que trata como notícias falsas, são oferecidos também textos de contraponto elaborados pelos verificadores de fatos, algo semelhante à ferramenta dos "artigos relacionados" constante também no Facebook. Esses textos trazem os "desmentidos" ou informações oficiais que corrigem a alegação colocada no conteúdo, como forma de trazer uma segunda perspectiva, apresentada como correta, sobre aquele determinado tema.

Assim como no Facebook, diante da política específica para reduzir a disseminação de desinformação sobre vacinas na plataforma, em setembro de 2019 foi adicionado o recurso que direcionava pessoas que realizaram buscas sobre o tema para fontes oficiais, como o site da Organização Mundial da Saúde (OMS).

71. INSTAGRAM. Combater a desinformação no Instagram. Publicada em: 19 dez. 2019. Disponível em: https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram/

No dia 13 de março de 2020, a empresa anunciou que passaria a disponibilizar informações da OMS e de autoridades nacionais de saúde no topo do feed. Os links traziam informações sobre a pandemia e sobre formas de prevenção. A rede social passou a disponibilizar um canal para os sites da OMS e de ministérios da saúde locais para quem busca informações sobre o coronavírus. Além disso, foram disponibilizadas novas imagens para compartilhamento no stories com medidas de prevenção, como lavar as mãos e manter distância de outras pessoas.

Em abril de 2020, o Facebook publicou uma atualização das medidas adotadas para combater a desinformação na plataforma e no Instagram relacionada à pandemia do novo coronavírus. Entre elas estavam: 1) direcionamento a mais de 2 bilhões de pessoas de informações da OMS e outras autoridades de saúde no Facebook e Instagram. Dessas, 350 milhões clicaram para obter mais informações; 2) trabalho com mais de 60 agências de checagem em mais de 50 idiomas em todo o mundo. Entre março e abril foram adicionados oito novos parceiros e atuação em 12 países; 3) avisos associados a conteúdos verificados foram sobre 40 milhões de posts em relação a quatro mil conteúdos em março; 4) remoção de "centenas de milhares" de conteúdos de desinformação "que poderiam ter levado a danos físicos no mundo real". Exemplo: afirmação de que beber água sanitária seria eficiente na eliminação do vírus do corpo e que o distanciamento social não auxilia no combate à pandemia.

2.2.4 MEDIDAS CORRELATAS

As Diretrizes da Comunidade condenam apoio ou exaltação de práticas como discurso de ódio, incluindo "incentivo à violência ou o ataque a alguém com base em raça, etnia, nacionalidade, sexo, gênero, identidade de gênero, orientação sexual, religião, deficiências ou doenças" e "mensagens indesejadas enviadas repetidamente e

conteúdo que ataque pessoas com a intenção de constrangê-las ou humilhá-las". São permitidas o que a rede social chama de "discussões mais intensas" sobre personalidades públicas, definidas como indivíduos com destaque na mídia. Em março de 2019, atentado em uma mesquita na Nova Zelândia gerou reações por parte do Facebook, que abarcaram as contas relacionadas no Instagram.

Como no Facebook, as chamadas "contas falsas" são removidas do Instagram. Embora o nome real não seja obrigatório, as Diretrizes da Comunidade desaprovam um usuário que se passa por outra pessoa ou "a criação de contas com o objetivo de violar nossas diretrizes ou enganar outras pessoas". A maior parte dessa exclusão ocorre no momento da tentativa de abertura da conta ou logo após o processo de criação. Conforme exemplificado na seção anterior, a empresa realizou diversas operações nos últimos três anos de derrubada de contas ou redes de contas, enquadradas na categoria de "comportamento inautêntico", ou dessa prática de forma coordenada.

Em julho de 2018, o Instagram alterou a política de desativação de contas. Até aquele momento, a remoção de um perfil poderia ocorrer a partir de um determinado percentual do que a empresa chama de "conteúdos irregulares". A nova regra passou a considerar também determinados números de violações em um dado período. Além disso, foram incluídas novas notificações para avisar o usuário sobre o risco de desativação da conta. De acordo com a empresa, passou a operar o sistema de apelações quando o usuário questiona a classificação do conteúdo como não falso.

Em agosto de 2020, o Instagram publicou em seu blog institucional que passaria a implantar mecanismos de confirmação de identificação caso fossem verificados padrões de possíveis comportamentos inautênticos[72]. A companhia disse que passaria a observar um "conjunto de sinais" para avaliar se uma conta precisaria

72. INSTAGRAM. Introducing New Authenticity Measures on Instagram. https://about.instagram.com/blog/announcements/introducing-new-authenticity-measures-on-instagram.

fornecer sua identificação. No texto, a empresa declarou que a mudança se aplicaria "a uma pequena quantidade de usuários" e não à maioria das pessoas que atuam na plataforma. Se o responsável pelo perfil não fornecer a identificação, suas publicações passam a ter a circulação reduzida.

Outro tipo de "falsidade" proibida na rede social são as "avaliações falsas". As Diretrizes da Comunidade vetam publicação que "envolva, promova, incentive, facilite ou admita a oferta, a solicitação e o comércio de avaliações falsas de usuários". Entre as medidas adotadas pela companhia contra essas práticas está também o ajuizamento de ações judiciais contra empresas que comercializam esse tipo de reações manipuladas.

Em novembro de 2018, o Instagram publicou texto explicando iniciativas para reduzir "atividades não autênticas"[73], atitude inclusa entre aquelas que violam as Diretrizes da Comunidade e os Termos de Uso da plataforma. A rede social passou a remover curtidas, seguidores e comentários "não autênticos de contas que utilizam aplicativos de terceiros para impulsionar sua popularidade". Tais conteúdos e comportamentos foram elencados entre aqueles passíveis de remoção para os sistemas automatizados da rede, a partir de sua tecnologia de aprendizagem de máquina. Usuários passaram a receber mensagens notificando quando o comportamento foi detectado e quando apps de terceiros têm acesso a sua conta, permitindo inclusive a revogação dessa autorização.

Em março de 2019, a firma anunciou derrubadas de redes de perfis não autênticos do FB e do Instagram que atuavam de forma coordenada, no Reino Unido e na Romênia[74]. Não foram encon-

73. INSTAGRAM. Reduzindo atividades não autênticas no Instagram. Publicado em: 19 nov. 2018. Disponível em: https://about.instagram.com/blog/announcements/reducing-inauthentic-activity-on-instagram/.

74. GLEICHER, Nathaniel. Removing Coordinated Inauthentic Behavior From the UK and Romania. Publicado em: 7 mar. 2019. Disponível em: https://about.fb.com/news/2019/03/removing-cib-uk-and-romania/.

tradas relações entre as duas, mas a lógica de atuação era semelhante: criavam redes de contas e enganavam pessoas sobre quem eram e o que estavam realizando na plataforma. No caso do Reino Unido, o Facebook argumentou que eram grupos que trocavam nomes de grupos e distribuíam mensagens que "geravam divisão" no espectro político do país, tanto de ultradireita quanto antiultradireita. Na Romênia, as páginas e perfis difundiam conteúdo em defesa do Partido Social-democrata (PSD). Em setembro, novas redes de contas do FB e do Instagram foram excluídas na Espanha[75]. Elas atuavam amplificando narrativas favoráveis ao Partido Popular (PP) e críticas a seus opositores.

Em abril de 2019, a plataforma informou que a companhia entrou com processo contra uma empresa e pessoas na Nova Zelândia que comercializavam likes, views e seguidores falsos no Instagram. "Ao ajuizar a ação, estamos mandando uma mensagem de que esses tipos de atividade fraudulenta não são tolerados em nossos serviços, e que vamos atuar para proteger a integridade da nossa plataforma", afirmou a diretora de fiscalização de plataforma, Jessica Romero[76].

Nas páginas, passou a funcionar a identificação dos responsáveis e a confirmação pelo FB, a exemplo de contas no Twitter. Também serão exibidas informações sobre o local do responsável e se foi fundida com outras páginas. Essas ferramentas valerão apenas para as páginas com muitos seguidores. As páginas de veículos e instituições estatais passarão a ser submetidas a um nível maior de transparência. Serão incluídos, por exemplo, rótulos marcando as páginas de veículos controlados em todo ou em parte por esses entes. Essas identificações serão aplicadas também no Instagram.

75. GLEICHER, Nathaniel. Removing inauthentic behaviour in Spain. Publicada em: 20 set. 2019. Disponível em: https://about.fb.com/news/2019/09/removing-coordinated-inauthentic-behavior-in-spain/.

76. ROMERO, Jessica. Preventing Inauthentic Behavior on Instagram. Publicado em: 25 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/preventing-inauthentic-behavior-on-instagram/.

Como pontuado anteriormente, uma das preocupações do Instagram nos últimos anos vem sendo evitar que a plataforma gere consequências prejudiciais para o bem estar de seus usuários, especialmente em situações de *bullying* e assédio. Uma das respostas nesse sentido foi ampliar as alternativas de controle sobre a experiência on-line e as interações dos usuários. A empresa criou ferramentas para que eles possam bloquear contas se considerarem que estas causam algum prejuízo ou danos a sua experiência na plataforma. Além disso, podem controlar os comentários feitos em suas publicações.

Outras ferramentas, embora não tenham sido criadas diretamente em razão da mitigação de desinformação, podem gerar impactos. Uma delas é a opção de tornar "mudo" determinado post ou perfil. Além das possibilidades de denúncia e de classificação de uma publicação, permite que um usuário deixe de receber os conteúdos temporariamente caso passe a considerar determinado perfil como indesejável, o que pode valer para o caso do envio de notícias enganosas.

Em outubro de 2018, os fundadores do Instagram, Kevin Systrom e Mike Krieger, anunciaram que deixariam o comando da rede social e que a liderança seria assumida por Adam Mosseri. Até então, ele era gerente de produto do News Feed do Facebook e comandou toda a reação depois das acusações de interferência nas eleições dos Estados Unidos em 2016 e o escândalo da Cambridge Analytica. Mosseri assumiu em 9 de outubro anunciando medidas para combater o *bullying* e "espalhar gentileza" na plataforma[77]. Embora a prática seja diferente da desinformação, as respostas informadas podem interferir nesta última. Sistemas automatizados da plataforma utilizando aprendizagem de máquina (*machine learning*) passaram a analisar imagens para identificar aquelas fo-

77. MOSSERI, Adam. Novas ferramentas para combater o bullying e espalhar gentileza no Instagram. Publicado em: 9 out. 2018. Disponível em: <https://about.instagram.com/blog/announcements/anti-bullying-tools-on-instagram/>. Acesso: 14 dez. 2020.

tos "relacionadas ao *bullying*".

Em abril de 2019, a plataforma iniciou um teste em alguns países no qual o número de likes em um determinado post deixaria de ser exibido. No Brasil, a medida foi implementada em julho. A iniciativa fazia parte de uma série de reações promovidas pela empresa como forma de combater o discurso de ódio e o *bullying* no interior da plataforma.

Com foco no *bullying* on-line, o Instagram divulgou em julho de 2019 que os sistemas automatizados passariam a alertar usuários quando avaliarem que um comentário a ser publicado pode ser ofensivo. O intuito era "incentivar" que pessoas repensassem sobre os comentários que postariam. Assim como no Facebook, a desativação de contas passou a considerar não somente o percentual de conteúdo irregular, mas também o número de violações das políticas da rede social.

Em outubro de 2019, foi implementada uma ferramenta para "restringir" comentários com funcionamento diferente do bloqueio de contas. Nela, uma pessoa que sinta-se incomodada com comentários de um terceiro pode limitar as atividades deste, deixando os comentários do terceiro visível apenas para ele.

Em maio de 2020, o Instagram anunciou o lançamento de um novo recurso, chamado "guias". De acordo com a empresa[78], a iniciativa foi motivada pela pandemia do novo coronavírus e pelo aumento da atividade na plataforma durante o confinamento. Os guias visam oferecer conteúdos sobre "bem-estar", conectando produtores de conteúdo com influência na plataforma e organizações especializadas no tema de modo a promover a difusão de publicações com dicas e explicações sobre o assunto. A ferramenta foi lançada em oito países, entre eles Brasil, Estados Unidos, Reino Unido e Alemanha.

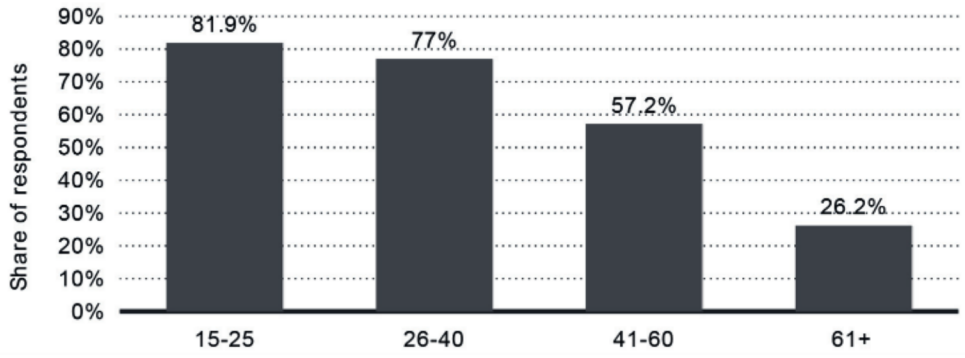
78. INSTAGRAM. Apoiar o bem-estar com guias do Instagram. Publicada em: 18 mai. 2020. Disponível em: https://about.instagram.com/blog/announcements/supporting-well-being-with-instagram-guides/.

2.3. WHATSAPP

O WhatsApp foi criado em 2009 por Brian Acton e Jan Koum, ambos veteranos do Yahoo. Em 2014, quando reunia 500 milhões de usuários e já havia desenvolvido ferramentas como compartilhamento de localização, conversas por áudios e possibilidade de criação de grupos, foi comprado pelo Facebook por US$ 19 bilhões. Em fevereiro de 2016, atingiu a marca de 1 bilhão de usuários ativos todos os meses. Em julho de 2017, já somava 1 bilhão de ativos todos os dias. Em 2018, 1,5 bilhão de usuários ativos todos os meses. Em março de 2020, ultrapassou a marca de dois bilhões de usuários em todo o mundo.

De acordo com a pesquisa *Social media usage in Latin America*, da empresa Statista, Índia e Brasil lideram a lista de países com base no número de usuários do WhatsApp em 2019 (em milhões), conforme demonstrado na figura abaixo:

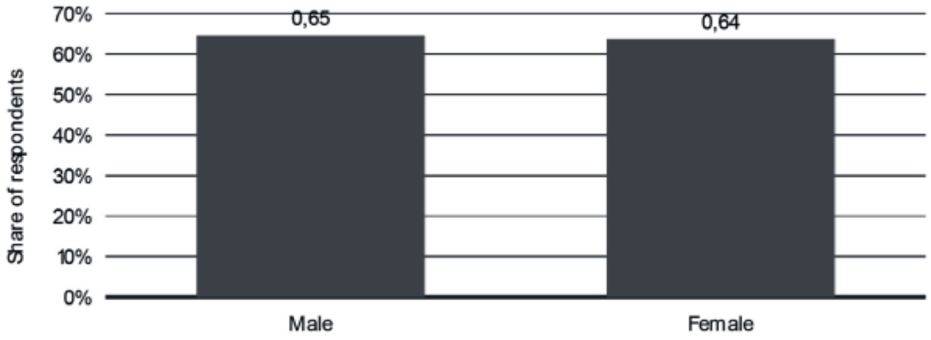
Países líderes com base no número de usuários do WhatsApp em 2019 (em milhões)



Fonte: Statista, 2020

Quanto ao gênero, a distribuição entre homens e mulheres é bastante próxima.

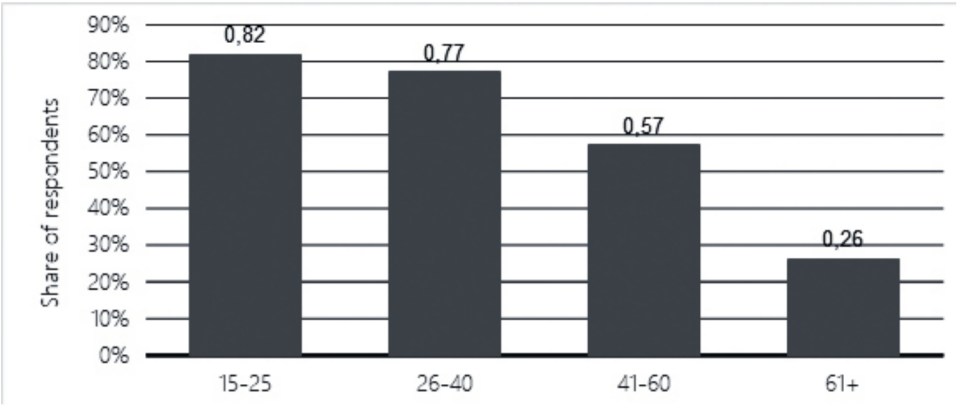
Percentual de utilizadores da WhatsApp na América Latina em 2018, por gênero



Fonte: Statista, 2020

Em relação à idade, há predomínio de usuários nas faixas, em primeiro lugar, dos 15 aos 25 anos e de 26 a 40 anos.

Percentual de utilizadores da WhatsApp na América Latina em 2018, por idade



Fonte: Statista, 2020

Em 2020, o número total de contas no Brasil já passa de 130 milhões. A relevância e a penetração da plataforma, que passou a disponibilizar recursos para realizar até pagamentos através dela, são inegáveis. A pesquisa *Panorama Mobile Time/Opinion Box sobre mensageria móvel no Brasil* revelou que o WhatsApp está

instalado em 99% dos *smartphones* brasileiros. Entre os usuários, 98% disseram que acessam o aplicativo todos os dias ou quase todos os dias. A *Global Mobile Consumer Survey Brasil 2019*[79] constatou que ele é o aplicativo mais utilizado no país e que mais de 80% dos brasileiros o acessam a cada hora.

Atestando a importância do conteúdo que nele circula, em dezembro de 2019, levantamento realizado pela Câmara dos Deputados e pelo Senado junto a 2.400 pessoas concluiu que 79% das pessoas utilizam o WhastApp como principal fonte de informação[80]. O relatório *Digital News Report 2020* do Instituto Reuters, por sua vez, mostrou que, no Brasil, o WhatsApp é apontado como a principal fonte de preocupação com as chamadas "fake news".

2.3.1. ABORDAGEM DO FENÔMENO

O WhatsApp viabiliza troca de conteúdos individualmente ou por meio de grupos e listas de transmissão, que permitem que uma mensagem seja enviada por um usuário para até outros 256. Todo o conteúdo é criptografado de ponta a ponta. Por tais características, apresenta-se como plataforma de mensagens privadas criptografadas. Ao ser questionada sobre a inexistência de política consolidada sobre desinformação dentro da plataforma e acerca de qual definição de desinformação é utilizada por ela, informou que, "como plataforma de mensagens privadas criptografadas de ponta a ponta, a WhatsApp não acessa as mensagens dos usuários e, portanto, não modera, julga, verifica os fatos, bloqueia ou retira qualquer conteúdo. Como tal, não há nenhuma definição de 'desinformação' dentro dos termos de serviço da WhatsApp".

79. Disponível em: https://canaltech.com.br/redes-sociais/whatsapp-nas-gracas-do-brasileiro-mais-de-80-checam-mensageiro-a-cada-hora-154069/
80. Disponível em: https://www12.senado.leg.br/radio/1/noticia/pesquisa-aponta-que-whatsapp-e-a-principal-fonte-de-informacao-de-79-dos-entrevistados

A corporação, seguindo tal entendimento, disse que não se aplicavam questionamentos sobre: diferenciação da desinformação de outras práticas como "discurso de ódio", "conteúdo abusivo" e "assédio", entre outras que possa usar para analisar conteúdos na rede; critérios e indicadores utilizados para a avaliação do conteúdo desinformativo; tempo médio para esta análise; como buscam/onde encontram conteúdo original para compará-los com suspeitas de mídias manipuladas; como são escolhidos os temas e momentos para os quais haverá ação específica, como no caso das vacinas ou da chegada do homem à lua; medidas como redução de alcance sobre determinados posts considerados desinformativos, entre outras.

Quanto ao estabelecimento de processos e/ou estrutura institucional para a integração e coordenação de ações sobre desinformação, respondeu que não possui procedimentos institucionais para definição de ações e tomada de decisões sobre o tema estabelecidos nem instância dedicada ao tratamento da questão. Não obstante, informou que, ao longo de 2020, o "WhatsApp treinou 41.228 pessoas em toda a América Latina para permanecerem seguras na plataforma e como combater spam, boatos e informações falsas no aplicativo".

Ainda que não trabalhe, como os pontos anteriores explicitam, com medidas voltadas especificamente à desinformação, ao ser perguntada sobre tratamento de conteúdos desinformativos relacionados a políticos, a companhia ressaltou que "as ações da WhatsApp foram reforçadas durante as eleições nos últimos anos, no Brasil e em toda a região, em parceria com coletivos de verificação de fatos compostos de várias redações locais (ou seja, no Brasil, a Comprova reuniu 24 redações em todo o país, enquanto no México, a Verificado reuniu 90 organizações), que trabalharam no monitoramento da desinformação e desinformação na preparação para as eleições de cada país". Acrescentou ainda que está analisando se outras medidas podem ser necessárias para evitar a desinformação.

No mesmo sentido, não informou sobre sanções que possam ser derivadas de ações de desinformação, apenas que só pode proibir contas como penalidade para o uso ilegítimo e/ou ilegal da plataforma. "Tais medidas podem ser aplicadas como resultado de detecção de padrões de uso que não correspondem aos do comportamento humano, utilizando ferramentas robustas de combate a SPAM e aprendizagem avançada de máquinas", diz o texto enviado à pesquisa. Esse tipo de medida encontra guarida nos documentos que orientam os usos do WhatsApp: os Termos de Serviço[81], a Política de Privacidade[82] e, em menor medida, para nosso caso, o Aviso de Privacidade – Brasil[83], que menciona adequação à Lei Geral de Proteção de Dados (LGPD) do país.

Tais documentos serão tratados com mais vagar posteriormente, no tópico sobre medidas correlatas. Para posicionar apenas a abordagem geral adotada pela plataforma, é útil ter em vista que, nos Termos de Serviço, ficam proibidos o uso de sistemas automatizados para gerenciar contas e envio de mensagens em massa, sendo mencionado entre os exemplos de uso que não é lícito e aceitável, e vedada utilização "envolvendo declarações falsas, incorretas ou enganosas". Já a Política de Privacidade do WhatsApp menciona, no tópico sobre "Proteção e Segurança", que "verificamos as contas e as atividades e promovemos a segurança dentro e fora de nossos Serviços, seja investigando atividades suspeitas e violações de nossos Termos ou garantindo que nossos Serviços sejam usados de forma lícita".

81. WHATSAPP, Termos de serviço do WhatsApp. Disponível em: https://www.whatsapp.com/legal/?lang=pt_br#terms-of-service

82. WHATSAPP. Política de Privacidade do WhatsApp. Disponível em: https://www.whatsapp.com/legal/?lang=pt_br#privacy-policy

83. WHATSAPP. Aviso de Privacidade – Brasil. https://www.whatsapp.com/legal/?lang=pt_br#brazil-privacy-notice

2.3.2. MODERAÇÃO DE CONTEÚDO

O WhatsApp informa oficialmente que não modera conteúdo. No entanto, desde abril de 2020, as "mensagens altamente encaminhadas" (Highly Forwarded Messages – HFM) – termo utilizado pela companhia para se referir a uma mensagem encaminhada através de uma cadeia de cinco ou mais chats, o que significa que ela está pelo menos a cinco conversas de distância de seu remetente original – só podem ser encaminhadas para um contato de cada vez. "Esta última mudança foi introduzida como uma forma de combater a disseminação de rumores, desinformação em meio à crise de saúde global da Covid-19", informou à nossa pesquisa, o que confirma a classificação dessa ação como de moderação de conteúdo associada à desinformação.

Em outra menção ao tema na entrevista, detalhou: "esta limitação foi introduzida após um aumento significativo na quantidade de encaminhamento, que pode contribuir para a disseminação da desinformação, com o objetivo de manter o WhatsApp como um lugar para conversas pessoais e privadas". Antes da adoção dessa medida no contexto da pandemia, o WhatsApp já vinha, desde meados de 2018, desenvolvendo mecanismos que incidem sobre a circulação de conteúdos, por meio de restrições ao encaminhamento de mensagens. Originalmente, uma mensagem poderia ser repassada a até 250 conversas (individuais ou grupos) de uma vez. Com a limitação naquele ano, o número máximo de chats para encaminhar um conteúdo passou a ser de 20. Ocorre que a plataforma não vinculava abertamente essa mudança à desinformação, por isso serão apresentadas aqui como medidas correlatas.

2.3.3. PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Quanto à promoção de informações e transparência, o site oficial do WhatsApp possui um espaço de "FAQ", um acrônimo da expressão inglesa Frequently Asked Questions, literalmente traduzida como "Questões Frequentemente Perguntadas". Dentro desta seção, no tópico "Segurança e Privacidade", há um texto intitulado "Dicas para ajudar a evitar a disseminação de boatos e notícias falsas". Ele apresenta informações resumidas sobre aprender a identificar mensagens encaminhadas; analisar fotos e arquivos de mídia cuidadosamente; procurar erros de ortografia; questionar opiniões tendenciosas e parciais; desconfiar das notícias que frequentemente viralizam; e verificar outras fontes. Esses conteúdos são acessíveis por meio do aplicativo, mas de forma pouco nítida. É preciso seguir os seguintes passos para chegar ao site: "Configurações", depois "FAQ", "Segurança e Privacidade" e, por fim, escolher "Dicas para ajudar a evitar a disseminação de boatos e notícias falsas" a partir da lista de títulos apresentada.

Como parte dos preparativos para as eleições brasileiras de 2018, o TSE assinou dois memorandos de entendimento com Google, Facebook e representantes da imprensa tradicional. As corporações se comprometeram a "prevenir e combater" as fake news, além de "apoiar a Corte em projetos de fomento à educação digital e em iniciativas de promoção do jornalismo de qualidade"[84]. Entre os compromissos, apoiar a prevenção de práticas dolosas de desinformação, bem como projetos de fomento à educação digital e iniciativas que promovam o jornalismo de qualidade. O memorando fazia referência à "possibilidade de replicação de práticas aptas a distorcer a liberdade do voto do eleitorado e a formação de escolhas conscientes por parte dos cidadãos".

84. AMARAL, Bruno do. TSE assina MoU com Google e Facebook contra fake news. Teletime. Publicado em 18 de junho de 2018. Disponível em: https://teletime.com.br/28/06/2018/tse-assina-mou-com-google-e-facebook-contra-fake-news/. Acesso: 10 dez. 2020.

Naquele período, o WhatsApp anunciou que premiaria "pesquisadores interessados em explorar questões relacionadas à desinformação (propagação de informações inverídicas) no WhatsApp"[85]. As abordagens foram divididas nos seguintes eixos: Processamento de informações de conteúdo problemático; Informações relativas a eleições; Efeitos de rede e viralidade; Alfabetização digital e desinformação; Detecção de comportamento problemático em sistemas criptografados. O estímulo não significa necessariamente impacto na adoção de medidas. "O objetivo desses prêmios de pesquisa é facilitar a pesquisa externa de alta qualidade sobre esses tópicos por acadêmicos e especialistas que estão nos países onde o WhatsApp é usado com frequência e onde há pesquisas relativamente limitadas sobre o tema", diz o comunicado. "Acreditamos que isso é importante para obter informações imparciais que explicam a variação social e cultural de como as pessoas usam nosso produto", apontou[86]. Premiações de US$ 50 mil (cerca de R$ 195 mil naquele período,) foram distribuídas aos autores de 20 projetos selecionados pela equipe do aplicativo em diferentes países: Brasil, Índia, Indonésia, Israel, México, Holanda, Nigéria, Cingapura, Espanha, Reino Unido e Estados Unidos. Do Brasil, foram agraciados apenas dois projetos: do Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio) e da Universidade Federal do Rio de Janeiro (UFRJ)[87].

85. WHATSAPP. Prêmio de Pesquisa WhatsApp para Ciências Sociais e Desinformação. Disponível em: https://www.whatsapp.com/research/awards/

86. FUNKE, Daniel. WhatsApp awards $1 million for misinformation research. Poynter. Publicada em 12 de novembro de 2018. Disponível em: <https://www.poynter.org/fact-checking/2018/whatsapp-awards-1-million-for-misinformation-research/>. Acesso: 10 dez. 2020.

87. AGRELA, Lucas. WhatsApp premia pesquisas de desinformação com US$ 1 milhão. Exame. Publicada em 13 de dezembro de 2018. Disponível em: <https://exame.com/tecnologia/whatsapp-premia-pesquisas-de-desinformacao-com-us-1-milhao/>. Acesso: 15 dez. 2020.

Em relação aos resultados e à incorporação de possíveis recomendações das pesquisas, a empresa informou que "o WhatsApp realizou dois workshops para os premiados para educá-los sobre o produto da plataforma, de modo que suas pesquisas fossem mais eficazes e, ao mesmo tempo, sustentassem uma comunidade global de acadêmicos e especialistas com foco na compreensão e no tratamento da desinformação. Todos os estudos são realizados de acordo com a ética das instituições acadêmicas dos pesquisadores e os processos de revisão acadêmica e não envolvem o uso de quaisquer dados de usuários da WhatsApp. A maioria dos estudos finais foi concluída recentemente ou está em seus estágios finais de produção. O WhatsApp está empenhado em analisar as conclusões desses projetos para promover a compreensão das medidas que podem ser tomadas tanto dentro de sua plataforma como através do apoio à sociedade civil, acadêmicos e especialistas para ajudar a resolver estas questões".

Quanto à sensibilização do público amplo, medidas foram adotadas também particularmente na Índia, onde houve inclusive ações de sensibilização nas ruas[88]. Ainda em 2018, também foi lançada campanha por meio de comerciais de TV naquele país para alertar sobre a propagação da desinformação[89]. Ações de promoção de informações ocorreram em menor escala no Brasil. Aqui, a busca pela conscientização do público foi por meio de campanhas como a "Compartilhe fatos, não rumores", projetada para alcançar 50 milhões de brasileiros, mas que o WhatsApp aponta que atingiu 8 milhões, com divulgação nos principais veículos de

88. GNIPPER, Patrícia. WhatsApp faz ação de rua na Índia na tentativa de combater fake news. Canal Tech. Publicada em 12 de outubro de 2018. Disponível em: <https://canaltech.com.br/apps/whatsapp-faz-acao-de-rua-na-india-na-tentativa-de-combater-fake-news-124688/>. Acesso: 15 dez. 2020.

89. ROSA, Natalie. WhatsApp lança seu primeiro comercial na TV para combater fake news. Canal Tech. Publicada em 03 de dezembro de 2018. Disponível em: <https://canaltech.com.br/apps/whatsapp-lanca-seu-primeiro-comercial-na-tv-para-combater-fake-news-128179/>. Acesso: 15 dez. 2020.

notícias do país, tanto nas versões impressas como digitais, assim como nas próprias plataformas da corporação, Facebook e Instagram[90]. A campanha ensinava a identificar fake news e quais passos tomar diante disso, dava dicas de como ficar seguro na plataforma e como a reportar e também ter maior discernimento em relação aos conteúdos. No final de 2019, foi veiculada uma nova edição, incluindo também mídia externa (outdoors) como parte dos canais de divulgação. Diretamente, ao longo de 2020, realizou treinamentos com 41.228 pessoas na América Latina sobre como permanecer seguro na plataforma e como combater spam, boatos e informações falsas no app.

Na esteira das ações do Facebook, corporação da qual faz parte, o WhatsApp também fez parcerias com projetos de checagem de fatos. Em agosto de 2018, às vésperas das eleições no Brasil, anunciou parceria com o Projeto Comprova, reunindo veículos jornalísticos brasileiros[91], do qual consta como parceiro institucional. Então, não foi firmada parceria direta para verificação de conteúdos, como fez o próprio Facebook por meio de aliança com Aos Fatos e Lupa, iniciativas que desde o primeiro semestre de 2018 passaram a ter acesso às notícias denunciadas como falsas pela comunidade no Facebook para analisar sua veracidade[92].

90. Dado apresentado por Ben Supple em seminário do TSE e registrado no livro: http://www.tse.jus.br/hotsites/catalogo-publicacoes/pdf/livro-digital-fake-news.pdf

91. O Comprova é coordenado pela Abraji (Associação Brasileira de Jornalismo Investigativo) a partir de uma iniciativa do First Draft, projeto do Centro Shorenstein da Harvard Kennedy School (EUA). As organizações de mídia envolvidas no Comprova são AFP, Band News, Band TV, Canal Futura, Correio do Povo, Exame, Folha de S. Paulo, GaúchaZH, Gazeta Online, Gazeta do Povo, Jornal do Commercio, Metro Brasil, Nexo Jornal, Nova Escola, NSC Comunicação, O Estado de S. Paulo, O Povo, Poder360, Rádio Band News FM, Rádio Bandeirantes, Revista Piauí, SBT, UOL e Veja. Assim como o Facebook, o WhatsApp é apresentado como parceiro institucional.

92. FACEBOOK. Facebook lança produto de verificação de notícias no Brasil em parceria com Aos Fatos e Agência Lupa. Publicada em 15 de maio de 2018. Disponível em: <https://about.fb.com/br/news/2018/05/facebook-lanca-produto-de-verificacao-de-noticias-no-brasil-em-parceria-com-aos-fatos-e-agencia-lupa/>. Acesso: 15 dez. 2020.

Antes das eleições na Índia, em 2019, o WhatsApp desenvolveu parceria com uma "empresa social" chamada Proto, a fim de desenvolver o *Checkpoint Tipline*, apresentado como projeto de pesquisa sobre desinformação. Embora em sua apresentação informe não se tratar de verificação de fatos, de acordo com informações divulgadas pela Proto[93], os usuários podem enviar um boato – incluindo link de texto, imagem ou vídeo – para o *Tipline*, por meio de um número disponível para tal fim. O usuário que fizer isso deve receber uma resposta de que o centro de verificação revisará o boato quanto à precisão. A velocidade da resposta depende de o boato ser ou não já conhecido. Além disso, um banco de dados com as checagens realizadas foi montado e relatórios foram produzidos com a expectativa de ajudar "o WhatsApp a determinar os próximos passos contra a desinformação na plataforma". A parceria está restrita à Índia.

A ampliação de medidas relacionadas à verificação ou checagem de fatos ocorreu em 2020, no contexto da pandemia do coronavírus. O WhatsApp anunciou então algumas medidas com o objetivo de conter a desinformação relacionada à situação. No dia 18 de março, criou o que foi chamado de *WhatsApp Coronavirus Information Hub*[94], em parceria com OMS, UNICEF, PNUD e IFCN (International Fact-Checking Network), que recebeu doação de US$ 1 milhão para expandir a presença de organizações de verificação de fatos no mensageiro.

O objetivo do hub, informou, é "fornecer orientações simples e acionáveis aos profissionais de saúde, educadores, líderes comunitários, organizações sem fins lucrativos, governos locais e empresas locais que dependem do WhatsApp para comunicar.

93. Disponível em: <https://www.checkpoint.pro.to/>. Acesso: 15 dez. 2020.

94. WHATSAPP. WhatsApp launches Coronavirus Information Hub to support health initiatives and donates $1m to International Fact-Checking Network for #CoronaVirus Facts Alliance. Publicada em 15 de março de 2020. Disponível em: <https://www.whatsapp.com/coronavirus/stories>. Acesso: 15 dez. 2020.

O site também oferece dicas e recursos gerais para os utilizadores de todo o mundo para reduzir a propagação de rumores e estabelecer uma ligação com informações de saúde precisas". Acrescentou que "tem trabalhado com vários ministérios nacionais da saúde e ONGs para fornecer informações factuais aos usuários por meio de texto em países como Singapura, Israel, África do Sul, Brasil e Indonésia. À medida que esses esforços prosseguirem, o centro será atualizado com os recursos mais recentes".

No dia 25 de março, novas ações foram anunciadas[95] pelo Facebook. Para o WhatsApp, a novidade foi a inclusão de fornecimento de informações sobre o tema, avançando, portanto, em uma ação editorial, ainda que com conteúdos organizados em parceria com órgãos e instituições. "As pessoas podem inscrever-se para receber o alerta de saúde da OMS sobre o WhatsApp, um relatório diário com os últimos números de casos de Covid-19. Inclui também dicas sobre como evitar a propagação da doença, bem como respostas às perguntas mais comuns que as pessoas podem facilmente enviar aos seus amigos e familiares", detalhou. Também divulgou em seu site[96] lista com "dicas" para evitar propagar fake news: 1) Cuidado com mensagens encaminhadas; 2) Desconfie de mensagens alarmantes; 3) Não clique em links; 4) Desconfie de erros e nomes; 5) Não tenha vergonha de informar ao remetente que se trata de notícia falsa; 6) Denuncie, acionando a opção "Denunciar contato" ou "Denunciar grupo".

Em abril, lançou o "Alerta de Saúde da OMS"[97], bot que dá

95. CLEGG, Nick. Combating COVID-19 Misinformation Across Our Apps. Publicada em 25 de março de 2020. Disponível em: <https://about.fb.com/news/2020/03/combating-covid-19-misinformation/>. Acesso: 15 dez. 2020.

96. EISHIMA, Rubens. WhatsApp lista dicas para não cair em fake news do coronavírus. Canal Tech. Publicada em 25 de março de 2020. Disponível em: <https://canaltech.com.br/apps/whatsapp-lista-dicas-para-nao-cair-em-fake-news-do-coronavirus-162395/>. Acesso: 15 dez. 2020.

97. JIN. Kang-Xing. Keeping People Safe and Informed About the Coronavirus. Publicada em 11 de dezembro de 2020. Disponível em: <https://about.fb.com/news/2020/12/

respostas a perguntas comuns sobre a Covid-19. O usuário que quiser utilizar precisa adicionar um número e mandar um "Hi". O serviço foi lançado inicialmente em inglês, e a plataforma pretende disponibilizá-lo nas seis línguas das Nações Unidas (inglês, árabe, chinês, francês, russo e espanhol). No Brasil, em março, o chat que havia era apenas uma parceria com o Ministério da Saúde. Lançado em março com o objetivo informado pela pasta de "combater a onda de fake news nas redes sociais", a ferramenta disponibiliza informações sobre a prevenção, diagnóstico, distância social, tratamentos e ações que acontecem no país, bem como dicas de profissionais de saúde.

No dia 4 de maio, a Rede Internacional de Verificação de Fatos (IFCN) lançou um *chatbot* no WhatsApp para, divulgaram os parceiros, conectar os milhões de usuários às mais de 80 organizações de verificação de fatos em todo o mundo, a fim de buscar informações junto a eles se determinado conteúdo sobre a Covid-19 já foi classificado como falso por verificadores profissionais. "O sistema é capaz de identificar o país do usuário por meio do código do país móvel e compartilhar as organizações de verificação de fatos mais próximas. Qualquer pessoa pode enviar informações para revisão diretamente aos verificadores de fatos locais ou visitar seu site para saber mais sobre as informações erradas que circulam na região", informou o Poynter[98].

Inicialmente disponível apenas em inglês, o *chatbot* passou, depois, a funcionar em espanhol e hindi. Em agosto de 2020, foi disponibilizado em português. Desenvolvido no Brasil em parceria com o ITS Rio, o serviço gratuito atende apenas casos de boatos

coronavirus/>. Acesso: 11 dez. 2020.

98. GRAU, Mel. New WhatsApp chatbot unleashes power of worldwide fact-checking organizations to fight COVID-19 misinformation on the platform. Poynter. Publicada em 04 de maio de 2020. Disponível em: <https://www.poynter.org/fact-checking/2020/poynters-international-fact-checking-network-launches-whatsapp-chatbot-to-fight-covid-19-misinformation-leveraging-database-of-more-than-4000-hoaxes/>. Acesso: 11 dez. 2020.

relacionados à pandemia do novo coronavírus. Para usá-lo, é preciso gravar o número +1 (727) 291 2606 na agenda do celular e mandar um "oi", então o bot vai iniciar uma conversa em português. De acordo com os desenvolvedores, ele sugerirá, principalmente: a busca por uma checagem específica; a apresentação das checagens mais recentes; e dicas de como combater a desinformação. No caso da busca por uma checagem específica, é necessário escrever "1". O chatbot vai explicar como fazer a busca, a partir de uma palavra-chave ou de perguntas. Ao escrever "2", o sistema fornece as checagens mais recentes. Importante ter em vista que o chatbot da IFCN apresenta mensagens dentro do próprio app – não precisando clicar em links externos.

"O WhatsApp implementou parcerias com verificadores de fatos que permitem aos usuários interagir com eles sobre o aplicativo, compartilhando informações duvidosas para serem revistas. Consiste em um mecanismo pelo qual, em primeiro lugar, os usuários encaminham aos verificadores de fatos as informações que desejam verificar (eles podem fazer isso diretamente às organizações listadas no site FAQ do WhatsApp, ou através do *chatbot* de verificação de fatos do IFCN). Em segundo lugar, o conteúdo relatado é analisado. Finalmente, o usuário recebe o feedback do verificador de fatos. Essas organizações obtêm acesso à API da WhatsApp, que lhes permite responder às submissões dos usuários e, eventualmente, refutar a desinformação", resume a empresa.

Parte de acordo do TSE com plataformas e outros agentes para combater a desinformação nas eleições de 2020, foi criado um *chatbot* no WhatsApp para denúncias. Para conversar com o assistente e receber informações sobre cuidados sanitários, dicas ao eleitor, regras do processo, notícias checadas e dados da Justiça Eleitoral, o usuário tem que adicionar o número 06196371078 ou usar um link e entrar em contato. Não foi feita notificação direta para os usuários, portanto eles também terão que tomar conhecimento do mecanismo. De acordo com o TSE, até o fim da eleição,

especificamente no dia 29 de novembro, o canal contou com mais de um milhão de usuários únicos. A plataforma disponibilizou figurinhas sobre voto consciente para as eleições de 2020.

A cooperação entre TSE e WhatsApp resultou na criação de um formulário para denunciar contas suspeitas de realizar disparos em massa, conduta proibida pela lei eleitoral e também pelos Termos de Serviço do aplicativo, o que está relacionado às medidas correlatas que serão explicitadas a seguir. Ambos divulgaram que, entre 27 de setembro e 29 de novembro, a plataforma de denúncias para contas suspeitas de disparos de mensagens em massa recebeu 5.180 registros, que foram enviados à plataforma. Desses, o WhatsApp baniu 1.042 números (29,5%) por violação de seus Termos de Serviço. Independentemente do canal com o TSE, foram banidas mais de 360 mil contas no Brasil por envio massivo ou automatizado de mensagens (abuso/spam), no período de setembro a novembro de 2020, a partir da operação do sistema da própria plataforma – um número muito superior ao de denúncias recebidas via TSE, o que confirma a centralidade das operações das próprias corporações, muitas baseadas em tecnologia de aprendizado de máquina, que identifica comportamento abusivo sem ter acesso ao conteúdo das conversas no aplicativo.

2.3.4. MEDIDAS CORRELATAS

Ainda que não estejam diretamente associadas à desinformação, algumas medidas implementadas pelo WhatsApp se relacionam com ela, a começar pelas regras que definiu como orientadoras dos usos do aplicativo. De acordo com os Termos de Serviço do WhatsApp, são consideradas violações de seus termos as seguintes ações: (a) de forma a violar, apropriar-se indevidamente ou infringir direitos do WhatsApp, dos seus usuários ou de terceiros, inclusive direitos de privacidade, de publicidade, de propriedade intelectual ou outros

direitos de propriedade; (b) de forma ilícita, obscena, difamatória, ameaçadora, intimidadora, assediante, odiosa, ofensiva em termos raciais ou étnicos, ou que instigue ou encoraje condutas que sejam ilícitas ou inadequadas, inclusive a incitação a crimes violentos; (c) envolvendo declarações falsas, incorretas ou enganosas; (d) para se passar por outrem; (e) para enviar comunicações ilícitas ou não permitidas, mensagens em massa, mensagens automáticas, ligações automáticas e afins; ou (f) de forma a envolver o uso não pessoal dos serviços, a menos que esteja autorizado pela companhia. O item destacado havia sido mencionado ao tratarmos do posicionamento da corporação quanto à desinformação.

No documento, detalha que as seguintes práticas causam prejuízos à plataforma e aos usuários: "(a) fazer engenharia reversa, alterar, modificar, criar trabalhos derivados, descompilar ou extrair códigos dos nossos Serviços; (b) enviar, armazenar ou transmitir vírus ou outros códigos nocivos usando nossos Serviços; (c) obter ou tentar obter acesso não autorizado aos nossos Serviços ou sistemas; (d) interferir ou interromper a integridade ou o desempenho de nossos Serviços; (e) criar contas por nossos Serviços usando meios não autorizados ou automatizados; (f) coletar informações de ou sobre os nossos usuários de maneira não permitida ou autorizada; (g) vender, revender, alugar ou cobrar por nossos Serviços; ou (h) distribuir ou disponibilizar os nossos Serviços em rede para ser usado por vários dispositivos ao mesmo tempo".

Embora não direta e expressamente voltadas ao combate à desinformação, outras medidas do WhatsApp guardam relação com esse problema, a exemplo das limitações ao encaminhamento, mencionada no tópico anterior. A alteração do repasse começou em julho de 2018, na Índia, local que, como vimos, registra o maior número de contas. Em comunicado oficial[99], a empresa de-

99. WHATSAPP. Mudanças no encaminhamento de mensagens. Publicada em 19 de julho de 2018. Disponível em: <https://blog.whatsapp.com/more-changes-to-forwarding>. Acesso: 11 dez. 2020.

talhou que aquele país é ainda onde as pessoas mais encaminham mensagens, fotos e vídeos e que, por isso, como parte do teste, a limitação foi maior do que em outros países: para 5 chats e não 20. Também foi removido "o botão de encaminhamento rápido que aparece ao lado de mensagens com arquivos de mídia", que era posicionado ao lado dos conteúdos multimídia, o que facilitava o acesso. O texto justificava: "acreditamos que essas mudanças – as quais continuaremos a avaliar – ajudarão a manter o WhatsApp dentro do princípio em que ele foi desenvolvido: um aplicativo de mensagens privadas".

O encaminhamento para até cinco contatos por vez foi adotado em todas as versões do aplicativo a partir de janeiro de 2019, inclusive no Brasil[100], com o argumento de que isso manteria "a privacidade da comunicação com contatos próximos" e a promessa de que a empresa seguiria buscando "formas de lidar com conteúdo viral". Outra alteração que também possui a finalidade de reduzir a viralização foi a remoção, em todo o mundo, do botão de "encaminhamento rápido"[101], que já havia sido removido na Índia.

No mesmo período em que começaram as mudanças no encaminhamento, julho de 2018, mensagens repassadas passaram a ser sinalizadas com uma seta. Oficialmente, a plataforma informou que isso ajudaria a "determinar se um amigo ou familiar realmente escreveu a mensagem que enviou ou se a mesma veio originalmente de outra pessoa". A companhia afirmou que encoraja o usuário a "pensar com cuidado antes de compartilhar mensagens encaminhadas. Lembre-se de que você pode bloquear ou denunciar um

100. WAKKA, Wagner. WhatsApp começa a limitar encaminhamento de mensagens a até 5 pessoas por vez. Canal Tech. Publicada em 17 de janeiro de 2019. Disponível em: <https://canaltech.com.br/apps/whatsapp-comeca-a-limitar-encaminhamento-de-mensagens-a-ate-5-pessoas-por-vez-131025/>. Acesso: 11 dez. 2020.

101. HAYASHI, Eduardo. WhatsApp para iOS ganha recurso para combater envio de fake news. Canal Tech. Publicada em 06 de agosto de 2018. Disponível em: <https://canaltech.com.br/apps/whatsapp-para-ios-ganha-recurso-para-limitar-proliferacao-de-fake-news-119530/>. Acesso: 11 dez. 2020.

contato como spam em um simples toque e que também pode sempre contatar o WhatsApp diretamente para ajuda", diz o texto[102].

Em agosto de 2019, o serviço começou a avisar sobre conteúdos compartilhados mais do que cinco vezes por uma seta dupla. A mudança incide tanto no diálogo entre usuários individuais quanto nos grupos[103]. O total de compartilhamentos, em si, não é exibido[104]. No comunicado oficial, a empresa não falou em fake news ou desinformação, apontou apenas que a etiqueta que já indica quando uma mensagem é encaminhada também passaria a incluir um aviso quando ela for compartilhada mais do que cinco vezes, o que a configura como HFM.

Um ano depois, no dia 3 de agosto de 2020, foi inserida uma lupa ao lado das mensagens assinaladas com duas setas (HFM), a fim de permitir que o usuário verifique na internet o conteúdo das mensagens encaminhadas com frequência. "Ao fornecer uma maneira simples de pesquisar na internet sobre o conteúdo desse tipo de mensagem, podemos ajudar nossos usuários a encontrar notícias ou outras fontes de informação sobre o conteúdo recebido. Esse recurso permite que os usuários carreguem a mensagem

102. CIRIACO, Douglas. Aviso que mensagem é encaminhada chega a todos os usuários do WhatsApp. Tecmundo. Publicada em 11 de julho de 2017. Disponível em: <https://www.tecmundo.com.br/software/132096-aviso-mensagem-encaminha-whatsapp.htm>. Acesso: 11 dez. 2020.

103. DEMARTINI, Felipe. WhatsApp começa a avisar usuários sobre mensagens encaminhadas muitas vezes. Tecmundo. Publicada em 05 de agosto de 2019. Disponível em: <https://canaltech.com.br/apps/whatsapp-comeca-a-avisar-usuarios-sobre-mensagens-encaminhadas-muitas-vezes-145835/>. Acesso: 11 dez. 2020.

104. Em fevereiro, vazou informação de que os usuários poderiam acessar mais informações sobre as mensagens. Reveladas pelo site WABeta Info, as informações de contagem de compartilhamento apareceriam nas mensagens também encaminhadas pelo usuário. Nesse caso, ao acessar o menu relacionado a um texto, ele poderia ver quantas vezes o conteúdo foi passado adiante. Entretanto, o contador não foi adotado, e o aplicativo passou a informar apenas que a comunicação segue adiante "com frequência". Disponível em: <https://canaltech.com.br/apps/whatsapp-vai-indicar-quantas-vezes-sua-mensagem-foi-encaminhada-135397/>.

diretamente pelo navegador do celular ou do computador sem que o WhatsApp tenha acesso ao conteúdo da mensagem"[105], informou a companhia, que disponibilizou a ferramenta para os usuários no Brasil, Espanha, Estados Unidos, Irlanda, Itália, México e Reino Unido que usam a versão mais recente do WhatsApp para Android e iOS e do WhatsApp Web/Computador.

Nos últimos anos, também foram adotadas medidas que modificaram as configurações dos grupos, as quais merecem ser mencionadas porque viabilizam que os usuários acessem mais informações sobre os grupos nos quais recebem as mensagens. No primeiro semestre de 2018, foram adicionados novos recursos, entre os quais: descrição do grupo (com um detalhe importante: quando uma nova pessoa entra em um grupo, a descrição passou a aparecer no topo da conversa) e controles de administrador (uma opção permitiu que os administradores restrinjam quem pode alterar o assunto, a imagem e a descrição do grupo).

Outra questão relevante: a partir dessa atualização, os criadores do grupo não podem mais ser removidos dele. Também foi incluído um mecanismo que impede que os usuários sejam adicionados novamente[106] ou possam voltar por vontade unilateral dos administradores aos grupos que deixaram. Em julho daquele ano, foi lançada uma configuração que permite que somente administradores possam enviar mensagens a um grupo. "Muitas pessoas utilizam grupos para receber comunicados e informações importantes, como pais e professores de uma escola, centros comunitá-

105. WHATSAPP. Apresentamos o novo recurso de pesquisar na internet. Publicada em 03 de agosto de 2020. Disponível em: <https://blog.whatsapp.com/search-the-web>. Acesso: 11 dez. 2020.

106. A medida que impede que usuários sejam incluídos em grupos sem autorização foi apresentada pelo diretor de políticas públicas, Pablo Bello, como exemplo de combate às notícias falsas em reunião com o TSE. ALECRIM, Anderson. Google, Facebook e Twitter prometem combater notícias falsas nas Eleições 2020. Tecnoblog. Publicada em 12 de novembro de 2019. Disponível em: <https://tecnoblog.net/314362/tse-google-facebook-twitter-whatsapp-combate-noticias-falsas-eleicoes-2020/>. Acesso: 13 mai. 2020.

rios e organizações sem fins lucrativos. Estamos apresentando esta nova configuração para que administradores tenham ferramentas melhores para casos como estes", explicou[107].

Em outubro de 2019, o app liberou[108] uma opção para impedir adição em grupos sem consentimento. O blog do WhatsApp havia anunciado, ainda no início de abril, novas configurações de privacidade para grupos[109] e mudança no sistema de convite, que dariam ao usuário poder de decisão sobre quem pode adicioná-los. Ficou assim: "o admin que não puder adicionar você a um grupo terá a opção de enviar um convite privado por conversa individual, e você poderá decidir entrar no grupo ou não. Você terá três dias para aceitar o convite antes que ele expire. Com esses novos recursos, os usuários terão mais controle sobre as mensagens que recebem em conversas em grupo", detalha o texto. O experimento foi inicialmente realizado com alguns usuários indianos, e aos poucos começou a ser liberado para todo mundo, tanto no iOS quanto no Android.

As ações de cunho preventivo não têm sido as únicas adotadas. O WhatsApp utiliza ferramentas robustas de tratamento de spam e aprendizagem avançada de máquinas para retirar mensagens automatizadas em massa e banir contas que fazem uso ilegítimo ou ilegal de sua plataforma. Como tratado em artigo técnico[110] publi-

107. WHATSAPP. Novas Configurações de Grupo para Administradores. Publicada em 02 de agosto de 2018. Disponível em: <https://blog.whatsapp.com/new-group-setting-for-admins>. Acesso: 11 dez. 2020.

108. YUGE, Claudio. WhatsApp finalmente libera opção para impedir adição em grupos sem consentimento. Publicada em 21 de outubro de 2019. Disponível em: <https://canaltech.com.br/apps/whatsapp-finalmente-libera-opcao-para-impedir-adicao-em-grupos-sem-consentimento-153133/>. Acesso: 11 dez. 2020.

109. WHATSAPP. Novas configurações de privacidade para grupos. Publicada em 5 de novembro de 2019. Disponível em: <https://blog.whatsapp.com/new-privacy-settings-for-groups>. Acesso: 11 dez. 2020.

110. Disponível em: <https://scontent.whatsapp.net/v/t61.22868-34/69510151_652112781951150_6923638360331596993_n.pdf/artigo-t%C3%A9cnico-Bloqueio-de-uso-n%C3%A3o-autorizado.pdf?_nc_sid=2fbf2a&_nc_ohc=aWUP9qEzDlMAX86Nw

cado em fevereiro de 2019, sistemas de *machine learning* têm sido utilizados para detectar e eliminar usuários com comportamentos inadequados, a exemplo do envio de mensagens em massa e da criação de múltiplas contas com o único objetivo de divulgar conteúdos questionáveis. De acordo com dados oficiais, cerca de dois milhões de contas são banidas globalmente todos os meses. Este sistema de aprendizagem de máquinas têm levado à proibição de 20% das contas consideradas problemáticas, segundo a empresa, já na criação, portanto antes mesmo de serem ativadas para uso.

Entre as informações avaliadas pelos sistemas de *machine learning* em relação a contas ativadas estão: o número de IP, o país do número de telefone, a idade da conta, o número de disparos desde a criação e a percepção do status "digitando". Se essa mensagem do status, por exemplo, não aparecer e muitas mensagens forem disparadas sem ativar esse recurso, ocorre o bloqueio da conta. Redes de conexão também dão pistas sobre contas falsas. Se estas já tiverem sido usadas para fazer spam, são bloqueadas a partir do sistema de inteligência artificial, sem intervenção humana. Também são consideradas: a velocidade de envio ou encaminhamento de mensagens, o nível de reciprocidade uma vez enviadas ou encaminhadas as mensagens (o que inclui não só o nível de interação recíproca direta entre um remetente e um destinatário, mas também se uma conta estabelece automaticamente conversas de grupo sem interações prévias) e se os destinatários estão na lista de contatos do remetente.

Além disso, é levado em conta o feedback negativo (possíveis reclamações, bloqueios do próprio usuário ou dos conteúdos), a partir da análise da resposta dos usuários. "O WhatsApp simplificou o processo de relatórios dos usuários que fazem contatos/encontros indesejados (ou seja, solicitações de 'novos' usuários ou grupos de usuários), e bloqueia contas com base no feedback

A0&_nc_ht=scontent.whatsapp.net&oh=d3e54d0156909bf84812750410062d28&oe=5F3834EF>. Acesso: 15 mai. 2020.

negativo desses relatórios", detalhou em entrevista à pesquisa. Se um usuário chegar a um limite máximo desse feedback, é banido definitivamente.

Em novembro de 2019, a empresa informou que baniria todos os membros de grupos que contenham nomes de atos ilegais ou que pareçam como tais. Isso porque, em fóruns na internet, foram registrados casos de nomes que eram irônicos de grupos que não praticavam ou estimulavam crime algum[111]. Os membros banidos da plataforma não tiveram a possibilidade de reaver a conta utilizada. Somando essas diferentes estratégias calcadas no banimento, a empresa calcula que, "em geral, 75% do uso ilegítimo é banido automaticamente", conforme apresentou à nossa pesquisa.

Adicionalmente, possui política voltada especificamente ao combate à exploração de crianças, em relação à qual afirma ter política de "tolerância zero". Na versão em inglês do Blog do WhastApp, há uma página dedicada ao tema, intitulada *How WhatsApp Helps Fight Child Exploitation*[112] (que podemos traduzir como "Como o WhatsApp ajuda a combater a exploração infantil"). Na página em português, há apenas um parágrafo sobre isso, parte das informações sobre *Informações para as autoridades policiais*. No texto em inglês, afirma que, nos três meses anteriores à data da publicação, 6 de fevereiro de 2019, havia banido aproximadamente 250.000 contas por mês suspeitas de compartilhamento de imagens exploradoras de crianças.

O WhatsApp diz que, para prevenir o compartilhamento de imagens de exploração infantil, utiliza todas as informações não

111. RIBEIRO, Gabriel Rodrigues. WhatsApp está banindo membros de grupos com zoeira além do limite; entenda. Publicada em 11 de novembro de 2019. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2019/11/11/whatsapp-esta-banindo-membros-de-grupos-com-zoeira-alem-do-limite-entenda.htm?cmpid=copiaecola>. Acesso: 11 dez. 2020.

112. WHATSAPP. How WhatsApp Helps Fight Child Exploitation. Publicada em 6 de fevereiro de 2019. Disponível em: <https://faq.whatsapp.com/general/how-whatsapp-helps-fight-child-exploitation?lang=ca>. Acesso: 11 dez. 2020.

criptografadas disponíveis, incluindo relatórios de usuários. Além disso, diz utilizar aprendizagem de máquinas para escanear e remover proativamente compartilhamento de imagens exploradoras de crianças de fotos de perfil, informações de grupo, relatórios de usuários. Para a detecção, "usamos a tecnologia de correspondência de fotos chamada PhotoDNA para escanear proativamente fotos de perfil para a conhecida CEI [imagens exploradoras de crianças]. Caso nossos sistemas detectem tal imagem, proibiremos o usuário e as contas associadas dentro de um grupo. Embora possamos não ser capazes de distinguir entre um usuário que envia ou recebe imagens CEI, nossa política de tolerância zero significa que proibimos todos os membros do grupo de usar o WhatsApp", diz o texto.

Na versão brasileira da política sobre o tema, afirma apenas que "denunciamos todos os casos aparentes de exploração infantil que aparecerem em nosso site, em qualquer lugar do mundo, ao NCMEC (Centro Nacional para Crianças Desaparecidas e Exploradas), incluindo conteúdo trazido a nosso conhecimento através de solicitações oficial"[113].

No fim de 2019, divulgou que processaria usuários que encaminhem mensagens em massa, argumentando violação de seus termos e condições de uso. A ação seria movida contra praticantes de envio de mensagens em massa ou automatizados, bem como o seu uso não-pessoal. Poderia partir de informações distribuídas fora da plataforma, conforme explicitou em nota: "afirmações de empresas que dizem ter a habilidade de utilizar o WhatsApp de forma que viole os nossos termos. Isso serve como uma notificação pública de que vamos tomar ações legais contra empresas onde nós tenhamos apenas evidências de abuso fora de nossa plataforma ou se esse abuso continuar após a referida data, ou ainda se tais empresas forem relacionadas com evidências de abuso antes dessa

113. WHATSAPP. Información para las fuerzas del orden. s/d. Disponível em: <https://faq.whatsapp.com/general/security-and-privacy/information-for-law-enforcement-authorities/>. Acesso: 11 dez. 2020.

data. Nós temos o compromisso de reforçar a natureza privada de nossa plataforma e manter os usuários livres de abuso"[114].

Sobre isso, detalhou em entrevista para esta pesquisa: "o WhatsApp também adota medidas para coibir a oferta de serviços de disparo de mensagens em massa, tendo enviado centenas de notificações a empresas que ofereciam tais serviços em violação aos termos de uso do aplicativo e às leis aplicáveis. O WhatsApp inclusive adota medidas judiciais para resguardar os seus direitos e colabora com as autoridades que investigam potenciais abusos em campanhas eleitorais". Parte dessa linha de ação, empresas que comercializam serviços que são contra as políticas da WhatsApp foram notificados para "cessar imediatamente suas atividades" e também o que a companhia considera "violação de direitos da marca". Como resultado disso, há ações judiciais em andamento, mas nenhuma ação criminal foi iniciada.

O WhatsApp afirma que "está comprometido em responder com rapidez e precisão às requisições das autoridades de investigação". Atualmente, a companhia disponibiliza[115] Sistema de Solicitação On-line para Autoridades de Investigação Criminal do WhatsApp permanentemente em seu site e em português, possibilitando que autoridades possam enviar requisições rapidamente. Por meio desse sistema, as autoridades podem preservar dados de contas; requisitar dados de usuários e enviar requisições emergenciais, quando há risco de morte ou danos corporais graves. Ainda nesta seara institucional, "com o objetivo de estabelecer um diálogo constante, apresentar com transparência as ferramentas disponíveis e os processos aplicáveis para obtenção de dados de

114. ARBULU, Rafael. WhatsApp quer processar usuários que encaminham mensagens em massa. Canal Tech. Publicada em 10 de junho de 2019. Disponível em: <https://canaltech.com.br/apps/whatsapp-quer-processar-usuarios-que-encaminham-mensagens-em-massa-141353/>. Acesso: 11 dez. 2020.

115. WHATSAPP. Law Enforcement Online Requests. s/d. Disponível em: <https://www.whatsapp.com/records/login/>. Acesso: 11 dez. 2020.

usuários para fins de investigação criminal, o WhatsApp realiza treinamentos para autoridades policiais, membros do Ministério Público e do Poder Judiciário em todo o pelo país", apresentou em entrevista. Segurança digital e fake news são alguns dos temas discutidos nesses encontros.

Por fim, pode ser considerada medida correlata ação de comunicação com a sociedade. No caso do WhatsApp, apenas uma foi registrada, além da "Compartilhe fatos, não rumores", citada ao debatermos a promoção de informações relacionadas à desinformação. Em fevereiro de 2020, foi lançada o que é apontado como primeira campanha publicitária do WhatsApp no país, intitulada "Fica só entre vocês"[116]. Com ela, buscou se posicionar como um apoio para os usuários estabelecerem "conexões verdadeiras por meio da comunicação privada". Para esta pesquisa, informou que se trata de uma campanha educacional voltada para privacidade e prevenção de golpes envolvendo o aplicativo. Os anúncios mostraram como se proteger de golpes relacionados ao app, como proceder caso seja vítima de uma abordagem suspeita e, por fim, como recuperar uma conta que já tenha sido roubada. A campanha foi veiculada na Internet e nos principais jornais e revistas do país, de acordo com o WhatsApp.

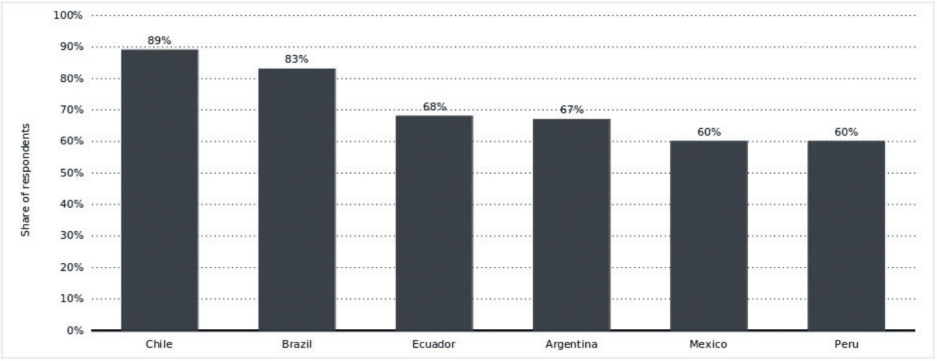
2.4. YOUTUBE

Lançado em 2005 e adquirido pela Google um ano depois, o YouTube passou de um repositório de vídeos amadores para a maior plataforma de vídeos on-line do mundo. Segundo o Statista, o YouTube registrou um número estimado de 1,68 bilhão de usuários ativos por mês em 2019, ano em que avançou também na oferta do aplicativo YouTube Music, que foi baixado mais de 13,3 milhões de

116. CLUBE DA CRIAÇÃO. Law Fica só entre vocês. WhatsApp lança sua 1ª campanha de marca do mundo. Publicada em: 17 de fevereiro de 2020. Disponível em: <https://www.clubedecriacao.com.br/ultimas/fica-so-entre-voces/>. Acesso: 11 dez. 2020.

vezes na Google Play Store e mais de quatro milhões na Apple App Store, tornando-o o aplicativo de música líder mundial. Em 2020, anunciou que já atingiu 1,8 bilhão de usuários ativos por mês, o que o deixa no lugar de segunda maior rede social do mundo, perdendo apenas para o Facebook. No Brasil, segundo dados do Global Digital 2019, um relatório anual da WeAreSocial, 95% dos usuários de Internet brasileiros assistem a vídeos no YouTube, o que equivale a 133 milhões de pessoas. Entre as crianças, segundo o Statista, um dos maiores portais estatísticos do mundo, a penetração era de 83% em 2019. Mais de mil canais contam com 1 milhão de inscritos.

Percentual de uso do YouTube entre crianças em países da América Latina em 2019



Fonte: Statista (2020)

2.4.1. ABORDAGEM DO FENÔMENO

O YouTube estabelece para seus usuários conjuntos de normas para quem publica conteúdo na plataforma: as Diretrizes da Comunidade[117], os Termos de Uso[118], as Políticas[119] do Programa

117. https://www.youtube.com/about/policies/#community-guidelines
118. https://www.youtube.com/t/terms
119. https://support.google.com/adsense/answer/48182?ctx=checklist

Google AdSense, as políticas de monetização do Youtube[120] e as diretrizes de conteúdo adequado para publicidade[121] – os três últimos válidos para quem quiser receber recursos com os vídeos que publicar no YouTube. Neste documento, trataremos apenas das regras relacionadas ao enfrentamento à desinformação e daquelas que podem impactar na sua disseminação, direta ou indiretamente.

Em entrevista a esta pesquisa, o YouTube afirmou que não trabalha com uma única definição de desinformação. Acredita que boa parte do que poderia ser incluído num dado conceito não necessariamente seria comprovadamente verdadeiro ou falso, e que nem todas as formas de desinformação devem ser proibidas. A plataforma, apesar de considerar a importância de diferentes taxonomias sobre o tema, avalia ainda que seria complicado analisar o volume de conteúdos que circulam pela plataforma com base nelas. Atualmente, cerca de 500 horas de vídeos são publicadas por minuto globalmente no YouTube. A empresa explica que o que não é permito está explícito e descrito nas Diretrizes da Comunidade e que busca ser o mais clara possível sobre o que faz. Utiliza, para o conjunto das publicações, uma abordagem que chama de "4Rs": Remover, Recomendar, Reduzir e Recompensar[122].

Moderações realizadas desde o início de 2020, relacionadas à Covid-19, por exemplo, são pautadas na análise de conteúdo que cause dano, como detalhado no próximo tópico. O YouTube também tem uma política de proibição de mídias manipuladas, que não podem circular na plataforma.

120. YOUTUBE. Políticas de monetização para canais do YouTube. S/D. Disponível em: <https://support.google.com/youtube/answer/1311392>. Acesso: 11 dez. 2020.

121. YOUTUBE. Diretrizes de conteúdo adequado para publicidade. s/d. Disponível em: <https://support.google.com/youtube/answer/6162278?hl=pt-BR&ref_topic=9153642>. Acesso: 11 dez. 2020.

122. YOUTUBE. Por dentro do YouTube. Os 4 Rs da Responsabilidade: Parte 1 - Removendo conteúdo nocivo. Publicado em: 25 de novembro de 2019. Disponível em: <https://blog.youtube/intl/pt-br/inside-youtube/os-4-rs-da-responsabilidade-parte-1/>. Acesso: 11 dez. 2020.

A empresa explica que usa uma combinação entre pessoas e tecnologias para marcar um conteúdo que requer revisão de suas equipes. Uma das tecnologias empregadas, as *hashes*, identifica cópias de conteúdos já sabidamente violadores das políticas da plataforma antes deles serem disponibilizados aos usuários. Desde 2017, o Google usa *machine learning* para detectar conteúdos potencialmente violadores com base na similaridade de padrões de conteúdos já avaliados. Categorias como spam e conteúdo adulto são mais facilmente identificáveis pela tecnologia, enquanto outras como discurso de ódio dependem mais de contexto e revisão humana. Cerca de 90% de todas as remoções são inicialmente indicadas por sistemas automatizados. Segundo a plataforma, os casos de apelação também mostram que o sistema é efetivo, porque poucas são revertidas após revisão humana.

A empresa informa que há pessoas nesta linha de frente 24 horas por dia, em todos os lugares do mundo, atuando em todas as etapas do processo de análise. No primeiro semestre de 2020, mais de 17 milhões de vídeos foram removidos em todo o mundo – só no Brasil, foram mais de 1,4 milhão de vídeos no mesmo período[123].

Em janeiro de 2018, o YouTube lançou a "Intelligence Desk", uma equipe que monitora notícias, redes sociais e denúncias de usuários para detectar novas tendências em conteúdos impróprios e preparar seus revisores antes que tais práticas ganhem larga escala. Entre janeiro de 2018 e junho de 2019, o número de visualizações de conteúdos posteriormente removidos pela plataforma reduziu 80%. No mesmo período, 48 atualizações foram feitas nas diretrizes do YouTube. Cerca de 10 mil pessoas trabalhavam na revisão e remoção de conteúdo em setembro de 2019[124].

123. YOUTUBE. Mitos e Fatos sobre desinformação e conteúdos impróprios no YouTube. s/d. Disponível em: <https://events.withgoogle.com/mitos-e-fatos-youtube/>. Acesso: 11 dez. 2020.

124. YOUTUBE. Inside YouTube. The Four Rs of Responsibility, Part 1: Removing harm-

Se o conteúdo violar as Diretrizes da Comunidade, ele é removido e o responsável receberá uma notificação no e-mail informando as políticas violadas e o que pode ser feito em seguida. Na primeira vez que a pessoa publica um conteúdo que viola as Diretrizes da Comunidade, o vídeo é removido, mas o canal recebe apenas uma advertência. Nas seguintes, é emitido um aviso[125]. A partir desse momento, por uma semana, o canal fica impedido de publicar vídeos, histórias ou fazer transmissões ao vivo, postar na comunidade, criar ou editar playlists e adicionar colaboradores a elas. Se o canal receber um segundo aviso no período de 90 dias após a primeira ocorrência, não poderá postar conteúdo por duas semanas. Cada aviso leva 90 dias para expirar depois de ser emitido. Três avisos em 90 dias levam à exclusão permanentemente do canal do YouTube – algo que também pode ocorrer diante de um único abuso grave, como comportamento predatório ou spam. Violações graves como a prática de discurso de ódio e de assédio podem levar ainda à suspensão de anúncios, à remoção de programas *premium* de monetização e de programas de parcerias e à perda de elegibilidade para ser recomendado no YouTube (como aparecer na *home page*, na lista de *trendings* e nas recomendações do "próximo" vídeo[126]).

A eventual exclusão do conteúdo não remove um aviso. E se o conteúdo é removido por outros motivos, como reivindicação por violação de privacidade de terceiros ou por um mandado, o usuário não recebe um aviso. Caso uma conta tenha sido impedida de usar algum dos recursos da plataforma, o YouTube também impede que a pessoa use outro canal para contornar essas penalidades.

ful content. s/d. Disponível em: <https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove>. Acesso: 11 dez. 2020.

125. GOOGLE. Conceitos básicos sobre os avisos das diretrizes da comunidade. s/d. Disponível em: <https://support.google.com/youtube/answer/2802032?hl=pt-BR>. Acesso: 11 dez. 2020.

126. GOOGLE. Responsabilidades do criador de conteúdo. s/d. Disponível em: <https://support.google.com/youtube/answer/7650329>. Acesso: 11 dez. 2020.

Em fevereiro de 2019, o YouTube modificou sua política de avisos visando deixar as regras mais claras para os usuários[127]. Na ocasião, a empresa informou que 98% dos usuários não desrespeitam as Diretrizes da Comunidade e que 94% não chegam a receber um segundo aviso. Uma das mudanças feitas no ano passado foi o estabelecimento da mesma penalidade para qualquer tipo de violação. Anteriormente, por exemplo, um primeiro aviso resultava na pena de 90 dias sem transmissões ao vivo, e somente no segundo aviso que o canal ficava impedido de publicar vídeos. Se um canal for excluído, o proprietário não pode acessar, ter ou criar outros canais no YouTube[128].

Nos Termos de Uso, o YouTube diz que qualquer conteúdo que violar o contrato, causar danos ao YouTube ou a terceiros poderá ser removido a critério da plataforma. O contrato também estabelece que o YouTube pode suspender ou encerrar o acesso ou a Conta do Google se: (a) o contrato for violado de forma significativa ou repetida; (b) for obrigado a fazê-lo para cumprir um requisito legal ou um mandado; ou (c) acreditar que houve uma conduta que cria (ou poderia criar) responsabilidade ou dano a qualquer usuário, terceiros, ao YouTube ou às suas afiliadas. Nesses casos, eles enviam uma notificação especificando o motivo da ação ao responsável pela postagem e/ou conta e com informações sobre como contestar essa decisão. Não notificam o usuário se acreditarem que isso: a) viola a legislação ou o pedido de uma autoridade legal; b) pode gerar responsabilidades legais para o YouTube ou suas afiliadas; c) prejudica uma investigação ou a integridade/operação do serviço; ou (d) prejudica qualquer usuário, terceiro, o YouTube ou suas afiliadas.

127. GOOGLE. Making our strikes system clear and consistent. Publicada em 19 de fevereiro de 2019. Disponível em: <https://youtube-creators.googleblog.com/2019/02/making-our-strikes-system-clear-and.html>. Acesso: 11 dez. 2020.

128. GOOGLE. Encerramentos de canal ou conta. s/d. Disponível em: <https://support.google.com/youtube/answer/2802168?hl=pt-BR>. Acesso: 11 dez. 2020.

A plataforma estimula que sua comunidade reporte conteúdos que considere inapropriados[129]. As denúncias são anônimas, e, quando recebidas, são avaliadas pela equipe do YouTube. A remoção não depende da quantidade de denúncias recebidas contra um conteúdo, e sim da análise feita pela empresa. O principal mecanismo disponível permite "flagear" o trecho de um vídeo apontando as diretrizes desrespeitadas, mas também é possível enviar um relatório mais detalhado sobre o conteúdo. A plataforma tem um programa de parceria[130] com indivíduos, órgãos públicos e ONGs considerados confiáveis para denunciar conteúdos que oferece a esses atores outras ferramentas, como a possibilidade de denunciar vários vídeos de uma só vez e ter acesso às decisões tomadas pela plataforma sobre eles. Os mesmos padrões de análise de conteúdo se aplicam. Entretanto, como o nível de precisão para identificar violações é maior por parte de quem participa do programa, suas denúncias são priorizadas pelos revisores do YouTube.

Opções de autopreenchimento de buscas na plataforma também podem ser denunciadas e possuem uma política própria[131], que vale para todos os serviços da empresa Google. Não são permitidos, por exemplo, autopreenchimentos associados a ódio contra grupos, assédio contra pessoas, ameaças ou que exponham informações privadas ou sensíveis.

Em entrevista à pesquisa em agosto de 2020, a empresa declarou que não tem todas as respostas sobre tudo e que confia no conhecimento de especialistas para tomar as decisões mais difíceis, sobretudo antes de alterar qualquer política ou diretriz da comunidade. Eles não são considerados imutáveis, entretanto, e fazem

129. GOOGLE. Report inappropriate content. s/d. Disponível em: <https://support.google.com/youtube/answer/2802027?hl=en&ref_topic=9387085>. Acesso: 11 dez. 2020.
130. GOOGLE. YouTube Trusted Flagger program. s/d. Disponível em: <https://support.google.com/youtube/answer/7554338?hl=en&ref_topic=9387085>. Acesso: 11 dez. 2020.
131. GOOGLE. Autocomplete policies. s/d. Disponível em: <https://support.google.com/websearch/answer/7368877>. Acesso: 11 dez. 2020.

parte de uma conversa permanente da plataforma com pesquisadores e organizações da sociedade civil sobre possibilidades de melhoria. Quem lidera este processo é o departamento de Confiança e Segurança.

Parte das informações relativas ao enfrentamento do YouTube às chamadas "fake news" foi agregada numa página especial lançada pela plataforma em outubro de 2020, intitulada "Fatos sobre desinformação: Mitos e Fatos sobre desinformação e conteúdo impróprio no YouTube"[132]. Nela, a empresa afirma que trata de desinformação "com diversas ferramentas" e que "várias políticas presentes nas diretrizes da comunidade são diretamente aplicáveis à desinformação".

2.4.2. MODERAÇÃO DE CONTEÚDO

Desde janeiro de 2019 nos Estados Unidos e a partir do final daquele ano no Brasil, o YouTube passou a reduzir o alcance do que chama de "desinformação danosa" e de "conteúdo limítrofe", deixando de recomendá-los aos usuários. Tratam-se de conteúdos que se aproximam mas não ultrapassam a linha estabelecida pelas Diretrizes da Comunidade. Assim, seguem no ar, mas como são motivos de atenção, tem seu alcance reduzido. Dentro dos exemplos dados pela plataforma para definir os dois conceitos estão a promoção de curas milagrosas para doenças sérias e alegações falsas sobre eventos como o 11 de setembro. Em entrevista a esta pesquisa, o YouTube exemplificou conteúdos limítrofes no caso da Covid-19 como porcentagens de mortalidade, taxas de contaminação ou teorias da conspiração sobre a origem do vírus, sobre as quais ninguém da comunidade científica teria certeza absoluta.

132. GOOGLE. Mitos e Fatos sobre desinformação e conteúdos impróprios no YouTube. s/d. Disponível em: <https://events.withgoogle.com/mitos-e-fatos-youtube/>. Acesso: 11 dez. 2020.

Segundo a plataforma, esses tipos de conteúdo representam, juntos, a 1% do que é assistido nos Estados Unidos (não há números públicos sobre o Brasil)[133]. Diariamente, a plataforma faz mais de 200 milhões de recomendações a seus usuários. Na ocasião do anúncio, o YouTube disse que tinha feito "centenas de mudanças" em seu sistema para melhorar a qualidade das recomendações ao longo de 2018[134]. Para analisar conteúdos limítrofes e que podem desinformar de maneira a causar danos, o YouTube trabalha com avaliadores externos (para áreas críticas, são especialistas certificados), que ajudam a "treinar" os sistemas de avaliação da empresa e diretrizes públicas de análise[135]. Cada vídeo recebe até nove opiniões de grupos diversos do ponto de vista cultural, regional, de gênero, raça, entre outros, e a partir de análises consensuais modelos de avaliação são desenvolvidos por sistemas de *machine learning*, que então ajudam a empresa a revisar centenas de milhares de vídeos por dia para decidir o que se enquadra nesse conteúdo limítrofe.

Em entrevista à revista Wire[136], em setembro de 2020, desenvolvedores do YouTube explicaram que construíram um conjunto de cerca de 40 questões para ajudar os avaliadores a definir se um conteúdo se aproxima das proibições da plataforma. A pergunta sobre se um vídeo apresenta conteúdo "intolerante para um

133. GOOGLE. The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. Publicada em 03 de dezembro de 2019. Disponível em: <https://youtube.googleblog.com/2019/12/the-four-rs-of-responsibility-raise-and-reduce.html>. Acesso: 11 dez. 2020.

134. GOOGLE. Continuing our work to improve recommendations on YouTube. Publicada em 25 de janeiro de 2019. Disponível em: <https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html>. Acesso: 11 dez. 2020.

135. GOOGLE. General Guidelines. Publicada em 20 de outubro de 2020. Disponível em: <https://static.googleusercontent.com/media/guidelines.raterhub.com/pt-BR//searchqualityevaluatorguidelines.pdf>. Acesso: 11 dez. 2020.

136. WIRED. YouTube's Plot to Silence Conspiracy Theories. Publicada em 09 de outubro de 2020. Disponível em: <https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories/>. Acesso: 11 dez. 2020.

grupo", baseado em raça, religião, orientação sexual, gênero ou origem, por exemplo, tem uma questão suplementar: "o vídeo é satírico?". Isso porque a política do YouTube proíbe discurso de ódio e mentiras sobre grupos étnicos, mas pode permitir conteúdo humorísticos sobre determinados comportamentos. Já a categoria que pergunta se um vídeo é impreciso ou enganoso é acompanhada de subcategorias como "teorias da conspiração sem fundamento", "informação comprovadamente imprecisa", "lenda urbana", "mito ou estória ficcional" e "contradição a consensos estabelecidos por especialistas".

De acordo com a reportagem, o sistema de inteligência artificial do YouTube não produz respostas binárias. Ele gera uma pontuação sobre a probabilidade de determinado vídeo se aproximar de um "conteúdo limítrofe", e isso é usado pelo algoritmo da rede social como um dos sinais para definir a recomendação ou não para um determinado usuário.

Em um ano de implementação da medida, de acordo com o YouTube, o tempo de visualização de conteúdos dessa natureza caiu 70%, em média, nos Estados Unidos. Em entrevista à *Folha de S. Paulo* em julho de 2020, a CEO da empresa, Susan Wojcicki, disse que as ações contra conteúdos limítrofes no Brasil reduziram sua circulação em 40% no país desde o início do ano[137]. Em outubro, a plataforma divulgou que, como resultado da mesma iniciativa, nos três primeiros meses de 2020 o consumo de notícias aumentou 75% no YouTube no Brasil. A empresa acredita que, assim, mantém o equilíbrio entre a liberdade de expressão e sua responsabilidade com seus usuários. Desde junho de 2019, nos Estados Unidos, se algum usuário assiste a um conteúdo considerado

137. SÁ, Nelson de. CEO do YouTube diz que esforços contra fake news vêm ganhando corpo. Folha de S. Paulo. Publicada em 5 de julho de 2020. Disponível em: <https://www1.folha.uol.com.br/ilustrada/2020/07/ceo-do-youtube-diz-que-esforcos-contra-fake-news-vem-ganhando-corpo.shtml?utm_source=newsletter&utm_medium=email&utm_campaign=newsfolha>. Acesso: 11 dez. 2020.

limítrofe, o YouTube também passa a oferecer fontes de credibilidade no painel "próximo vídeo".

Dentro da sua política de spam, golpes e práticas enganosas[138], o YouTube também proíbe o que chama de "mídia manipulada", definida como "conteúdo que foi editado ou adulterado para enganar os usuários e que possa apresentar sérios riscos de danos significativos". Entre eles, vídeos fora de contexto, para inventar eventos, com traduções incorretas que estimulem tensões geopolíticas, para fazer parecer que algum político morreu e para fraudar resultados eleitorais.

O YouTube declara adotar uma série de medidas para permitir que a plataforma seja considerada uma fonte confiável de informações e notícias sobre as eleições, ao mesmo tempo que um espaço aberto para um discurso político "saudável". Nesse contexto, remove vídeos que violem as diretrizes gerais da plataforma para conteúdos relacionados às eleições, com destaque para os vídeos objeto de manipulação técnica (como edições que tirem uma informação de contexto), além de conteúdos que podem enganar os eleitores sobre processos de votação ou sobre elegibilidade de candidatos. O YouTube também considera que sua política de exclusão de canais por falsificação da identidade de uma pessoa ou canal, por falsear seu país de origem ou sua relação com algum ente público também são medidas importantes durante processos eleitorais.

É importante mencionar, ainda, o Grupo de Análise de Ameaças (TAG, na sigla em inglês), que trabalha para identificar alvos de práticas de hackeamento e desinformação patrocinados e apoiados por governos[139], em um esforço para combater a influência de ope-

138. GOOGLE. Políticas de spam, práticas enganosas e golpes. s/d. Disponível em: <https://support.google.com/youtube/answer/2801973?hl=pt-BR>. Acesso: 11 dez. 2020.

139. GOOGLE. Protecting users from government-backed hacking and disinformation. Publicada em 26 de novembro de 2019. Disponível em: <https://blog.google/technology/safety-security/threat-analysis-group/protecting-users-government-backed-hacking-and-disinformation/>. Acesso: 11 dez. 2020.

rações coordenadas por atores internos e externos – principalmente em eleições. O TAG detecta essas ameaças e informa os usuários do Gmail, Drive e YouTube sobre riscos de segurança digital trazidos por adversários sofisticados como governos e/ou grupos ligados a eles que podem desenvolver operações para invadir os sistemas da empresa. Ao todo mais de 270 grupos de 50 países, incluindo o Brasil, são monitorados e a disseminação de desinformação coordenada é uma de suas práticas. Uma das campanhas de desinformação identificadas pelo TAG em 2019, por exemplo, foi contra o Movimento Free Papua, nas províncias da Indonésia. O YouTube excluiu 28 canais da plataforma nesse processo. A cada bimestre, são publicados relatórios sobre os resultados da ação do Grupo, que foca sua ação não em vídeos específicos, mas redes que atuam de maneira coordenada na plataforma.

A empresa aumentou os esforços de segurança para as eleições presidenciais dos Estados Unidos em 2020[140], oferecendo treinamento para mais de 3 mil autoridades eleitorais no país sobre como prevenir ataques digitais e ampliando o Programa de Proteção Avançada para usuários do Gmail. O YouTube também limitou o direcionamento de anúncios políticos na plataforma, "adotando medidas para eliminar o micro-targeting dos eleitores", passou a proibir conteúdos que minam a confiança em processos democráticos, como chamados para fraudar votações ou falsas alegações sobre resultados eleitorais, e conteúdos contendo informações hackeadas, cuja divulgação possa interferir em processos democráticos. As regras valeram para as eleições municipais no Brasil[141].

140. GOOGLE. An update on our 2020 U.S. election efforts. Publicada em 13 de agosto de 2020. Disponível em: <https://blog.google/outreach-initiatives/civics/update-our-2020-us-election-efforts/>. Acesso: 11 dez. 2020.

141. GOOGLE. Como o YouTube apoia o processo eleitoral. Publicada em 28 de outubro de 2020. Disponível em: <https://blog.youtube/intl/pt-br/news-and-events/como-o-youtube-apoia-o-processo-eleitoral/>. Acesso: 11 dez. 2020.

Diante da pandemia da Covid-19, no dia 6 de março de 2020, o CEO do Google e da Alphabet, Sundar Pichai, divulgou um comunicado[142] informando como a empresa estava ajudando no combate ao novo coronavírus. O CEO disse estar trabalhando para "remover rapidamente qualquer conteúdo que alegue prevenir o coronavírus em vez de procurar tratamento médico". Alguns anúncios buscando capitalizar em cima da epidemia passaram a ser bloqueados ("bloqueamos dezenas de milhares de anúncios nas últimas semanas", disse), incluindo os que vão de encontro a consensos científicos. O YouTube já tem uma política que impede anúncios relacionados a eventos trágicos, mas nem toda a publicidade em torno da pandemia foi restrita[143]. No dia 11, a plataforma permitiu que canais certificados pudessem receber anúncios em vídeos que discutam a pandemia e que novos produtores de conteúdo poderiam integrar este rol nas semanas seguintes. Logo depois, um segundo comunicado relatou que milhares de vídeos perigosos ou com informação enganosa sobre o coronavírus já haviam sido removidos do YouTube e que anúncios de máscaras médicas e respiradores haviam sido temporariamente bloqueados[144].

Em nova atualização, a monetização foi liberada para todos os vídeos que seguirem as diretrizes de conteúdo. Entre os vídeos relacionados ao coronavírus que não seguem as políticas e que continuam limitados ou sem anúncios estão: filmagens angustiantes, que mostrem pessoas sofrendo claramente por causa da Covid-19; desinformação médica (exemplos: conteúdo que incenti-

142. PICHAI, Sundar. Coronavirus: How we're helping. Publicada em 6 de março de 2020. Disponível em: <https://blog.google/inside-google/company-announcements/coronavirus-covid19-response/>. Acesso: 11 dez. 2020.

143. GOOGLE. Coronavirus: Monetization update on COVID-19 content. s/d. Disponível em: <https://support.google.com/youtube/answer/9803260?hl=en&ref_topic=6151248>. Acesso: 11 dez. 2020.

144. PICHAI, Sundar. COVID-19: How we're continuing to help. s/d. Publicada em 15 de março de 2020. Disponível em: <https://blog.google/inside-google/company-announcements/covid-19-how-were-continuing-to-help/>. Acesso: 11 dez. 2020.

ve testes ou exames do coronavírus que não têm base científica, declarações falsas ou sem embasamento sobre a causa da doença, promoção de medicamentos e remédios perigosos, propagação de informações sobre a origem ou o contágio da Covid-19 que não estão de acordo com o consenso científico); e pegadinhas e desafios relacionados à Covid-19, que promova atividades perigosas, como as que envolvem exposição proposital ao vírus ou que geram pânico[145].

Na ocasião, a plataforma informou, como práticas recomendadas para produzir conteúdo sobre o coronavírus, fazer checagem de fatos usando organizações confiáveis como a OMS e compartilhar conteúdo relacionado à Covid-19 "com as melhores intenções possíveis".

Em maio, a plataforma proibiu o envio de conteúdos disseminando informações médicas incorretas que contrariassem as orientações da Organização Mundial da Saúde (OMS) ou as informações médicas das autoridades locais sobre a Covid-19, tratamento, prevenção, diagnóstico e transmissão do novo vírus. Como exemplo, indicou vídeos com afirmações de que a Covid-19 não existe ou que as pessoas não morrem em decorrência da doença; que incentivem o uso de medicamentos caseiros, orações ou rituais como substitutos do tratamento médico adequado; que afirmem que há cura garantida; que algum remédio disponível atualmente previne o contágio; que a Covid-19 não é transmitida em determinados climas ou localidades; que algum grupo ou indivíduo tem imunidade ao vírus ou que não pode transmiti-lo; que o distanciamento social e o auto-isolamento não são eficazes na redução do contágio do vírus, entre outros.

A plataforma informou que poderia abrir "algumas exceções caso o objetivo do vídeo seja condenar ou combater as informa-

145. GOOGLE. Atualização sobre a monetização no conteúdo relacionado à COVID-19. s/d. Publicada em 15 de março de 2020. Disponível em: <https://support.google.com/youtube/answer/9803260?hl=pt-BR&ref_topic=6151248>. Acesso: 11 dez. 2020.

ções equivocadas que violam nossas políticas. Esse contexto precisa ficar evidente nas imagens ou no áudio do próprio vídeo. Informar esses detalhes no título ou na descrição não é suficiente".

Foi por desrespeitar diretrizes como essas que um vídeo do ideólogo Olavo de Carvalho dizendo que "o coronavírus não existe", e dois vídeos do presidente Jair Bolsonaro com informações comprovadamente falsas sobre a eficácia da cloroquina no combate à Covid-19 foram tirados do ar[146]. O mesmo ocorreu no final de julho, com um vídeo publicado pelo presidente norte-americano Donald Trump, com informações falsas sobre o uso da cloroquina para a cura da Covid-19[147]. Em entrevista à pesquisa, o YouTube reforçou que as decisões de cumprimento às diretrizes da plataforma são tomadas a partir da análise do conteúdo e independem de quem fala ou publica tais materiais.

Em outubro de 2020, a plataforma divulgou que está banindo qualquer conteúdo que sugira que uma vacina contra a Covid-19 pode matar pessoas, causar infertilidade ou ser usada para implantar microchips em quem for vacinado[148]. Em função das medidas adotadas contra conteúdo perigoso ou enganoso sobre o coronavírus, o YouTube removeu mais de 200 mil vídeos entre fevereiro e outubro de 2020.

146. G1. Facebook e Instagram removem vídeo de Jair Bolsonaro por violação de regras. G1. Publicada em: 30 mar. 2020. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2020/03/30/facebook-e-instagram-removem-video-de-jair-bolsonaro-por-violacao-de-regras.ghtml.

147. G1. Facebook remove publicação de Trump pela primeira vez por violar política de fake news sobre a Covid-19. G1. Publicada em 5 ago. 2020a. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2020/08/05/facebook-remove-publicacao-de-trump-pela-primeira-vez.ghtml.

148. HART, Roberts. YouTube Bans Covid-19 Vaccine Misinformation. Publicada em 14 de outubro de 2020. Disponível em: <https://www.forbes.com/sites/roberthart/2020/10/14/youtube-bans-covid-19-vaccine-misinformation/#252be4a968a9>. Acesso: 11 dez. 2020.

De acordo com o Relatório de Transparência divulgado em agosto de 2020[149], o segundo trimestre de 2020 foi o com maior número de conteúdos excluídos desde 2017, quando os relatórios começaram a ser publicados. O YouTube afirmou que, ao contar com a capacidade de revisão humana muito reduzida devido à pandemia, foi "forçado a fazer uma escolha entre a possível sub-execução ou a potencial aplicação excessiva". A primeira opção poderia limitar a aplicação das medidas ao que pudesse contar a capacidade de revisão humana disponível, o que manteria um alto nível de precisão, mas resultaria em menos conteúdo sendo removido, incluindo alguns que violassem as políticas da plataforma. A segunda opção – que acabou sendo escolhida para garantir que conteúdos prejudiciais saíssem rapidamente do ar – poderia resultar em vídeos que eventualmente não violassem as políticas da empresa sendo removidos por erro da tecnologia e falta da revisão humana.

O resultado foi o dobro da quantidade de vídeos removidos em relação ao trimestre anterior. De acordo com o relatório, "a decisão de aplicar um nível mais baixo de precisão, por cautela, levou a um aumento de mais de 3 vezes nas remoções de conteúdos que nossos sistemas suspeitavam estar vinculado ao extremismo violento ou potencialmente prejudicial às crianças". Os vídeos removidos sem revisão humana no período – à exceção daquele com alta taxa de certeza – não contaram para o sistema de alertas que a plataforma aplica até decidir remover uma conta por abuso na violação dos padrões da comunidade. Em decorrência dessa opção, os recursos contra remoções indevidas também cresceram e o número de vídeos restabelecidos após recurso aumentou de 25%, no primeiro trimestre de 2020, para 50% no segundo.

149. YOUTUBE. Aplicação responsável das nossas políticas durante a COVID-19. Publicada em 25 de outubro de 2020. Disponível em: <https://blog.youtube/intl/pt-br/inside-youtube/aplicacao-responsavel-das-nossas-politicas-durante-covid-19/>. Acesso: 11 dez. 2020.

2.4.3. PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Em 2017, o YouTube passou a priorizar o que chama de "vozes autorizadas", como canais de jornalismo, nas buscas por informações e nos "próximos vídeos" da plataforma. A empresa diz que, em média, em nível global, 93% dos vídeos mostrados nos 10 principais resultados da busca são de canais de alta credibilidade. Desde julho de 2018[150], se um usuário assistir ou pesquisar conteúdo jornalístico no YouTube, ele também terá exibida em sua página inicial a estante "Principais notícias". A seção inclui conteúdo relevante de fontes de notícias que seguem as Políticas de conteúdo do Google Notícias[151].

Um painel de informações com as últimas notícias ainda é exibido quando há um evento importante em curso[152]. A estante "Últimas notícias" traz vídeos do YouTube e de fontes que também seguem as Políticas de conteúdo do Google Notícias, com link direto para o site jornalístico. A plataforma também lembra aos usuários que informações sobre casos em andamento podem mudar rapidamente. A seção pode ser dispensada se o usuário não tiver interesse e foi criada com o objetivo de fortalecer fontes de informação de credibilidade dentro da plataforma. Segundo o YouTube, em 2019, houve um crescimento de 60% na visualização de canais de notícias confiáveis[153]. Desde o início de 2020, o aumento teria sido de 75%.

150. YOUTUBE. Building a better news experience on YouTube, together. Publicada em 09 de julho de 2018. Disponível em: <https://blog.youtube/news-and-events/building-better-news-experience-on>. Acesso: 11 dez. 2020.

151. GOOGLE. Políticas de conteúdo do Google Notícias. s/d. Disponível em: <https://support.google.com/news/publisher-center/answer/6204050?hl=pt-BR>. Acesso: 11 dez. 2020.

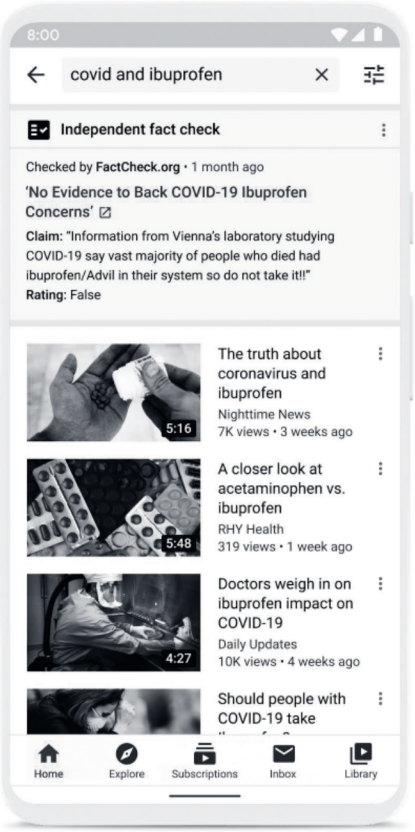
152. GOOGLE. Destaques jornalísticos no YouTube. s/d. Disponível em: <https://support.google.com/youtube/answer/9057101?hl=pt-BR&ref_topic=9257092>. Acesso: 11 dez. 2020.

153. YOUTUBE. The Four Rs of Responsibility, Part 2: Raising authoritative content and reducing borderline content and harmful misinformation. s/d. Disponível em: <https://youtube.googleblog.com/2019/12/the-four-rs-of-responsibility-raise-and-reduce.html>. Acesso: 11 dez. 2020.

Para alguns temas e/ou episódios, desde julho de 2018, o YouTube inclui painéis de informação trazendo dados de contexto extra aos vídeos publicados. São informações de fontes terceirizadas – como a Wikipedia ou a Enciclopédia Britânica –, sobre assuntos históricos e cientificamente comprovados (como a chegada do homem à lua ou a eficácia de vacinas) ou ainda sobre teorias da conspiração, que podem ser objeto de práticas de desinformação. Na avaliação da empresa, os chamados painéis de informação "ajudam o usuário a decidir a quais vídeos assistir". Desde sua introdução, mais de 3,5 bilhões de links foram disponibilizados nesses painéis.

A depender da busca realizada no YouTube, um painel com informações sobre a checagem do fato também pode ser exibido. A checagem é feita por "editores independentes" parceiros da plataforma, que informam se o conteúdo pesquisado é verdadeiro,

falso ou "parcialmente verdadeiro"[154]. O painel "Checagem de fatos independente" traz o nome do editor que fez a verificação, o fato checado, um resumo do que o editor encontrou nesse processo, link para o artigo do editor com mais informações e data de publicação do artigo da checagem de fatos. Quando houver verificações de vários editores, vários resultados são exibidos. A checagem, entretanto, não é sobre cada vídeo individualmente, mas baseada no termo utilizado pelo usuário para a sua busca.



Se nada for exibido, é porque nenhum "editor qualificado" pelo YouTube publicou um artigo relevante para a pesquisa. A busca é

154. GOOGLE. Ver checagem de fatos nos resultados da pesquisa do YouTube. s/d. Disponível em: <https://support.google.com/youtube/answer/9229632?hl=pt-BR>. Acesso: 11 dez. 2020.

baseada no sistema de tageamento do projeto Claim Review[155]. O YouTube divulga quais os critérios utilizados para credenciar checadores, mas informa que não endossa ou cria qualquer verificação exibida nos painéis de informação. Diz que, se o usuário não concordar com os dados de um artigo específico, deve entrar em contato com o site que publicou o conteúdo. O painel de checagem começou na Índia e no Brasil, em março de 2019, e só passou a funcionar para os usuários dos Estados Unidos no dia 28 de abril de 2020, em função da pandemia do coronavírus[156].

No caso das recomendações feitas no painel "próximo vídeo" e nas buscas, o YouTube afirma indicar vídeos noticiosos ou de canais de credibilidade quando o tema são eventos ou notícia sobre política. Em 2020, segundo a empresa, "milhões" de buscas e recomendações já funcionavam adotando esse tipo de ranqueamento baseado em "vozes de autoridades". A prateleira "Notícias", dentro de "Em Alta" também destaca canais jornalísticos "de qualidade" e os painéis de informação indicam as fontes de financiamento de vídeos de canais que recebem recursos públicos ou governamentais.

Nas eleições de 2018 nos Estados Unidos e para o Parlamento Europeu em 2019, quando usuários pesquisavam por um candidato/a na plataforma, o YouTube oferecia junto aos resultados um painel com informações adicionais sobre a pessoa, como afiliação partidária e distrito pelo qual concorria, e com o canal oficial da candidatura no YouTube, quando disponível. O mesmo aconteceu nas eleições de 2020 nos Estados Unidos, incluindo informações sobre como se registrar para votar[157].

155. Disponível em: <https://www.claimreviewproject.com/>; Acesso: 15 de dez. 2020.
156. YOUTUBE. Expanding fact checks on YouTube to the United States. Publicado em 28 de abril de 2020. Disponível em: <https://youtube.googleblog.com/2020/04/expanding-fact-checks-on-youtube-to-united-states.html>. Acesso: 11 dez. 2020.
157. STOREY, Amanda. An update on our 2020 U.S. election efforts. Publicada em 13 de agosto de 2020. Disponível em: <https://blog.google/outreach-initiatives/civics/update-our-2020-us-election-efforts/>. Acesso: 15 de dez. 2020.

Nas eleições municipais de 2020 no Brasil, a plataforma realizou lives em parceria com o TSE sobre a importância da checagem de fatos e informações, o ecossistema de enfrentamento às chamadas "fake news" e os mitos mais frequentes sobre eleições, como os relacionados à segurança das urnas eletrônicas.

Outra disponibilização de informações feita pelo YouTube trata dos donos dos canais. Se um canal for propriedade de uma empresa de notícias financiada por um governo ou órgão público, um painel informativo será mostrado na página de exibição dos vídeos[158]. O painel indica a origem do financiamento da empresa e fornece um link para a página da Wikipédia sobre ela. O objetivo é fornecer mais informações sobre as fontes do conteúdo jornalístico que o usuário assiste no YouTube. O painel informativo da empresa jornalística não é exibido nos resultados da pesquisa do YouTube, somente quando o usuário acessa a página daquele canal.

Em março de 2018, o Google lançou seu programa de apoio ao jornalismo na era digital, que conta com medidas para "melhorar a experiência com notícias" no YouTube. A Google News Initiative (GNI), que em dois anos apoiou cerca de 1.200 veículos na América Latina[159], foi anunciada pela empresa como uma ferramenta de combate à desinformação na medida em que apoia o desenvolvimento de novos produtos, financia projetos jornalísticos e também agências de checagem. Em julho de 2018, 25 milhões de dólares foram investidos pela GNI no YouTube, visando o estabelecimento de um grupo de trabalho para enfrentar desafios como as chamadas "fake news"; o apoio para organizações jornalísticas manterem e aprimorarem sua produção audiovisual, buscando

158. GOOGLE. Painel informativo da empresa jornalística. s/d. Disponível em: <https://support.google.com/youtube/answer/7630512?hl=pt-BR&ref_topic=9257092>. Acesso: 15 de dez. 2020.

159. ZARAGOZA, Lucia. O apoio da GNI e YouTube ao jornalismo em vídeo. Publicada em 25 de novembro de 2020. Disponível em: <https://blog.youtube/intl/pt-br/news-and-events/o-apoio-da-gni-e-youtube-ao-jornalismo-em-video/ >. Acesso: 15 de dez. 2020.

inovações para o vídeo on-line; e a oferta de treinamento e formação para editores de vídeo em todo o mundo, em temas como formato, desenvolvimento de audiência e integração técnica.

No Brasil, uma das iniciativas apoiadas é a do projeto Comprova, que recebeu mais investimentos da empresa nas eleições de 2020 para aumentar o trabalho de verificação de fatos realizado pela rede de 24 veículos de comunicação. A agência de checagem *Aos Fatos*, também apoiada pelo Google, produziu um boletim em seu canal no YouTube durante as eleições, analisando conteúdos veiculados na plataforma sobre o pleito deste ano. Uma parceria com o Insper e ProJor ofereceu um curso de formação sobre cobertura das eleições e uso de ferramentas para isso. O Google também lançou uma parceria com a agência de notícias *ÉNóis*, focada em vozes periféricas, para trazer mais diversidade para veículos de mídia e ampliar o jornalismo de qualidade na plataforma. Por fim, a empresa também patrocina o EducaMídia, um dos maiores programas de educação midiática no país, em parceria com o TSE.

Nos Estados Unidos, o Google modificou suas regras de propaganda política, para dar maior visibilidade a quem comprou anúncios eleitorais no YouTube, no Google e em outras plataformas do grupo durante o processo eleitoral deste ano[160]. Em entrevista à pesquisa, o Google informou que se trata de um esforço geral sobre transparência de anúncios e que a implementação dessas medidas no Brasil vai depender dos desafios que estão enfrentando em outros países e dos requisitos colocados no processo.

Diante de um aumento de 260% em todo o mundo nas buscas pela Covid-19, o Google lançou um "Alerta SOS" na sua ferramenta de busca na primeira semana de fevereiro, com as últimas notícias sobre o assunto, mais dicas e links de informações da Organização Mundial de Saúde (OMS). No YouTube, Pichai disse

160. MILLER, Leslie. How YouTube supports elections. Publicada em 3 de fevereiro de 2020. Disponível em: <https://youtube.googleblog.com/2020/02/how-youtube-supports-elections.html>. Acesso: 15 de dez. 2020.

que a homepage da plataforma estava sendo usada para direcionar usuários ao site da OMS e outras autoridades sanitárias e centro de controle de epidemias locais. Um painel de informações gerais sobre saúde foi lançado em fevereiro no mesmo sentido, disponibilizado ao lado de conteúdos relacionados e nas buscas por coronavírus[161]. Os painéis mostram informações em texto da OMS sobre saúde e outros conteúdos, como sintomas, prevenção e tratamento. Em abril, a Covid-19 ganhou um painel próprio, com links em idioma local para fontes de informação como ministérios da saúde e centros de controle de infecções[162]. Até o final de agosto, o painel de informações sobre a Covid-19 recebeu mais de 300 bilhões de impressões. Nos Estados Unidos, a ferramenta passou a disponibilizar um link para usuários realizarem uma autoavaliação em sua busca por tratamento médico, em parceria com o Centro de Controle de Doenças (CDC). A iniciativa deve ser expandida para outros países.

No dia 19 de março de 2020, uma sessão de notícias sobre a Covid-19 passou a ser disponibilizada na homepage do YouTube, trazendo vídeos de notícias sobre a pandemia produzidos por veículos de informação e autoridades sanitárias locais. Também em março, várias plataformas firmaram uma declaração conjunta de compromisso com o combate à desinformação sobre a pandemia: Facebook, Google, Twitter, Microsoft, Reddit, YouTube e LinkedIn. Um botão foi adicionado à guia "Explorar", levando os espectadores ao conteúdo dos criadores que estão participando da campanha #FiqueEmCasa #Comigo.

Em julho, nos Estados Unidos, novos painéis de informações sobre saúde relacionados à ansiedade e depressão foram lançados,

161. GOOGLE. Health information panels. s/d. Disponível em: <https://support.google.com/youtube/answer/9795167?hl=en&ref_topic=9257092>. Acesso: 15 de dez. 2020.
162. GOOGLE. Information panel giving topical context. s/d. Disponível em: <https://support.google.com/youtube/answer/9004474?hl=en&ref_topic=9257092>. Acesso: 15 de dez. 2020.

para ajudar as pessoas a ter acesso a dados confiáveis com mais facilidade. A iniciativa chegou ao Brasil no dia 10 de agosto[163]. Em novembro de 2020, os painéis passaram a incluir informações sobre vacinas contra a Covid-19.

Em abril de 2020, o Google informou, via GNI, a doação de 6,5 milhões de dólares para agências de checagem de fatos e organizações sem fins lucrativos que combatem a desinformação em todo o mundo[164], entre elas o projeto First Draft, a aliança CoronaVirusFacts e, no Brasil, o projeto Comprova. Os recursos também foram destinados a iniciativas que trabalham com acesso à informação, apoiam o desenvolvimento científico e criam bancos de dados para jornalistas que cobrem a pandemia. O Centro de Formação da GNI está oferecendo workshops de checagem e jornalismo de dados em 10 idiomas. Créditos em anúncios foram doados à OMS e ONGs para disseminação de informação em áreas impactadas pela pandemia. No mesmo mês, em resposta a demissões e reduções de equipes de jornalismo provocadas pela pandemia, a GNI lançou um Fundo de Apoio Emergencial ao Jornalismo[165], voltado a pequenas e médias iniciativas de informação local. A ajuda vai de alguns milhares a dezenas de milhares de dólares, a depender do tamanho dos projetos. O link para uma página com todas as atualizações tomadas pela plataforma diante da Covid-19 é exibida permanentemente no cabeçalho do YouTube[166].

163. GOOGLE. Atualizações sobre o coronavírus 2019 (COVID-19). Publicado em 17 de novembro de 2020. Disponível em: <https://support.google.com/youtube/answer/9777243?hl=pt-BR&ref_topic=6151248>. Acesso: 15 de dez. 2020.

164. MANTZARLIS, Alexios. COVID-19: $6.5 million to help fight coronavirus misinformation. Publicado em 2 de abril de 2020. Disponível em: <https://www.blog.google/outreach-initiatives/google-news-initiative/covid-19-65-million-help-fight-coronavirus-misinformation/>. Acesso: 15 de dez. 2020.

165. GINDRAS, Richard. A Global Journalism Emergency Relief Fund for local news. Publicado em 15 de abril de 2020. Disponível em: <https://www.blog.google/outreach-initiatives/google-news-initiative/global-journalism-emergency-relief-fund-local-news/>. Acesso: 15 de dez. 2020.

166. GOOGLE. Atualizações sobre o coronavírus 2019 (COVID-19). Publicado em

2.4.4. MEDIDAS CORRELATAS

Nas Diretrizes da Comunidade, o YouTube afirma não aceitar conteúdo que promova ou apoie a violência e incite o ódio contra indivíduos ou grupos com base em raça ou origem étnica, religião, deficiência, gênero, idade, nacionalidade, classe social, orientação sexual, identidade e expressão de gênero, situação de imigração ou condição de veterano de guerra[167]. Em 2017, vídeos de conteúdo supremacista deixaram de ser recomendados, compartilhados e de poder receber curtidas, o que resultou numa redução de 80% de sua visualização.

Em junho de 2019, a política para discurso de ódio foi atualizada[168] e tais vídeos passaram a ser totalmente proibidos. Conteúdos com declarações como "Que bom que [um evento violento] aconteceu. Eles mereciam isso"; "[Pessoas com tais características] são animais", exaltando ideologias como o nazismo ou negando eventos violentos bem documentados, como o Holocausto, passaram a ser proibidos. A plataforma diz que removeu, desde então, dezenas de milhares de vídeos, assim como encerrou centenas de canais com ameaças explícitas de violência[169].

Em outubro de 2020, a plataforma anunciou outro passo para conter o discurso de ódio de agressão: passou a remover mais conteúdos com teorias conspiratórias usadas com frequência para

15 de novembro de 2020. Disponível em: <https://support.google.com/youtube/answer/9777243?p=covid19_updates&visit_id=637239423136764001-2016920031&rd=1>. Acesso: 15 de dez. 2020.

167. GOOGLE. Política de discurso de ódio. s/d. Disponível em: <https://support.google.com/youtube/answer/2801939?hl=pt-BR>. Acesso: 15 de dez. 2020.

168. YOUTUBE. Our ongoing work to tackle hate. Publicado em 5 de junho de 2019. Disponível em: < https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html>. Acesso: 15 de dez. 2020.

169. BICKERT, Monica. Removing Holocaust Denial Content. Publicada em 12 out. 2020. Disponível em: <https://about.fb.com/news/2020/10/removing-holocaust-denial-content/>. Acesso em 8 dezembro 2020.

justificar atos reais de violência contra uma pessoa ou grupo específico[170], como vídeos relacionados à teoria QAnon. Conteúdos que discutem esse assunto podem continuar no ar somente se não tiverem como alvo nenhuma pessoa ou grupo protegido.

Para conteúdos que se aproximam do discurso de ódio, mas não são caracterizados como tal – na linha da política sobre conteúdo limítrofe, já descrita –, o YouTube pode limitar a disponibilização de recursos como comentários, vídeos sugeridos ou curtidas. O vídeo será mantido no ar atrás de um alerta e também perderão a possibilidade de serem monetizados.

Dentro da política contra assédio e ameaças[171], o YouTube impede a circulação de vídeos e comentários considerados abusivos, que ameacem ou intimidem pessoas, persigam um indivíduo com insultos prolongados ou maliciosos a fim de desumanizar alguém em função de seus atributos intrínsecos, violem a privacidade de uma pessoa ou incentivem repetidamente a audiência a ter comportamento abusivo/violento. Qualquer pessoa flagrada em uma dessas situações pode ser banida permanentemente do YouTube. Em dezembro de 2019, a política foi atualizada[172] para incluir neste rol: a) ameaças implícitas (conteúdos que simulem violência contra um indivíduo ou linguagem sugerindo que violência física pode ocorrer); b) comportamento abusivo de um canal (mesmo que vídeos individualmente não possam ser caracterizados assim); c) comentários tóxicos. Desde a atualização, os donos dos canais podem escolher se deixarão a ferramenta

170. YOUTUBE. Teorias conspiratórias prejudiciais no YouTube. Publicado em 15 de outubro de 2020. Disponível em: <https://blog.youtube/intl/pt-br/news-and-events/teorias-conspiratorias-prejudiciais-no-youtube/>. Acesso: 15 de dez. 2020.

171. GOOGLE. Política contra assédio e bullying virtual. s/d. Disponível em: <https://support.google.com/youtube/answer/2802268?visit_id=1-636215053151010017-1930197662&rd=1&hl=pt-BR>. Acesso: 15 de dez. 2020.

172. HALPRIN, Matt. An update to our harassment policy. Publicada em 11 de dezembro de 2019. Disponível em: <https://youtube.googleblog.com/2019/12/an-update-to-our-harassment-policy.html>. Acesso: 15 de dez. 2020.

de comentários habilitada ou não e, deixando, podem escolher quais comentários ficam visíveis.

Questionado pela pesquisa sobre a definição de "comentários tóxicos" e sua relação com culturas e realidades locais, o YouTube destacou que todas as Diretrizes da Comunidade valem tanto para vídeos quanto para comentários, que diferentes perspectivas são consideradas em todas as análises de conteúdo e que 693 milhões deles foram removidos no último bimestre. Informaram, entretanto, que precisam explicar melhor para os usuários o que entendem por "tóxico".

A plataforma recomenda que os usuários que recebem ameaças específicas e sintam que sua segurança está em risco também denunciem o problema às autoridades locais. Porém, diferencia o assédio de comportamento "irritante", diante do qual recomenda que tais pessoas sejam "ignoradas". Em 2018, por exemplo, o YouTube disponibilizou uma ferramenta para que os criadores de vídeos pudessem moderar os comentários publicados em suas páginas. Eles podem escolher se todos os comentários requerem aprovação antes da publicação ou se a medida se aplica para comentários que contenham links ou que possam conter conteúdo ofensivo.

O YouTube permite a publicação de conteúdo com assédio se o objetivo principal for de natureza educativa, documentária, científica ou artística, incluindo, por exemplo, discussões relacionadas a líderes políticos e econômicos, performances ensaiadas e educação para o assédio. Para conteúdos violentos, o YouTube reafirma a importância da análise de contexto, principalmente para conteúdos educativos ou jornalísticos, e entende que é seu dever explicar para os criadores a importância de e como fornecer essas informações aos espectadores[173].

Em entrevista à pesquisa, a plataforma citou o exemplo de filmagens feitas por jornalistas cidadãos na Síria que podem ser

173. GOOGLE. A importância do contexto. s/d. Disponível em: < https://support.google.com/youtube/answer/6345162>. Acesso: 15 de dez. 2020.

confundidas com conteúdo violento ou extremista se não vierem acompanhadas de contexto. A existência de contexto também foi o que permitiu a plataforma manter no ar vídeos de canais de notícias com trechos dos passeios do presidente Bolsonaro fazendo comentários sobre a Covid-19. Sozinhos, os vídeos foram removidos. Em conteúdos acompanhados com especialistas provendo o contexto sanitário da discussão, puderam seguir no ar.

Dentro da política de conteúdo perigoso ou nocivo[174], a plataforma não permite vídeos que incentivem atividades ilegais ou perigosas, que possam causar danos físicos graves ou de morte. Entre elas a promoção de curas, remédios, tratamentos ou substâncias nocivas.

O YouTube também proíbe, como parte de suas Diretrizes, spam, golpes e práticas enganosas[175] na descrição, no título, na miniatura dos vídeos e nas tags informadas que induzam os usuários a acreditar que o conteúdo é algo diferente do que é e visem aumentar visualizações. Não são permitidos, por exemplo, conteúdos postados de maneira repetitiva ou de forma excessiva; que direcionem o público a sites que espalham malware ou que forneçam informações incorretas sobre candidatos ou sobre como participar de eleições. Também não são autorizados spam de comentários nem comentários com o objetivo de coletar informações pessoais dos usuários. A plataforma informa em seu site que a lista não é exaustiva.

Contas criadas com o intuito de se passar por outro canal ou outra pessoa também podem ser removidas de acordo com a política de falsificação de identidade[176]. Dentro dela, além da proteção

174. GOOGLE. Política de conteúdo perigoso ou nocivo. s/d. Disponível em: <https://support.google.com/youtube/answer/2801964?hl=pt-BR>. Acesso: 15 de dez. 2020.
175. GOOGLE. Políticas de spam, práticas enganosas e golpes. s/d. Disponível em: <https://support.google.com/youtube/answer/2801973?hl=pt-BR>. Acesso: 15 de dez. 2020.
176. GOOGLE. Política de falsificação de identidade. s/d. Disponível em: <https://support.google.com/youtube/answer/2801947?hl=pt-BR>. Acesso: 15 de dez. 2020.

a marcas, uma falsificação é apontada, entre outros, quando o canal copia o perfil ou a aparência geral de outro; quando o conteúdo aparenta ter sido postado por outra pessoa; quando canais ou conteúdos causem dúvidas sobre a origem dos produtos e serviços anunciados; quando um canal é configurado usando o nome e a imagem de uma pessoa e depois passa a postar comentários em outros canais; ou quando um canal falsifica outro canal de notícias existente.

Dentro da política de envolvimento falso[177], o YouTube não permite nenhuma ação que aumente de maneira artificial o número de visualizações, marcações "Gostei", comentários ou outras métricas, seja pelo uso de sistemas automáticos ou pela reprodução de vídeos a espectadores que não queiram vê-lo. Além disso, é proibido postar conteúdo criado apenas para incentivar a interação dos espectadores. A plataforma considera o engajamento legítimo quando a intenção principal dos usuários humanos é interagir com o conteúdo sem nenhum tipo de coerção ou enganação, e sem que o único propósito do envolvimento seja conseguir ganhos financeiros. O tráfego classificado como artificial não é contabilizado no YouTube e pode resultar em avisos à conta. Contas suspensas e inscritos identificados como spam não são contabilizados no total de inscrições ou visualizações.

Em seus Termos de Uso, o YouTube restringe ainda o acesso ao serviço por meio de qualquer meio automatizado (como robôs, *botnets* ou *scrapers*), exceto: (a) no caso de mecanismos de pesquisa públicos compatíveis com o arquivo robots.txt do YouTube; ou (b) mediante permissão prévia por escrito do YouTube. A plataforma informa explicitamente que o conteúdo publicado é de responsabilidade da pessoa ou da entidade que envia o material ao serviço.

177. GOOGLE. Política de envolvimento falso. s/d. Disponível em: < https://support.google.com/youtube/answer/3399767?hl=pt-BR>. Acesso: 15 de dez. 2020.

Dentro da política de privacidade[178], a remoção de um conteúdo que traga informações pessoais ou que mostre alguém claramente identificável sem seu consentimento também pode ocorrer. O primeiro passo orientado pelo YouTube é que a pessoa atingida tente resolver o problema diretamente com o usuário que fez a postagem. Caso um acordo não seja possível, a solicitação de remoção pode ser feita via "Procedimento para Reclamação relativa à Privacidade", que considerará o interesse do público, a relevância das informações e o consentimento como fatores essenciais para a decisão.

A difamação não aparece explicitamente como uma violação às Diretrizes da Comunidade. Mas o YouTube disponibiliza um formulário[179] para envio, pela pessoa atingida, de reclamações sobre conteúdos difamatórios dentro do rol de "Ferramentas e discursos de segurança".

O Relatório de Transparência do YouTube[180] correspondente ao período de outubro a dezembro de 2019 (quarto trimestre do ano) mostra que mais de 5,8 milhões de vídeos foram removidos da plataforma. Dessas remoções, somente 542 mil resultaram de denúncias não automatizadas. Mais de 5,3 milhões de vídeos tiveram violações à política do YouTube detectadas automaticamente e 64,7% deles foram removidos antes de serem visualizados. Entre as violações praticadas estão o uso de spam, conteúdo enganoso e golpes (52% das remoções); atos perigosos ou nocivos (3,1%); conteúdo abusivo ou de incitação ao ódio (1,5%); assédio e bullying virtual (0,6%). A remoção de 108.779 vídeos foi contestada. Desses, 23.471 foram restabelecidos e voltaram ao ar.

178. GOOGLE. Como proteger sua identidade. s/d. Disponível em: <https://support.google.com/youtube/answer/2801895?hl=pt-BR>. Acesso: 15 de dez. 2020.

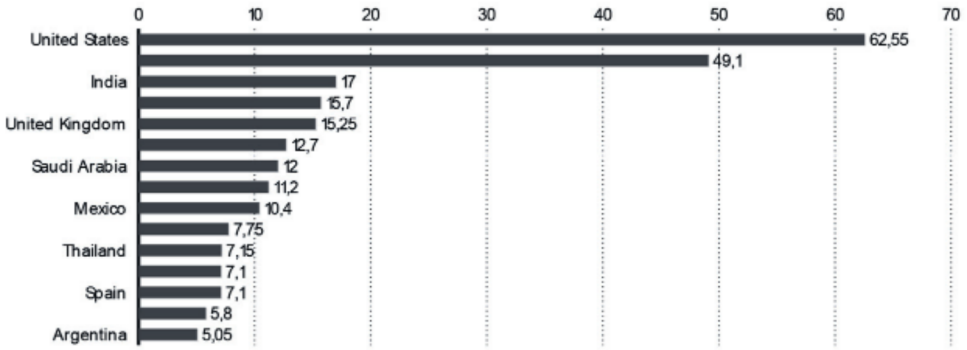
179. GOOGLE. Difamação. s/d. Disponível em: <https://www.youtube.com/reportingtool/defamation?visit_id=636215053151010017-1930197662&rd=1>. Acesso: 15 de dez. 2020.

180. GOOGLE. Cumprimento das diretrizes da comunidade do YouTube. s/d. Disponível em: <https://transparencyreport.google.com/youtube-policy/removals>. Acesso: 15 de dez. 2020.

O Brasil foi, no período, o 5º país em número de vídeos removidos (331.406), atrás de EUA, Holanda, Índia e Indonésia, e à frente da Rússia. Mais de 2 milhões de canais também foram removidos, 89,1% pela prática de spam. A falsificação de identidade levou à exclusão de 2.129 canais (0,1%) e incitação ao ódio, de 4.114 (0,2%). Já em relação aos comentários, 24% das remoções (133,3 milhões) foram por abuso ou incitação ao ódio, e 8,3% por assédio (44,6 milhões). Não há dados referentes à redução do alcance de conteúdos por desinformação.

2.5. TWITTER

Palco de declarações de políticos e de intensa discussão sobre assuntos polêmicos do cotidiano, o Twitter vem crescendo em influência, por mais que estudos mostrem que a plataforma tem perdido em número de usuários. De acordo com relatório da rede, fechou o primeiro trimestre de 2020 com 186 milhões de usuários ativos por dia. Segundo o Statista, em julho de 2020 o Brasil era o quarto país em número de usuários, com mais de 15 milhões deste total. A maior parte de líderes mundiais assim como ministros e celebridades possuem contas no Twitter, que, em muitos países, se transformou num canal estratégico de divulgação de informações oficiais no campo da política.



Fonte: Statista, 2020

2.5.1. ABORDAGEM DO FENÔMENO

Em 2017, depois das investigações realizadas pelo Congresso dos Estados Unidos sobre a influência da Rússia nas eleições norte-americanas e da forte pressão sofrida pelas plataformas digitais para que tomassem medidas diante da prática de desinformação, o Twitter publicou um pronunciamento[181] sobre o tema, no qual afirma que a empresa não deve ser considerada um "árbitro da verdade" e que a natureza aberta e em tempo real da plataforma é um "poderoso antídoto contra a disseminação de todos os tipos de informações falsas", ao permitir que jornalistas, especialistas e cidadãos engajados corrijam e questionem conteúdos em segundos. Nesse sentido, a plataforma não possui uma política para validar a autenticidade ou a precisão de um conteúdo.

Em entrevista respondida por escrito a esta pesquisa (ver anexos), o Twitter Brasil declarou que não trabalha com uma definição de desinformação, e que não tem uma política para o tema em vigor. "Ao servir a conversa pública, nosso objetivo é ajudar as pessoas a encontrar informações confiáveis no Twitter e limitar a disseminação de conteúdo potencialmente prejudicial. Em 2019, consultamos o público sobre nossa abordagem a conteúdo disputado. Ouvimos que o Twitter não deve determinar a veracidade dos Tweets e que o Twitter deve fornecer contexto para ajudar as pessoas a se decidirem nos casos em que o teor de um Tweet é contestado. Portanto, nosso foco tem sido fornecer contexto, não verificar fatos", afirmou a empresa.

O fornecimento de contexto, explica, ocorre em função do potencial de dano que uma informação representa. Assim, o foco do Twitter tem sido em possíveis danos causados na veiculação de mídias manipuladas, no que chama de "integridade cívica" (re-

181. CROWELL, Colin. Our approach to bots and misinformation. Publicado em 14 de junho de 2017. Disponível em: <https://blog.twitter.com/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html>. Acesso: 15 de dez. 2020.

lacionado sobretudo a informações sobre processos eleitorais) e em questões relacionada à Covid-19. "A probabilidade, a gravidade e o tipo de dano potencial – junto com o alcance e a escala – são levados em consideração", afirma a plataforma. Para isso, trabalha com três amplas categorias: 1) Informações enganosas – declarações ou afirmações que tenham sido confirmadas como falsas ou enganosas por especialistas no tema em questão, como autoridades em saúde pública e eleitorais; 2) Afirmações disputadas – declarações ou afirmações cuja precisão, veracidade ou credibilidade é contestada ou desconhecida; 3) Declarações não verificadas – informações (que podem ser verdadeiras ou falsas) que ainda não foram verificadas e confirmadas no momento em que são compartilhadas.

Ao mesmo tempo, em seus comunicados oficiais, o Twitter afirma que proíbe o uso de robôs – vistos como uma ferramenta positiva em diferentes setores – para a manipulação das funcionalidades da plataforma, influência nos *trending topics*, amplificação ou supressão artificial de informações, realização de disparos em massa e comportamento *spammer*. Entre os processos que afirma desenvolver de maneira permanente para evitar tais usos dos robôs está a redução da visibilidade de "tweets e contas potencialmente *spammers*" enquanto a plataforma analisa se há alguma violação de seus termos de uso. Havendo atividade suspeita nesse sentido, as contas são suspensas.

O Twitter também disse, em 2017, agir contra aplicativos que usam sua API pública para automatizar atividade na plataforma, freando robôs "potencialmente manipuladores" na origem. API (em inglês *Application Programming Interface*) é a interface de programação que permite acesso à plataforma, para viabilizar a comunicação entre dois sistemas ou plataformas distintas. O Twitter afirmou não poder revelar como faz isso para que possa ser eficiente em seu desafio de barrar a prática. Frequentemente, critica pesquisas independentes que apontam o impacto de robôs

na plataforma, dizendo que são imprecisas e metodologicamente frágeis. A empresa explica que estudos que usam sua API pública e não são revisadas por pares fazem avaliações abrangentes de comportamentos de contas usando apenas sinais públicos, como localização, conteúdo da conta, com que frequência uma conta tweeta e as contas que ela segue. De acordo com a plataforma, o trabalho preventivo realizado para desafiar contas de manipulação da plataforma (até 10 milhões de contas por semana) não é visível na pequena amostra disponível na API pública do Twitter. Tratam-se de medidas adotadas para enfrentar violações em geral às diretrizes da comunidade[182], mas que também impactam no combate à desinformação, como veremos no tópico "Medidas correlatas".

De acordo com o Twitter Brasil, a Central de Segurança da plataforma é a responsável por receber denúncias de violações às regras do Twitter e por cuidar dos métodos de aplicação adotados para combater tais violações. Mas há um conjunto de equipes envolvidas em um processo de decisão sobre novas regras a serem adotadas e sua implementação. Em 2016, o Twitter estabeleceu um Conselho de Confiança e Segurança, integrado por 40 especialistas de organizações de diversas áreas que auxiliam a empresa na definição de novos produtos e regras para a plataforma[183]. Os membros incluem advogados, acadêmicos e pesquisadores de temas como infância, educação para a mídia, cidadania digital, organizações de base e entidades comunitárias que atuam na prevenção de abusos, assédios e bullying, assim como com saúde mental e prevenção ao suicídio.

Em dezembro de 2019, a plataforma aumentou os membros do Conselho, visando incluir uma gama maior de vozes, e mu-

182. TWITTER. As Regras do Twitter. s/d. Disponível em: <https://help.twitter.com/pt/rules-and-policies/twitter-rules>. Acesso: 15 de dez. 2020.

183. CARTES, Patricia. Announcing the Twitter Trust & Safety Council. Publicado em 9 de fevereiro de 2016. Disponível em: <https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html>. Acesso: 15 de dez. 2020.

dou sua estrutura interna, para favorecer maior debate interno[184]. O Conselho passou a contar então com diferentes grupos, focados em temas como segurança e assédio on-line, direitos humanos e digitais e exploração sexual infantil. "Focar nesses assuntos nos ajudará a nos adaptarmos e a respondermos mais rapidamente a novos riscos que precisamos considerar", declarou Nick Pickes, do Departamento de Políticas Públicas da empresa, que pretende expandir sua interpretação de desumanização. O objetivo do Twitter é incluir especialistas de países e perspectivas ainda não representados no Conselho. De acordo com o Twitter Brasil, no momento existem cinco organizações da América Latina que fazem parte do Conselho de Confiança e Segurança, além de várias organizações globais que têm representação na região.

2.5.2. MODERAÇÃO DE CONTEÚDO

Como parte do esforço de conter a desinformação, "garantir conversas saudáveis e seguras", o Twitter lançou no início de 2020 uma política contra mídias (vídeos, áudios e imagens) chamadas de sintéticas ou manipuladas[185], que poderão ser marcadas num tweet ou removidas da plataforma, a depender de cada caso. As marcações tiveram início no dia 5 de março. O objetivo é informar o usuário se o conteúdo visualizado é real, alterado ou fabricado de maneira enganosa, para iludir as pessoas sobre sua autenticidade. A política inclui o enfrentamento às chamadas "deepfakes".

184. TWITTER. Fortalecendo nosso Conselho de Confiança e Segurança. Publicado em 13 de dezembro de 2019. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2019/fortalecendo-nosso-conselho-de-confianca-e-seguranca.html>. Acesso: 15 de dez. 2020.

185. TWITTER. Política de mídia sintética e manipulada. s/d. Disponível em: <https://help.twitter.com/pt/rules-and-policies/manipulated-media>. Acesso: 15 de dez. 2020.

Desde sua adoção, o Twitter recebe denúncias e/ou avalia se o conteúdo foi substancialmente editado de modo a alterar fundamentalmente sua composição; se qualquer informação visual ou de áudio (como novos frames de vídeo, áudio com *overdubbing* ou legendas modificadas) que foi adicionada ou removida; e se foram fabricadas ou simuladas mídias que mostram uma pessoa real. Quando não pode determinar de modo confiável se a mídia foi alterada, não toma providências.

A plataforma também considera se o contexto em que a mídia é compartilhada pode causar confusão ou sugere uma intenção deliberada de enganar as pessoas. Eles avaliam o contexto que acompanha a mídia (texto do tweet, metadados associados à mídia, perfil da pessoa e websites vinculados) para verificar se está claro que ela foi alterada ou fabricada. Por fim, se o conteúdo pode afetar a segurança pública ou causar danos graves imediatos, o tweet não pode ser compartilhado e está sujeito à remoção. Entre os danos considerados estão ameaças à segurança física de uma pessoa ou grupo; risco de violência em massa ou difusão de instabilidade civil; e ameaças à privacidade ou à capacidade de uma pessoa ou grupo de se expressar ou participar de eventos cívicos.

Como resposta, o Twitter pode: a) aplicar uma marcação ao conteúdo; b) exibir um aviso para as pessoas antes que elas compartilhem ou curtam o conteúdo; c) reduzir a visibilidade do conteúdo e/ou impedir que ele seja recomendado; e/ou d) fornecer um link para mais explicações ou esclarecimentos.

Previamente à implementação da política, o Twitter realizou uma consulta pública sobre seus termos que apontou[186], globalmente, que: a) 70% dos usuários considerariam inaceitável que

186. TWITTER. Construindo regras publicamente: nossa nova abordagem para mídia sintética e manipulada. Publicado em 4 de fevereiro de 2020. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2019/construindo-regras-publicamente-nova-abordagem-para-midia-sintetica-e-manipulada.html>. Acesso: 15 de dez. 2020.

a plataforma não fizesse nada diante de mídias manipuladas; b) 90% concordam com marcações de aviso em conteúdos significativamente alterados, com o recebimento de alertas antes de publicarem mídias manipuladas e com remoções em casos claros de danos intencionais; c) mais de 75% defendem que contas tendo compartilhado tais conteúdos deveriam ser obrigadas a deletá-los ou serem suspensas.

Ao divulgar um recorte da consulta sobre os Estados Unidos, a plataforma informou que apenas metade americanos que responderam à pesquisa aceitariam a remoção de mídias manipuladas em casos onde o dano intencional não esteja claro, temendo um risco de censura por parte da plataforma.

Apesar de ter maior impacto em países em que o voto não é obrigatório, o Twitter possui uma política para integridade nas eleições[187], que restringe a publicação ou o compartilhamento de conteúdo que possa diminuir o comparecimento dos eleitores às urnas (incluindo intimidações), enganar as pessoas sobre quando, onde e como votar ou informar falsas afiliações partidárias. Em abril de 2019, foi criada uma ferramenta específica para facilitar a denúncia de conteúdos que tentam enganar votações[188]. Em setembro de 2020, a política foi atualizada para passar a rotular ou remover informações falsas ou enganosas que tenham como intuito minar a confiança do público em uma eleição ou processo cívico[189]. Por exemplo, informações que podem gerar confusão

187. TWITTER. Política de integridade cívica. Publicado em outubro de 2020. Disponível em: < https://help.twitter.com/pt/rules-and-policies/election-integrity-policy>. Acesso: 15 de dez. 2020.

188. TWITTER. Strengthening our approach to deliberate attempts to mislead voters. Publicado em 24 de abril de 2019. Disponível em: <https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html>. Acesso: 15 de dez. 2020.

189. TWITTER. Expandindo nossas políticas para proteger ainda mais a conversa cívica. Publicado em 10 de setembro de 2020. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2020/expandindo-nossas-politicas-para-proteger-ainda-mais-a-conver-

sobre normas ou instituições que regem um processo eleitoral; informações não verificadas sobre fraude eleitoral; alegações enganosas sobre resultados; conteúdos que incitem conduta ilegal para impedir transições pacíficas de poder. Postagens que violem a política e forem rotulados terão visibilidade reduzida.

Nos Estados Unidos, desde 2018, o Twitter ainda exibe para seus usuários, de acordo com sua localização, as informações sobre como e onde se registrar para votar. A iniciativa #BeAVoter ("seja um eleitor") também envolve a criação de um portal de suporte específico para as autoridades eleitorais trocarem informações críticas com a plataforma o mais rápido possível[190].

O Twitter também passou a sinalizar, naquele país, as contas de candidatos/as, assim como os tweets e retweets publicados por elas[191]. No início, os candidatos precisavam manifestar interesse em terem suas contas e postagens rotuladas. Depois, o Twitter passou a notificar as candidaturas e elas tinham sete dias para negarem a rotulação antes dela ser aplicada. Os selos também foram adotados nas eleições da União Europeia no ano seguinte, durante as quais a plataforma criou um grupo interno de especialistas para conduzir o trabalho em torno da integridade eleitoral, visando identificar sobretudo ameaças de atores maliciosos[192]. Em agosto de 2018, com base nessa política, o Twitter removeu 50 contas

sa-civica.html >. Acesso: 15 de dez. 2020.

190. HARVEY, Del Harvey; ROTH, Yoel. An update on our elections integrity work. Publicado em 1 de outubro de 2018. Disponível em: <https://blog.twitter.com/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html>. Acesso: 15 de dez. 2020.

191. COYNE, Bridget. Introducing US Election Labels for Midterm Candidates. Publicado em 23 de maio de 2018. Disponível em: <https://blog.twitter.com/official/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html>. Acesso: 15 de dez. 2020.

192. WHITE, Karen. The world's window into the European election conversation. Publicado em 19 de fevereiro de 2019. Disponível em: <https://blog.twitter.com/en_us/topics/company/2019/world-window-EU-election-conversation.html>. Acesso: 15 de dez. 2020.

falsamente identificadas como de integrantes do Partido Republicano dos Estados Unidos e tweets com informações erradas sobre afiliações partidárias. Na mesma ocasião, 770 contas originárias do Irã foram removidas por comportamento coordenado tentando influenciar o debate eleitoral[193].

Em 29 de janeiro de 2020, considerando o início da campanha eleitoral para presidência nos Estados Unidos, o Twitter colocou no ar uma ferramenta para permitir aos usuários denunciarem informações enganosas sobre como participar em uma eleição e em eventos cívicos[194]. Na ocasião, a empresa reforçou que o compartilhamento de informação enganosa sobre isso viola as políticas da plataforma e que a ferramenta ajuda a identificar e a remover desinformação que possa impedir a participação em eventos como censos e eleições.

Foi com base nesta política que o Twitter marcou, em maio de 2020, duas postagens do presidente Trump como factualmente imprecisas. Elas falavam sobre suspeitas de fraude com o uso de cédulas enviadas pelo correio no processo eleitoral. Os posts foram averiguados por agências de checagem independentes e a marcação foi acompanhada de links para que os usuários pudessem se informar por outras fontes, no que o Twitter chamou de "fatos sobre votações por correio"[195]. A plataforma disse que sua prioridade é garantir a integridade do debate público sobre eleições em

193. HARVEY, Del Harvey; ROTH, Yoel. An update on our elections integrity work. Publicado em 1 de outubro de 2018. Disponível em: <https://blog.twitter.com/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html>. Acesso: 15 de dez. 2020.

194. KANE, Kevin; COYNE, Bridget. US Census search prompt launch & efforts to support Census conversations. Publicado em 11 de fevereiro de 2020. Disponível em: <https://blog.twitter.com/en_us/topics/company/2020/us-census-search-prompt.html>. Acesso: 15 de dez. 2020.

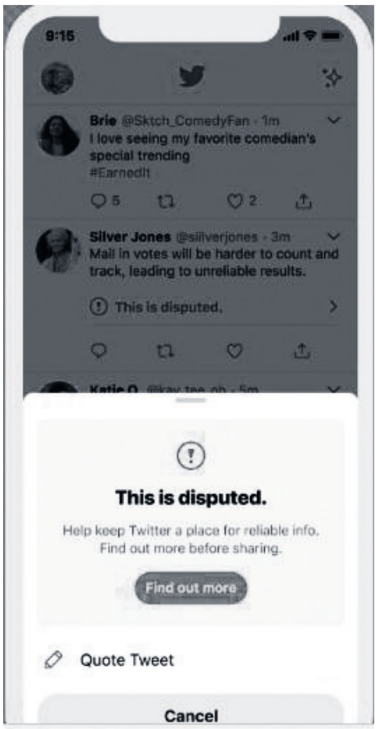
195. SANTINO, Renato. O Twitter na mira de Trump: até onde o presidente dos EUA pode interferir na rede social? Publicado em 29 de maio de 2020. Disponível em: <https://olhardigital.com.br/noticia/o-twitter-na-mira-de-trump-ate-onde-o-presidente-dos-eua-pode-interferir-na-rede-social/101486>. Acesso: 15 de dez. 2020.

todo mundo, usando para isso pessoas e tecnologias de *machine learning* para combater a desinformação.

A 50 dias antes do término das eleições nos Estados Unidos, o Twitter lançou uma *hub* (página especial) para facilitar o acesso a informações oficiais e fontes sobre o processo na plataforma. Os usuários do país puderam visualizar o *hub* no alto da aba "Explorar" da rede social, e o conteúdo estava disponível em inglês e em espanhol[196]. Em outubro, informou que rotularia tweets que declarassem falsamente a vitória de qualquer candidato – a medida de moderação direcionaria as pessoas para sua página especial – e que removeria postagens que incentivassem violência ou convidassem pessoas a interferir nos resultados das eleições e no bom funcionamento dos locais de votação[197].

A política de rotulação também foi ampliada durante as eleições dos EUA. Em primeiro lugar, passa a exibir, para quem tentasse retweetar algum post que tivesse aviso de informação enganosa, uma notificação direcionando para informações de credibilidade sobre o assunto em questão. O objetivo foi permitir que as pessoas tomassem "decisões mais informadas sobre o que elas querem ou não disseminar a quem as segue". Em segundo lugar, para rotular e restringir tweets com informações enganosas de figuras públicas, incluindo candidatos, que alcançassem engajamento significativo ou viessem de contas com mais de 100 mil seguidores. As pessoas tiveram que clicar em um alerta para visualizar esses tweets e só puderam retweetá-los com comentário. Curtidas, retweets e respostas foram desabilitados.

196. COYNE, Bridget; TOIZER, Sam. Helping you find accurate US Election News and Information. Publicado em 15 de setembro de 2020. Disponível em: https://blog.twitter.com/en_us/topics/company/2020/2020-election-news.html>. Acesso: 15 de dez. 2020.
197. BEYKPOUR, Kayvon. Passos adicionais diante das eleições nos EUA. Publicado em 9 de outubro de 2020. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2020/passos-adicionais-diante-das-eleicoes-nos-eua.html>. Acesso: 15 de dez. 2020.



Entre 27 de outubro de 11 de novembro de 2020, 300 mil posts com informações potencialmente enganosas ou controversas foram rotulados de acordo com a política de integridade cívica. O número, segundo a empresa, representa 0,2% de todos os tweets relacionados às eleições nos EUA no período. Desses, 456 foram cobertos com uma mensagem de alerta e tiveram as ferramentas de engajamento limitadas. A plataforma estima uma redução de 29% nas menções a esses tweets devido à medida. Avisos sobre a segurança do voto por correio e sobre lentidão nas apurações foram visualizados 389 milhões de vezes[198].

O Twitter informou que vai analisar a continuidade da medida de inclusão de comentários prévios aos compartilhamentos, que foi adotada globalmente no período, começando pelo Brasil, também em função da campanha eleitoral no país.

198. GADDE, Vijaya; BEYKPOUR, Kayvon. An update on our work around the 2020 US Elections. Publicado em 12 de novembro de 2020. Disponível em: <https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html>. Acesso: 15 de dez. 2020.

A plataforma afirma, em seus comunicados, que a política para integridade nas eleições é uma medida de combate à desinformação. Entretanto, segundo a empresa, não são consideradas violações a essa política conteúdos cujo uso excessivo foi constatado nas eleições de 2018 e de 2020 no Brasil, como "declarações incorretas sobre um representante público eleito, candidato ou partido político; e conteúdo orgânico polarizado, tendencioso, hiperpreconceituoso ou que contenha pontos de vista controversos sobre as eleições ou os políticos".

Sobre as eleições de 2018 no Brasil, não houve regras específicas. A plataforma organizou sessões de perguntas e respostas com os candidatos e divulgou balanços de tweets publicados, temas mais discutidos e candidatos mais populares na plataforma[199]. Num comunicado, falou das parcerias com as emissoras de TV na cobertura dos debates presidenciais e com o TSE para divulgação de resultados e da curadoria de tweets criada pela empresa para que o usuário acompanhasse "curiosidades e os assuntos mais relevantes de cada dia", assistisse ao vivo aos debates e sabatinas[200]. A plataforma também criou um canal para o recebimento de eventuais requisições a serem encaminhadas pelas autoridades eleitorais brasileiras. Os perfis oficiais dos candidatos à Presidência da República contaram com o selo azul de verificação do Twitter. E um grupo interno formado por engenheiros, advogados, pesquisadores de dados e integrantes das equipes de integridade eleitoral e políticas públicas foi criado para acompanhar o processo.

199. TWITTER. Como foi a conversa sobre o 1º turno das #Eleições2018 no Twitter. Publicado em 12 de novembro de 2020. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2018/como-foi-a-conversa-sobre-o-1-turno-das-eleicoes-2018-no-twitter.html>. Acesso: 15 de dez. 2020.

200. TWITTER. Como seguir as conversas sobre as #Eleições2018 no Twitter. Publicado em 3 de setembro de 2018. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2018/como-seguir-as-conversas-sobre-as-eleicoes2018-no-twitter.html>. Acesso: 15 de dez. 2020.

Antes da campanha de 2018 começar, o Twitter realizou uma enquete com usuários no Brasil para entender o papel da plataforma no debate político. Os resultados indicaram que quase 70% costumam acessar o Twitter com alguma regularidade para se informar sobre política, sendo que 47% usam frequentemente e 22% de vez em quando. Apenas 18% dos usuários disseram não usar o Twitter para se informar sobre política[201]. Em junho daquele ano, mais de 60% dos usuários do Twitter no Brasil acreditavam que as ideias defendidas pelos candidatos à Presidência em seus perfis oficiais na plataforma podiam contribuir com a sua decisão sobre em quem votar.

Em entrevista concedida em julho de 2020 a esta pesquisa, o Twitter informou que, como parte dos preparativos específicos para as eleições municipais, aperfeiçoaram e aplicam continuamente a tecnologia anti-spam para abordar redes de automação mal-intencionada voltadas para assuntos relacionados a eleições; que manteriam linhas de comunicação abertas com as autoridades para escalar os problemas que surgissem; monitorariam tendências e picos nas conversas relacionadas às eleições de 2020 para possíveis atividades de manipulação; e verificariam os principais candidatos e as principais contas de partidos como uma proteção contra a falsa identidade.

Por fim, vale destacar que, até o final de 2019, o Twitter também autorizava a veiculação de propaganda política durante os períodos de campanha eleitorais. As regras para veiculação de anúncios políticos tiveram início em 2017. Em abril de 2018, a plataforma firmou o "Honest Ads Act", ao lado de outras empresas e, em seguida, lançou uma política específica[202], que valeu primeiro para os Estados Unidos, depois, em março de 2019, para os países da

201. https://blog.twitter.com/pt_br/topics/company/2018/twitter-e-as-eleicoes-2018-no-brasil.html

202. https://blog.twitter.com/official/en_us/topics/company/2018/Increasing-Transparency-for-Political-Campaigning-Ads-on-Twitter.html

União Europeia, Índia e Austrália[203]. Quem fosse anunciar tinha que ter registro aprovado na plataforma – o que envolvia comprovar que estava baseado nos Estados Unidos – e os conteúdos eram rotulados.

Em junho de 2018, foi lançado um Centro de Transparência de Anúncios, no qual era possível fazer buscas nas propagandas veiculadas pelas contas ainda em funcionamento[204]. Para anúncios veiculados nos Estados Unidos, era possível consultar informações sobre pagamentos, valor investido no anúncio e dados demográficos de direcionamento da propaganda. De acordo com a plataforma, todos os anúncios veiculados no Twitter no mundo passaram a ser adicionados ao Centro, mas somente os referentes à campanha dos Estados Unidos estavam visíveis no momento do lançamento da iniciativa.

Em agosto do mesmo ano, a política foi incrementada para incluir a identificação de anúncios temáticos, que fizessem referência a uma eleição ou claramente identificassem um candidato ou que defendessem mudanças legislativas de importância nacional – como aborto, direitos civis, mudanças climáticas, armas, imigração, impostos, comércio e seguridade social[205]. Para anunciar sobre tais temas, organizações tinham que exibir um nome, foto e descrição de perfil consistente com as apresentações das organizações em outros sites e plataformas. O objetivo era garantir que não tentassem omitir sua identidade de maneira a confundir usuários de que a conta pertencia a uma organização política. Contas de órgãos de imprensa poderiam submeter uma aplicação de exceção da política, que começou em setembro de 2018, somente nos Estados Unidos.

203. https://blog.twitter.com/en_us/topics/company/2019/transparency-political-ads.html
204. https://blog.twitter.com/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html
205. https://blog.twitter.com/en_us/topics/company/2018/Announcing-new-US-issue-ads-policy.html

Em agosto de 2019, o Twitter passou a restringir, em todo o mundo, anúncios de órgãos de mídias estatais, dizendo que acreditava haver uma diferença entre dialogar com contas que o usuário escolhe seguir e receber anúncios na plataforma de contas que você não segue[206]. A mudança passou a valer para mídias financiadas ou editorialmente controladas pelo Estado, e não incluiu radiodifusores públicos. Organizações como Repórteres Sem Fronteiras, Freedom House, Comitê para a Proteção dos Jornalistas e Unesco contribuíram para o desenho dos critérios de autonomia dos veículos, independência, controle financeiro, influência e interferência sobre jornalistas e editores e pressão política externa. Em novembro, todos os anúncios políticos foram banidos da plataforma.

Desde a proibição, é possível fazer um download do arquivo com os conteúdos veiculados entre 24 de maio de 2018 e 22 de novembro de 2019. Entretanto, os detalhes da política não estão mais disponíveis no site do Twitter desde que ela foi descontinuada. O CEO da empresa, Jack Dorsey, justificou a decisão dizendo que "os anúncios políticos na Internet trazem desafios totalmente novos ao discurso cívico: *micro-targeting* e otimização de mensagens baseadas em *machine learning*, informações enganosas não verificadas e deep fakes".

Com o avanço da pandemia da Covid-19 e mais de 15 milhões de tweets sobre o tema publicados no início de 2020, antes do final de janeiro de 2021 o Twitter anunciou as medidas que passou a adotar para ajudar seus usuários a encontrar informações de credibilidade sobre o vírus[207]. Entre elas, se comprometeu a barrar "anúncios inapropriados" que busquem se "aproveitar de maneira oportunista" da pandemia. Proibiu, por exemplo, a promoção de

206. https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html

207. https://blog.twitter.com/en_us/topics/company/2020/authoritative-information-about-novel-coronavirus.html

máscaras médicas e de álcool em gel. A menção a vacinas, tratamentos e kits de teste só é permitida a título de informação, vinda por editores de notícias considerados isentos de acordo com a política de "Conteúdo de Anúncios Políticos"[208]. Ao mesmo tempo, deram créditos em anúncios para organizações sem fins lucrativos que promovam campanhas de checagem de fatos e de veiculação de informações sobre saúde.

No dia 1º de abril de 2020, ao informar que a demanda da plataforma nunca foi tão grande[209], o Twitter anunciou que aumentaria o uso de *machine learning* e de automação para coibir conteúdos potencialmente abusivos[210]. Ao admitir que tais sistemas podem carecer de contexto em suas análises, podendo levar a erros, a plataforma disse que não vai suspender permanentemente nenhuma conta apenas com base na decisão de sistemas automatizados. Em vez disso, priorizará a revisão humana para casos de maior impacto e, por isso, informou aos usuários que os retornos em caso de denúncias podem demorar mais do que o normal. A possibilidade de apelação diante de uma decisão tomada pela plataforma segue válida.

O Twitter também ampliou sua definição de "dano" para incluir conteúdos que vão diretamente de encontro às recomendações das autoridades mundiais e locais de saúde. A prioridade seguirá sendo, segundo o Twitter, remover conteúdos que façam um chamado claro a ações que possam colocar diretamente a saúde das pessoas em risco. Postagens encorajando pessoas a não praticarem o distanciamento social em áreas onde ele tiver sido recomendado, a não usarem máscara, falando de curas, diagnósticos, medidas de

208. https://business.twitter.com/pt/help/ads-policies/ads-content-policies/political-content.html

209. O número de mensagens diretas enviadas teria subido 30% desde o dia 6 de março de 2020.

210. https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html

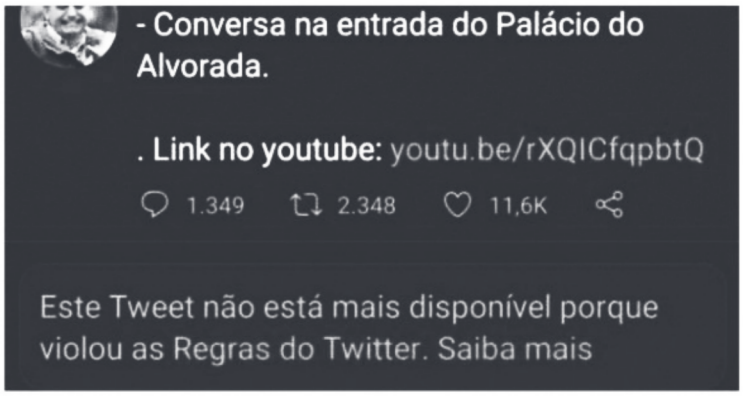
proteção sabidamente inefetivas para o vírus ou de tratamentos perigosos, postagens alegando que determinados grupos ou nacionalidades são mais ou menos susceptíveis ao vírus, negando fatos científicos sobre a transmissão do vírus durante seu período de incubação, que gerem confusão sobre autodiagnóstico, que possam gerar pânico ou distúrbio social (incluindo a destruição de infraestrutura essencial) e que se passem por mensagens oficiais de governos ou autoridades de saúde (como uma conta de paródia de alguma autoridade) serão removidas ou seus responsáveis receberão pedidos para que as removam. As decisões estão sendo tomadas em coordenação com parceiros que incluem governos e autoridades sanitárias.

O Twitter disse que colocou um "severo sistema de triagem de conteúdo", para que os tweets mais perigosos possam ser identificados e removidos. Informou, entretanto, que não terá como agir contra cada post que traga informações incompletas ou controversas sobre a Covid-19, como conteúdos que expressem esperança com estudos de potenciais remédios contra a doença.

A política para anunciantes, que já considerava inadequados para publicidade conteúdos "perigosos ou explorativos; degradantes ou provocadores; com deturpações, ataques pessoais, de mau gosto, com tópicos ou eventos delicados"[211], entre outros, também foi atualizada para a Covid-19. Publicidade contendo referências implícitas ou explícitas ao novo coronavírus só são permitidas em casos de "ajustes às práticas comerciais em resposta à Covid-19" e em "atendimento aos clientes e funcionários com relação à Covid-19". Anúncios de serviços públicos relacionados ao tema também são permitidos. No dia 17 de março, o Twitter assinou, em conjunto com outras plataformas, um comunicado em que afirma seu compromisso no combate à desinformação.

211. https://business.twitter.com/pt/help/ads-policies/ads-content-policies/inappropriate-content.html

Desde a adoção de medidas relacionadas à pandemia, o Twitter removeu mais de 14.900 postagens contendo conteúdo potencialmente danoso sobre o tema e seus sistemas questionaram 4,5 milhões de contas. As postagens em vídeo do presidente Bolsonaro passeando no Distrito Federal no dia 29 de março foram apagadas dentro dessa regra. Nos materiais, o presidente citava o uso da cloroquina como forma de tratamento da doença.



O mesmo ocorreu com postagens do senador Flávio Bolsonaro e do ministro do Meio Ambiente, Ricardo Salles, em que mostravam um vídeo do Dr. Drauzio Varella sobre o coronavírus fora de contexto. A imprensa mostrou, entretanto, no dia 8 de maio de 2020, que links para o documentário "Plandemic", de 26 minutos, repleto de teorias da conspiração relacionadas ao novo coronavírus e que chegou a ser assistido por mais de 8 milhões de pessoas, eram facilmente encontrados no Twitter.

Em julho do mesmo ano, um comunicado da plataforma trouxe mais informações e detalhes sobre como o Twitter passou a analisar informações enganosas sobre a Covid-19 para decidir a remoção ou não de postagens. Três critérios passaram a ser considerados: 1) o conteúdo faz uma afirmação sobre um fato relacionado ao coronavírus com a intenção de influenciar o comportamento de terceiros?; 2) a afirmação é comprovadamente falsa ou enganosa segundo especialistas, como autoridades públicas de saúde ou

compartilhada de maneira a causar confusão? ;3) a crença nesta informação poderia causar dano, como aumento à exposição ao vírus ou preconceito contra comunidades específicas?

Conteúdos que se enquadrarem nos três critérios podem ser removidos pela plataforma e contas que violem essa regra repetidamente podem ser suspensas. A plataforma divulgou os tipos mais comuns de afirmações que podem se enquadrar nesses critérios. Casos de conteúdos com risco de dano menos severo podem receber um rótulo ou aviso. No mesmo mês, o Twitter removeu mais um post de Donald Trump, com um vídeo publicado pelo site de notícias de direita *Breitbart* com a coletiva de imprensa de um grupo recém-criado nos Estados Unidos, chamado America's Frontline Doctors, com informações falsas sobre a hidroxocloroquina[212]. O filho do presidente também tuitou o vídeo e teve sua conta suspensa temporariamente. Novas mudanças nas normas poderão ser feitas durante a pandemia. Há uma página especial para verificar as alterações: covid19.twitter.com.

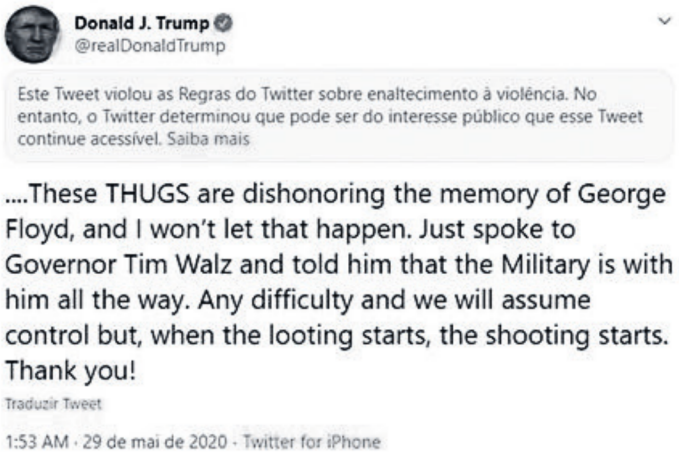
Por fim, cabe destacar que, desde 2019, o Twitter adota regras específicas quando se trata de contas e postagens do que chama de "líderes globais"[213]. A plataforma declara que quer ser um fórum onde as pessoas possam dialogar com lideranças políticas diretamente, mas diz que não deixa de aplicar suas regras gerais a esses tweets. Exceção é feita quando existe, na avaliação da empresa, um claro interesse público em manter disponíveis postagens de funcionários do governo e de potenciais e atuais detentores de cargos eletivos, em casos como: se a preservação do tweet no serviço tornará as figuras políticas responsáveis por suas declarações; se a remoção do tweet esconder um contexto ou impedir que as pessoas entendam uma questão de interesse público; se existirem ou-

212. https://www.businessinsider.com/facebook-youtube-twitter-take-down-false-covid--video-shared-trump-2020-7

213. https://blog.twitter.com/pt_br/topics/company/2019/lideres-mundiais-no-twitter-principios-e-abordagem.html

tras fontes de informação sobre essa declaração disponíveis para que o público possa se manter informado; e se o tweet fornece um contexto ou perspectiva única não disponível em outro lugar, e necessário para uma discussão mais ampla[214] [215].

A conta precisa ser verificada e ter mais de 100 mil seguidores. Uma análise do conteúdo é feita pela equipe de "Confiança & Segurança" e submetida a diretores de diferentes setores da plataforma. Considerado o interesse público, o post é disponibilizado por trás de um aviso que fornece contexto e informa a violação praticada. Não é possível interagir com o mesmo, ele não aparece em buscas e o algoritmo da plataforma deixa de recomendá-lo, "para encontrar o equilíbrio certo entre permitir a liberdade de expressão, promover a responsabilidade e reduzir o dano potencial causado por esses Tweets", explicou a plataforma, em entrevista a esta pesquisa.



Em maio de 2020, um post do presidente norte-americano, Donald Trump, considerado como violador da política contra a incitação à violência, foi mantido no ar com base a política de

214. https://help.twitter.com/en/rules-and-policies/public-interest

215. https://blog.twitter.com/pt_br/topics/company/2019/mais-clareza-sobre-conteudos-de-interesse-publico-no-twitter.html

exceções ao interesse público[216]. O interesse público é analisado em balanço com o dano potencial do conteúdo e, para alguns temas, não há admissão de interesse público possível. Entre eles, estão a promoção do terrorismo, ameaças diretas a indivíduos ou grupos, divulgação de informações privadas ou imagens íntimas. Conteúdos que reduzam o comparecimento, impeçam ou confundam as pessoas sobre quando, onde ou como votar também não se enquadram. O Twitter reconhece que decisões nesse campo são difíceis de serem tomadas e que por isso evoca muito raramente o interesse público para manter postagens (em 2018, foram menos de 5 vezes). Declara ainda estar em constante diálogo com organizações da sociedade civil e atores externos para incorporar visões locais no desenvolvimento de sua política de cumprimento das normas internas da plataforma e que respeita as leis dos países no processo de análise.

2.5.3. PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Em maio de 2019, após o surto de sarampo nos Estados Unidos, o Twitter lançou uma ferramenta intitulada #ConheçaosFatos, que passou a inserir um aviso na plataforma todas as vezes que usuários pesquisam palavras-chave associadas a vacinas[217]. O aviso direciona as pessoas a páginas de autoridades de saúde pública dos países e foi lançado simultaneamente nos Estados Unidos, Canadá, Reino Unido, Coreia do Sul, Japão, Indonésia, Singapura e em toda a América Latina, incluindo o Brasil. Aqui, a busca indica a página do Ministério da Saúde. A ferramenta já estava em operação para buscas relacionadas a suicídio e auto-

216. https://g1.globo.com/mundo/noticia/2020/05/29/twitter-marca-mensagem-de-trump-sobre-protestos-de-minneapolis-por-enaltecer-a-violencia.ghtml

217. https://blog.twitter.com/pt_br/topics/company/2019/encontre-informacoes-confiaveis-sobre-saude-publica-no-twitter.html

mutilação e a previsão era expandir para outros temas relacionados a saúde pública.

No mesmo anúncio, a plataforma informou que deixaria de sugerir buscas que pudessem direcionar os usuários a comentários e informações não-críveis sobre vacinas e a coibir anúncios com afirmações enganosas sobre curas, tratamentos, diagnóstico e prevenção de certas doenças, num esforço contra a disseminação artificial de conteúdos sem credibilidade sobre a segurança e efetividade das vacinas.

Já durante a pandemia da Covid-19, o Twitter lançou uma ferramenta de busca que apresenta fontes oficiais como primeiras respostas para quem pesquisa por #coronavírus dentro da plataforma e também bloqueou a publicação de resultados autossugeridos que possam levar as pessoas a conteúdos sem credibilidade. As medidas foram uma expansão da ferramenta #ConheçaosFatos. No dia 4 de março de 2020, uma melhoria da ferramenta foi anunciada[218], para garantir que erros de digitação na busca não impeçam o #ConheçaosFatos de ser apresentado ao usuário. O mecanismo já estava em operação em 70 países. Na ocasião, o Twitter reafirmou sua postura de "tolerância zero" para tentativas de manipulação e de abusos de seus serviços nesta conjuntura crítica, dizendo que seguirá atuando proativamente para garantir que *trends*, buscas e outras áreas de serviço estejam protegidas de comportamentos maliciosos. Em maio, adicionou uma nova aba na função #Explorar para facilitar o acesso às últimas informações sobre a pandemia. A aba inclui páginas de fontes de credibilidade, comunicados dos serviços públicos de saúde, tweets de especialistas e jornalistas em saúde e stories sobre pessoas que estão desenvolvendo ações de solidariedade. A ferramenta está disponível em 15 países, incluindo o Brasil, e os últimos números (julho 2020) informam que ela já foi visualizada por mais de 160 milhões de pessoas.

218. https://blog.twitter.com/pt_br/topics/company/2019/intensificando-nosso-trabalho-em-torno-do-coronavirus.html

No início de junho, a plataforma passou a aplicar um rótulo em alguns tweets sobre Covid-19 e 5G, encorajando os usuários a checarem informações sobre a questão. Um clique no rótulo leva para uma página com links de vários veículos de notícias com informações desconstruindo a relação entre o novo coronavírus e a tecnologia 5G. No início do ano, a teoria da conspiração resultou em diversos ataques a torres de 5G em diferentes países. No mesmo mês, iniciou o teste de uma nova ferramenta para a versão do aplicativo para Android que mostra um anúncio para usuários que retweetem um artigo sem abri-lo anteriormente. O anúncio, via o perfil de suporte da plataforma, explica o teste em curso e pergunta se o usuário não quer abrir o texto antes de retuitá-lo.

A empresa anunciou ainda a garantia de que o recurso "Eventos" contenha informações de credibilidade sobre a Covid-19 e que esteja disponível no topo da timeline de todas as pessoas. A plataforma também passou a usar seu marcador azul de verificação para identificar contas confiáveis e que possam fornecer informações de credibilidade sobre o coronavírus. A verificação está sendo feita em parceria com organizações como a OMS e vale para entidades e pessoas a elas filiadas.

No dia 24 de março, o Twitter anunciou a doação de um milhão de dólares para o Comitê para a Proteção dos Jornalistas (CPC) e para a International Women's Media Foundation[219]. Os recursos serão usados para garantir a continuidade do trabalho das organizações no apoio a jornalistas durante a crise econômica da Covid-19.

Em 29 de março de 2020, a plataforma lançou um mecanismo de acesso para permitir que desenvolvedores e pesquisadores aprovados possam estudar a conversa pública sobre a Covid-19 em tempo real[220]. Serão disponibilizados dezenas de milhões de

219. https://blog.twitter.com/pt_br/topics/company/2019/protegendo-e-apoiando-os-jornalistas-durante-a-covid19.html

220. https://blog.twitter.com/pt_br/topics/company/2019/possibilitando-estudo-da-conversa-publica-em-tempos-de-crise.html

tweets sobre a conversa pública global em torno da crise, que têm o potencial de ajudar a entender melhor a pandemia e também a disseminação de desinformação em torno dela. A escala do conjunto de dados requer experiência prévia no trabalho com dados do Twitter e ter a infraestrutura necessária para processar, armazenar e analisar os milhões de tweets gerados todos os dias. Dada a experiência e os recursos computacionais necessários para lidar com esses dados, e reconhecendo sua sensibilidade, o Twitter criou um aplicativo dedicado para acessar os conteúdos na plataforma e uma política própria para regular sua gestão.

No dia 11 de maio do mesmo ano, a plataforma informou que passará a rotular tweets e a publicar mensagens de aviso com informações adicionais, explicações e clarificações sobre situações controversas que possam trazer algum dano menos severo ou confundir o usuário[221]. A ideia é facilitar a oferta de informações factuais e o que o Twitter chama de "contexto adicional de fontes confiáveis" para que as pessoas possam tomar decisões informadas sobre a pandemia. Os rótulos levarão a uma página de informações de uma fonte confiável e seguem a linha já adotada pela empresa para mídias manipuladas, podendo ser aplicados a tweets publicados anteriormente à adoção da medida. A depender da probabilidade de dano e do tipo da informação enganosa, avisos também podem ser aplicados ao post. Antes de as pessoas verem o conteúdo do tweet, serão informadas de que ele conflita com as diretrizes de autoridades e especialistas em saúde.

O monitoramento está sendo feito de maneira proativa pelas equipes da plataforma e também por parceiros do Twitter, que ajudam a identificar conteúdos que possam causar danos. As postagens marcadas dessa maneira não terão seu alcance potencializado pelos sistemas da plataforma. O comunicado explica que, "dado o dinamismo da situação, vamos priorizar a revisão e a inclusão de

221. https://blog.twitter.com/pt_br/topics/company/2019/atualizando-nossa-abordagem-a-informacoes-enganosas.html

avisos e alertas em conteúdos que possam levar a maior exposição ou transmissão" do vírus. É possível que pessoas não logadas na plataforma visualizem esses posts sem rótulos.

Também no âmbito dos esforços de promoção de informações e transparência, durante as eleições presidenciais de 2018 e 2020 no Brasil, o Twitter apoiou institucionalmente o Projeto Comprova, que reuniu jornalistas de 24 veículos de comunicação do país para investigar informações enganosas e deliberadamente falsas durante a campanha eleitoral. Além de treinar os jornalistas para o melhor uso de ferramentas da plataforma, o Twitter ofereceu em 2018, por meio do programa "Ads For Good", créditos gratuitos para que o projeto promovesse seu conteúdo na plataforma. A plataforma também apoiou campanhas de comunicação da Justiça Eleitoral, para maximização de informações críveis e confiáveis sobre o processo eleitoral.

Em 2020, assim como nos Estados Unidos, uma página especial sobre as eleições foi disponibilizada. Outras dez páginas com recorte estadual, trazendo tweets de contas selecionadas sobre eleições nas dez maiores cidades do estado, também foram ao ar para usuários do Amazonas, Bahia, Ceará, Goiás, Minas Gerais, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul e São Paulo. Qualquer busca sobre eleições na plataforma teve como primeiro resultado um link para a página do TSE[222]. Foram realizadas sessões de treinamentos para as equipes de comunicação do sistema de Justiça Eleitoral sobre as melhores práticas do Twitter, além de treinamento para magistrados e servidores dos TREs, com explicações sobre o funcionamento da plataforma, as políticas da comunidade e aspectos práticos de contencioso eleitoral.

O Twitter também investe e apoia ações de educação para a mídia como forma de enfrentar a prática da desinformação. Entre as iniciativas estão campanhas usando hashtags específicas durante a Semana Global de Educação para a Mídia e a Informação,

222. https://blog.twitter.com/pt_br/topics/company/2020/como-acompanhar-as-eleicoes--2020-no-twitter.html

que acontece em outubro; a publicação de um manual educativo intitulado "Ensinando e aprendendo com o Twitter", disponível em 9 idiomas e voltado para educadores, com as melhores práticas de educação para a mídia apontadas pela Unesco[223]; concessão de bônus para anúncios de organizações da sociedade civil que trabalham com esse tema (Ads for Good); e participação em conferências sobre o tema.

Na Semana Global de Educação para a Mídia de 2019, o Twitter desenvolveu um conjunto de ações na Europa, como apoio a uma campanha da Autoridade de Radiodifusão da Irlanda, um workshop com estudantes franceses sobre a representação da mulher no esporte, analisando tweets de diferentes fontes, e um encontro para apoiar idosos em seu uso da plataforma no Reino Unido. No mesmo ano, por ocasião das eleições da União Europeia, funcionários do Twitter deram treinamentos e distribuíram materiais de educação midiática para partidos políticos e escritórios locais dos países membros da UE. No Brasil, apoiam campanhas do Instituto Palavra Aberta para divulgação de material educativo para consumo de mídia, e realizam reuniões, eventos e diálogos com acadêmicos e ativistas que estão debruçados sobre o tema da qualidade da informação.

Por fim, a plataforma desenvolve parcerias para que pesquisadores independentes analisem a influência, na plataforma, de atores estrangeiros em campanhas eleitorais, oriundos sobretudo de países como a Rússia, Irã e Venezuela, visando uma melhor compreensão das táticas e ameaças em curso.

223. https://blog.twitter.com/pt_br/topics/company/2019/twitter-e-unesco-lancam-guia.html

2.5.4. MEDIDAS CORRELATAS

A plataforma reconhece que, se as pessoas sofrem assédio no Twitter, sua capacidade de expressão poderá ser colocada em risco, e que há grupos de pessoas sofrem assédio on-line de maneira desproporcional, como "mulheres, pessoas de cor, lésbicas, gays, bissexuais, transexuais, homossexuais, intersexuais, indivíduos assexuados, comunidades marginalizadas e historicamente sub-representadas". Por isso, o Twitter tem uma política específica contra propagação de ódio[224], desenhada ouvindo pessoas de 30 países. Em linhas gerais, "não é permitido promover violência, atacar diretamente ou ameaçar outras pessoas com base em raça, etnia, nacionalidade, orientação sexual, sexo, identidade de gênero, religião, idade, deficiência ou doença grave" e manter contas "cuja finalidade principal seja incitar lesões a outros com base nessas categorias". Para fins deste levantamento, apontaremos os aspectos que podem ser mais diretamente relacionados com o enfrentamento à prática de desinformação.

A empresa proíbe: a) conteúdos que fazem ameaças violentas (declarações de intenção de causar ferimentos que podem resultar em lesões corporais graves ou mortes) a um alvo identificável. Quem compartilha ameaças violentas tem a conta suspensa de maneira imediata e permanente; b) conteúdo que almeja, espera, promove ou expressa um desejo de morte, lesões corporais ou doença grave em relação a uma categoria ou indivíduos dessa categoria protegida; c) direcionar a alguém conteúdo que incite o medo ou dissemine estereótipos sobre uma categoria protegida; d) declarações, apelidos, tratamentos sexistas e racistas recorrentes e/ou não consensual ou outro conteúdo com intenção de desumanizar, degradar ou reforçar estereótipos negativos. É proibida a desumanização de um grupo com base em religião ou casta (desde 2019),

224. https://help.twitter.com/pt/rules-and-policies/hateful-conduct-policy

idade, deficiência ou doença grave (desde março de 2020)[225] e raça, etnia ou origem (desde dezembro de 2020)[226]. No último caso, o processo demorou mais de um ano, e envolveu um grupo global de especialistas que considerou nuances e fatores como contexto histórico e regional[227].

Mídias que retratam imagens de propagação de ódio não são permitidas em vídeos ao vivo, na bio da conta ou em imagens de capa ou do perfil. O envio de imagens de propagação de ódio não solicitadas a um indivíduo é uma violação da política de comportamento abusivo no Twitter. A plataforma pode solicitar que a pessoa remova o conteúdo e passe por um período no modo somente leitura antes que possa tweetar novamente. Violações subsequentes podem resultar em suspensão permanente da conta.

Também não é permitido pelo Twitter se envolver no assédio direcionado a alguém nem incitar outras pessoas a fazer isso[27]. Para a plataforma, é considerado comportamento abusivo intimidar, envergonhar, rebaixar outras pessoas ou silenciar vozes. Também não são tolerados insultos "excessivamente agressivos" (expressões sexistas, racistas ou linguagem semelhante) ou conteúdo que deseje, espere ou expresse um desejo de morte, danos físicos graves ou doença fatal em relação a uma pessoa ou grupo de pessoas. Nem a publicação sem consentimento ou envio de conteúdos de nudez/adulto sem solicitação.

Em alguns casos, a empresa entende que os tweets devem ser analisados em contexto, o que envolve contatos diretos da equipe da plataforma com a pessoa afetada antes da tomada de alguma medida corretiva. Comportamentos abusivos com base em raça, etnia, nacionalidade, orientação sexual, sexo, identidade de gêne-

225. https://blog.twitter.com/pt_br/topics/company/2019/atualizando-nossas-regras-contra-conduta-de-odio.html

226. https://blog.twitter.com/pt_br/topics/company/2020/expandindo-nossas-regras-contra-propagacao-de-odio.html

227. https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

ro, religião, idade, deficiência ou doença grave podem ser analisados dentro da política contra a propagação de ódio. São analisadas denúncias tanto das pessoas envolvidas diretamente quanto de espectadores. O infrator pode receber a solicitação de excluir o conteúdo, passar por um período de somente leitura na plataforma ou ser suspendido totalmente.

Até 2018, o Twitter não tomava medidas próprias para combater conteúdos que violassem sua própria política. A plataforma apenas revisava postagens denunciadas por outros usuários. A partir de uma abordagem pró-ativa, por meio de tecnologias de automação, 20% dos tweets que demandaram alguma medida contra abusos foram, em 2018, detectados antes de qualquer denúncia. O número cresceu depois para 38% e, de acordo com o último Relatório de Transparência do Twitter, lançado em outubro de 2019[228], mais de 50% dos tweets abusivos publicados entre janeiro e junho daquele ano foram detectados tecnologicamente.

Devido à combinação da análise pró-ativa com o uso do que o Twitter chama de "dados de representação" (informações disponibilizadas pelo usuário na identificação da conta), também houve um aumento de 105% em contas bloqueadas ou suspensas por violar as regras da plataforma. Os relatórios da plataforma, entretanto, não detalham as análises feitas para o bloqueio dessas contas, tratam apenas de grandes blocos de violações. Assim, não é possível afirmar que não houve cerceamento à liberdade de expressão nesses casos. Os casos de propagação de ódio, por exemplo, tiveram um aumento de 48% de denúncias, e a plataforma acionou 133% mais contas por esse motivo em relação ao relatório anterior. No caso de conteúdo abusivo, as contas reportadas foram em número 22% maior, mas o Twitter agiu contra 68% a mais de casos.

O Twitter também afirma combater o spam e a automação mal-intencionada em grande escala. O foco está cada vez mais em

228. https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html

identificar proativamente contas e comportamentos problemáticos, por meio de ferramentas de *machine learning* que agem em redes de contas que disseminam spam ou mensagens automatizadas. Em entrevista à pesquisa, o Twitter Brasil informou que a plataforma "refinou seus sistemas de detecção e prioriza a identificação de atividades suspeitas da conta, como tweetar um volume excepcional de tweets com a mesma hashtag ou mencionar a mesmo @ sem que uma resposta da conta seja endereçada, e exige que um indivíduo usando a plataforma confirme que está no controle da conta".

A política do Twitter contra spam e manipulação da plataforma[229] é definida como "usar o Twitter para se envolver em atividades em massa, agressivas ou enganosas que induzem os outros ao erro e/ou prejudicam sua experiência". Entre as atividades vetadas estão: a) spam por motivos comerciais; b) engajamento não autêntico, que tenta fazer com que contas ou conteúdo pareçam mais populares ou ativo de que são; e c) atividade coordenada, que tenta influenciar artificialmente as conversas por meio do uso de várias contas, contas falsas, automação e/ou scripts.

A operação de contas falsas (uso de repertório, bios, fotos e informações de perfil falsos ou enganosas, incluindo a localização do perfil) é considerada uma violação dessa política. Em outubro de 2018, entraram nesse rol o uso de fotos de banco de imagens ou roubadas, bios copiadas e o uso de informações de perfil intencionalmente enganosas, incluindo a localização do perfil[230]. A atualização foi motivada pelas eleições parlamentares nos Estados Unidos. Usar/operar contas sobrepostas (operar várias contas como personas idênticas ou semelhantes ou conteúdo significativamente semelhante), contas que interagem mutuamente (para inflar ou manipular a importância de tweets) e coordenar contas para publicar conteúdo duplicado (hashtags ou tweets idênticos/signifi-

229. https://help.twitter.com/pt/rules-and-policies/platform-manipulation

230. https://blog.twitter.com/en_us/topics/company/2018/an-update-on-our-elections-integrity-work.html

cativamente semelhantes) e criar engajamento falso (retweets, curtidas, menções) também.

Na mesma atualização das regras acima citada, o Twitter informou que, se consegue vincular uma conta conhecida por violar suas normas a uma entidade, adotará medidas contra outras contas dessa mesma organização. A plataforma passou a suspender, por exemplo, contas que deliberadamente imitam ou substituem contas suspensas anteriormente. Proíbe ainda a coordenação ou recompensa (likes e retweets) de outras pessoas para entrarem em engajamento artificial, mesmo que essas pessoas usem somente uma conta.

Dentro da mesma política, não é permitido inflar artificialmente e de forma recíproca seus seguidores ou engajamentos (vender, comprar ou trocar seguidores ou engajamentos); transferir ou vender contas; seguir de maneira indiscriminada um grande número de contas não relacionadas em um curto período, especialmente de forma automatizada; e seguir e, em seguida, deixar de seguir grandes quantidades de contas com o objetivo de inflar a própria contagem de seguidores ("rotatividade de seguir").

O Twitter também restringe o uso de suas funcionalidades de outras maneiras, entre elas: a) impede o envio em massa um alto volume de respostas, menções ou mensagens diretas agressivas e não solicitadas; b) o uso de uma hashtag popular com a intenção de subverter ou manipular uma conversa para direcionar o tráfego ou a atenção para contas, websites, produtos, serviços ou iniciativas; c) a publicação de links de malware, phishing ou links enganosos.

Não são proibidos, por essa política, o uso de pseudônimos, de contas de paródia e a operação de várias contas com identidades e finalidades distintas, que podem interagir umas com as outras, desde que não violem outras regras. As consequências da violação da política dependem da gravidade e do histórico de violações anteriores. As contas podem ser limitadas temporariamente, solicitadas a fornecer informações adicionais (por exemplo, um número de celular),

receberem solicitação de exclusão de um ou mais tweets até serem suspensas definitivamente. Se a suspeita é de conta falsa, o Twitter pode solicitar o fornecimento da cópia do documento de identidade.

Em 2018, o Twitter fez mudanças no TweetDeck e em sua API que resultaram na redução em 90% das contas que usavam o TweetDeck apenas para distribuir spams e na exclusão de 142 mil aplicativos do acesso à API da plataforma. Em junho daquele ano, a plataforma anunciou quatro iniciativas para combater o spam e a automação mal-intencionada[231]:

1) Redução da visibilidade de perfis suspeitos em métricas de tweets e contas: visando combater uma forma comum de comportamento automatizado que é seguir as contas de maneira coordenada e em volume, quando um perfil for colocado em "modo de leitura" porque está agindo de forma suspeita, os números de seguidores e de engajamento da conta serão removidos até que ela passe por um teste, como confirmar um número de telefone. Um aviso passou a ser exibido em perfis em modo de leitura;

2) Aprimoramento no processo de entrada na plataforma: requisição de endereço de e-mail e/ou número de telefone para registro no Twitter;

3) Auditoria em contas existentes para analisar indícios de automação na criação de novas contas: essas contas são usadas, principalmente, para inflar números de seguidores e, em muitos casos, parecem ter seguido de forma automatizada ou em massa contas verificadas ou de influenciadores, que são sugeridas a novos usuários no fluxo de criação de novas contas;

4) Expansão dos sistemas de detecção de comportamento mal-intencionado: o Twitter passou a automatizar processos em que vê atividades suspeitas de contas, como um volume excepcionalmente alto de tweets com a mesma hashtag ou a utilização do mesmo @ sem uma resposta da conta mencionada.

231. https://blog.twitter.com/pt_br/topics/company/2018/como-o-twitter-esta-combatendo-spam-e-automacao-mal-intencionada.html

De acordo com o Relatório de Transparência de outubro de 2019[232], ao todo foram mais de 124 mil contas que sofreram sanções no período pelo fato de se passarem por outra pessoa, marca ou organização, agindo de maneira confusa ou maliciosa. Entre janeiro e março de 2019, 100 mil contas foram suspensas por tentarem recriar outras já excluídas da plataforma – um aumento de 45% em relação ao mesmo período de 2018. O número de contas suspensas em 24 horas após uma denúncia cresceu 3 vezes e as respostas a apelações foram 60% mais rápidas em função de mudanças no procedimento adotado pelo Twitter. Números brutos não foram divulgados.

Em outubro de 2018, em função de seu compromisso com o Congresso dos Estados Unidos de prover informações regulares sobre a investigação da plataforma em interferência estrangeira em debates políticos, o Twitter passou a divulgar arquivos compilados de tweets e mídias (imagens e vídeos) referentes ao que chama de "operações de informação pública apoiadas por Estados". Os arquivos contêm milhões de postagens feitas em atividades coordenadas por contas relacionadas ou apoiadas por governos que foram posteriormente removidas da plataforma, por violarem as regras do Twitter. O objetivo da divulgação é dar transparência a tentativas de manipulação do Twitter, para influenciar eleições e outros episódios cívicos, praticadas por entidades, nacionais ou estrangeiras, apoiadas por Estados, e permitir sua análise por pesquisadores em todo o mundo. "Ao tornar esses dados abertos e acessíveis, procuramos empoderar pesquisadores, jornalistas, governos e o público para aprofundar a compreensão de questões importantes que afetam a integridade da conversa pública no meio digital, especialmente em torno das eleições"[233], declarou a empresa.

232. https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html

233. https://blog.twitter.com/pt_br/topics/company/2019/operacoes-de-informacao-no-twitter-principios-processo-edivulgacao.html

A equipe de integridade da empresa utiliza ferramentas proprietárias de código aberto para identificar quando a tentativa de manipulação coordenada pode estar ocorrendo, bem como os atores responsáveis por ela. E trabalha com governos, autoridades e empresas para melhorar a compreensão dos envolvidos nas operações de informação. A equipe é formada por cientistas de dados, linguistas, analistas de políticas, cientistas políticos e especialistas técnicos em questões de abuso e antispam. O foco é identificar comportamentos enganosos e de spam e diferenciá-los do discurso legítimo de indivíduos e partidos políticos. A divulgação só em feita em casos de atribuição definitiva a Estados e a atores apoiados por eles.

Em outubro de 2018, foram divulgados dados de mais de 3.800 contas russas e de 770 contas iranianas. Em junho do ano seguinte, um novo conjunto de arquivos foi tornado público, com conteúdos de quatro contas da Rússia que teriam se envolvido nas eleições presidenciais dos EUA de 2016; 130 contas falsas da Espanha criadas para influenciar inorganicamente o movimento catalão de independência; além de outras do Irã e Venezuela. Em outubro do mesmo ano, 1.594 contas de redes do Irã, Arábia Saudita, Cuba, Tailândia e Rússia foram suspensas. Em 2020, o Twitter declarou que milhares de pesquisadores já haviam acessado os dados. Ainda não houve nenhuma divulgação sobre o Brasil.

Em agosto de 2020, a plataforma passou a incluir selos para identificar contas de membros de alto escalão de governos e contas de órgãos de comunicação estatal, seus editores chefes e diretores[234]. Incluem-se neste rol veículos que sejam total ou parcialmente financiados pelo poder público. O objetivo é trazer mais informação para o público.

Em abril de 2018, após o tiroteio na sede do YouTube nos Estados Unidos e da difusão de informações enganosas que o seguiram na plataforma, o Twitter publicou um comunicado esclarecendo

234. https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html

como analisa conteúdos durante episódios de grande repercussão[235]. Entre os critérios usados nessas ocasiões, independentemente de onde ocorram, está a análise se o conteúdo foi postado para assediar alguém, violando as regras sobre comportamento abusivo; para incitar medo contra um grupo social protegido, como descrito na política contra discurso de ódio; se deturpa a representação de alguém podendo causar dano real, de acordo com as regras sobre ameaças; se manipula a plataforma, violando a política de spam; e se o responsável pela conta já foi previamente suspenso pela plataforma.

Ao constatar tais práticas nessas ocasiões, o Twitter requisita imediatamente aos usuários a remoção das postagens e, proativamente, via automação, encaminha conteúdos e contas para revisão e impede que pessoas que já foram suspensas criem contas adicionais. Ao mesmo tempo, a sessão "Momentos" dá destaque a informações confiáveis. O mecanismo está disponível em 5 idiomas e em 16 países, incluindo o Brasil.

Em 2018, por fim, o Twitter deu início a um processo para definir e medir a "saúde" do debate público na plataforma[236]. Depois de receber 230 propostas de diferentes países, algumas foram selecionadas para receber o investimento da empresa. Pesquisadores liderados por professores das universidades de Leiden (Holanda) e de Syracuse (Estados Unidos) iniciaram um trabalho para medir como comunidades se formam em torno de discussões políticas no Twitter e os desafios que podem surgir conforme as discussões de desenvolvem. Nesse contexto, a próxima etapa da pesquisa incluirá fenômenos como as chamadas "câmaras de eco" – para mostrar até onde as pessoas se engajam no diálogo com pontos de vista diferentes do seu no Twitter – e a "incivilidade" no discurso.

235. https://blog.twitter.com/en_us/topics/company/2018/Serving-the-Public-Conversation-During-Breaking-Events.html

236. https://blog.twitter.com/pt_br/topics/company/2018/mensurando-a-saude-das-conversas.html

Os pesquisadores apontam que, enquanto a falta de "normas de polidez", apesar de problemática, tem funções importantes no diálogo político, o discurso intolerante (como discurso de ódio, racismo e xenofobia) são ameaças à democracia. A ideia é desenvolver algoritmos capazes de distinguir esses dois comportamentos. Outra iniciativa, liderada pelas universidades de Oxford (Reino Unido) e Amsterdam (Holanda), pesquisa como a exposição a diferentes perspectivas e backgrounds no Twitter pode reduzir o preconceito e a discriminação, reforçando os efeitos positivos da interação on-line. Os resultados ainda não foram divulgados nem adotados na estrutura da plataforma.

CAPÍTULO 3

ANÁLISE DAS MEDIDAS ADOTADAS PELAS PLATAFORMAS DIGITAIS

Neste capítulo, as medidas apresentadas por plataforma digital serão discutidas, tendo como marco de análise as referências apresentadas no capítulo primeiro e as categorias utilizadas na pesquisa. Por fim, será feita análise comparativa das ações.

3.1 FACEBOOK

A atenção sobre o Facebook como locus de desinformação ganhou força com as eleições dos Estados Unidos em 2016. Isso não significa que não houvesse problemas ou percepção deles antes. Mas esse episódio teve papel fundamental para jogar a plataforma no escrutínio público. As suspeitas de seu impacto sobre a eleição de Donald Trump foram denunciadas por organizações da sociedade, por políticos e em estudos acadêmicos (ALLCOTT e GENTZKOW, 2017). Gunther et al. (2018) apontaram para uma influência das notícias falsas no pleito. Por meio de entrevistas focadas em eleitores de Barack Obama em 2012, encontraram um apoio à

candidata democrata, Hillary Clinton, de 77% apenas. Além disso, 25% e 10% dos entrevistados relataram acreditar, respectivamente, em fatos falsos divulgados na campanha: uma doença de Clinton e o apoio do Papa a Donald Trump. Outros estudos, contudo, minimizaram o impacto do fenômeno no pleito.

Nelson (2017) mapeou sites conhecidos por disseminar notícias falsas e veículos tradicionais nos Estados Unidos e concluiu que a audiência destes é significativamente maior do que a dos primeiros. Na amostra analisada em novembro de 2016 – mês do pleito eleitoral-, 40,1 mil pessoas visitaram sites tradicionais, considerados pelo autor como fontes seguras de notícias, enquanto 3,1 mil acederam notícias de páginas conhecidas por conteúdos falsos. O site Buzzfeed registrou que nos meses antes daquele pleito os conteúdos falsos com melhor desempenho geraram mais engajamento do que as histórias de maior popularidade de veículos de mídia tradicional do país: 8,7 milhões de compartilhamentos contra 7,3 milhões (SILVERMAN, 2016)[237]. Naquele momento, o Facebook era a rede mais usada pela população estadunidense: 79%, contra 32% para o Instagram, 31% para o Pinterest, 29% para o LinkedIn e 24% para o Twitter (PEW RESEARCH CENTER, 2016).

A primeira reação do líder da empresa, Mark Zuckerberg, foi classificar a afirmação como uma "ideia muito louca" (SOLON, 2016)[238]. Pouco tempo depois, publicou que a empresa toma o problema da desinformação "seriamente" e estava comprometida com o seu combate. A mudança de postura não foi um ato de benevolência, mas uma resposta a pressões vindas de diversos lados. Nos Estados Unidos, foi aberta uma investigação no

237. SILVERMAN, Henry. The next phase in fighting misinformation. Publicado em: 10 abr. 2019. Disponível em: https://about.fb.com/news/2019/04/tackling-more-false-news-more-quickly/.

238. SOLON, O. (2016). Facebook's fake news: Mark Zuckerberg rejects 'crazy idea' that it swayed voters.The Guardian. Retrieved from https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-us-election-mark-zuckerberg-donald-trump.

Congresso em 2017 sobre a influência da Rússia na política do país por meios digitais, especialmente os sites de redes sociais (CALABRESI, 2017)[239], sendo o Facebook o mais questionado deles.

Outro fator de pressão sobre a companhia foi o escândalo da empresa de marketing político britânica Cambridge Analytica. Em 2018, um ex-funcionário da firma britânica veio a público mostrar como ela obteve dados de dezenas de milhões de usuários de um desenvolvedor de um app de enquete utilizado dentro do Facebook e os utilizou para influenciar a eleição dos EUA de 2016 e o referendo do Reino Unido (CADWALLADR; GRAHAM-HARISSON, 2018)[240]. No total foram manipulados dados de 87 milhões de pessoas, sendo mais de 440 mil brasileiros[241]. A empresa foi objeto de investigações nos Estados Unidos, Reino Unido e no Brasil pelo episódio, tendo sido multada nos três.

Mesmo após alguns anos de providências adotadas, a rede social ainda é reconhecida por usuários como espaço de propagação de desinformação. Segundo o *Digital News Report 2020,* do Instituto Reuters, o Facebook foi a rede social mais apontada nas Filipinas (47%), Estados Unidos (35%) e Quênia (29%), entre outros países. No Brasil, o WhatsApp foi mencionado como principal local por onde mensagens falsas são disparadas (35%), enquanto o Facebook é o segundo canal mais citado (24%). O YouTube é objeto de maior preocupação na Coreia do Sul, enquanto o

239. CALABRESI, M. Inside Russia's Social Media War on America. Time, 18 mai. 2017. Disponível em: <http://time.com/4783932/inside-russia-social-media-war-america/>. Acesso em: 04 jan. 2019

240. CADWALLADR, C. GRAHAM-HARISSON, E. Revealed:50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian, 17 mar. 2018. Disponível em: <https://www.theguardian.com/news/2018/mar/17/cambridge-analyticafacebook-influence-us-election>. Acesso em: 02 nov. 2018

241. G1. Facebook eleva para 87 milhões o nº de usuários que tiveram dados explorados pela Cambridge Analytica. G1. Publicada em: 4 abr. 2018. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/facebook-eleva-para-87-milhoes-o-n-de-usuarios-que-tiveram-dados-explorados-pela-cambridge-analytica.ghtml.

Twitter ocupou essa posição no Japão. Mais da metade (56%) dos participantes do levantamento se mostrou preocupada em como identificar o que é real e o que é falso no consumo de informações. O Brasil foi o país onde esse receio apareceu de forma mais presente (84%), seguido do Quênia (76%) e da África do Sul (72%).

3.1.1 ABORDAGEM DO FENÔMENO

Este cenário é importante para entender as medidas adotadas conforme apresentado no capítulo 2. A plataforma tomou providências nos últimos anos até configurar sua atual abordagem do fenômeno. Ela envolve ações diretas para conteúdos desinformativos, abarcando medidas de moderação de conteúdo (como a redução de alcance e remoção em casos específicos) e a promoção de informações (como a rotulagem e oferta de artigos com outras perspectivas sobre um conteúdo desinformativo).

Também compõem a estratégia da companhia o que chamamos aqui de "medidas correlatas". Entre elas estão a derrubada de contas "falsas" ou por motivos listados nos padrões da comunidade (como comportamento inautêntico) e a remoção de conteúdos que podem estar relacionados à desinformação, como discurso de ódio. Ao caracterizar sua abordagem "remover, reduzir e informar" (FACEBOOK, 2019), o Facebook reforça o papel das medidas correlatas como parte da estratégia.

A abordagem não está apresentada como uma política específica (como existe sobre dados ou sobre anúncios), mas fragmentada em diferentes políticas, no serviço de "ajuda" e em anúncios de medidas em seu site institucional. Os padrões da comunidade elencam entre conteúdos passíveis de sanções as notícias falsas (*false news*). A política de anúncios elenca a proibição de conteúdos falsos entre suas vedações. E parte importante das ações está em textos institucionais, como visto no capítulo 2. Assim, não há

um local em que as diretrizes da companhia sobre o tema estejam reunidas de forma clara aos usuários, organizações da sociedade e autoridades, um problema básico ao não cumprir essa importante prática de transparência.

Em relação ao conceito, o Facebook não divulga uma definição específica, e adota diferentes termos, como notícias falsas (tratadas tanto como *false news* quanto *fake news*) e desinformação (nominada tanto como *minsinformation* quanto *disinformation*). Nos padrões da comunidade, a explicação relativa às notícias falsas (*false news*) traz de forma genérica a abordagem e reconhece a questão como "desafiadora e delicada" (indicando a "linha tênue" entre essas mensagens, paródia e sátiras). Se, por um lado, a demarcação é complexa e demanda contexto, como dito anteriormente, a ausência do conceito é mais uma insuficiência, uma vez que não deixa claro que tipos de conteúdos serão apontados enquanto tais e moderados de acordo com as providências relacionadas a essas publicações.

3.1.2 MODERAÇÃO DE CONTEÚDO

A empresa não classifica os conteúdos como desinformativos, repassando essa atribuição para agências de checagem. Essa decisão é importante, uma vez que a evocação dessa prerrogativa conferiria um poder ainda maior à plataforma, reforçando seu papel mediador não só do debate público como da aferição de verdade no seu interior. Esses agentes vêm promovendo uma expansão da mediação das atividades sociais (VALENTE, 2019). Se já operam como controladores de fluxos informacionais e de discursos on-line, a regulação da verdade implicaria um passo adicional preocupante neste contexto.

Por outro lado, as agências de checagem não são entes neutros para essa aferição. O trabalho de verificação é operado a partir

de critérios e categorias próprios de cada entidade, muitas vezes diferentes. A definição da "verdade" é tema de longa polêmica nas ciências sociais e exatamente por isso implica uma complexidade na caracterização da desinformação. Por isso, essa prática pelas organizações de checagem não estão isentas de vieses, perspectivas e questionamentos.

Uma tentativa de uniformizar e contemplar as diversas modalidades foi a categorização de conteúdos (falso, parcialmente falso, título falso etc.). Mesmo assim, não foi possível avaliar na presente pesquisa a aplicação de marcações. Não há também nos espaços institucionais e relatórios de transparência da companhia o número de verificações realizadas e contestadas.

Contudo, há diversos casos contestados de classificações como conteúdos falsos pelas agências. Um episódio que ganhou visibilidade foi a atribuição de falsidade ao portal da Revista Fórum pela Agência Lupa ao noticiar o envio de um terço atribuído ao Papa ao ex-presidente Luiz Inácio Lula da Silva em 2018. A medida foi denunciada pelo editor do site em Comissão Geral da Câmara (BITTENCOURT, 2018)[242]. Os riscos de abuso na identificação se relacionam com possíveis impactos para a liberdade de expressão das contas nas medidas de moderação. No caso da Revista Fórum, o conteúdo foi removido, não teve apenas seu alcance reduzido.

Um relatório do grupo Popular Information (ATKIN, 2020) apontou um conteúdo negacionista sobre mudanças climáticas que foi classificado como parcialmente falso, mas teve o aviso retirado posteriormente. Conforme a organização, a empresa teria recebido o pedido de um político republicado dos Estados Unidos para retirar a classificação. Para além dos erros na checagem, denúncias

242. BITTENCOURT, Julinho. Editor da Fórum denuncia censura do Facebook em comissão da Câmara. Revista Fórum. Publicada em: 19 jun. 2018. Disponível em: https://revistaforum.com.br/politica/editor-da-forum-denuncia-censura-do-facebook-em-comissao-da-camara/. Acesso: 14 dez. 2020.

(SOLON, 2020)[243] mostraram como a companhia reagiu politicamente diante de críticas de conservadores nos Estados Unidos para afrouxar a marcação e consequente moderação de conteúdos desinformativos disseminados por páginas desses grupos no país.

Tomando os referenciais de direitos humanos discutidos na presente pesquisa, a decisão de não remoção de conteúdos falsos é acertada, pelos motivos já expostos anteriormente. Contudo, a plataforma passou a flexibilizar essa diretriz, incorporando situações em que a derrubada de mensagens passou a ser admitida, como nos riscos de violência iminente e dano físico, vacinas, temas relacionados à pandemia do novo coronavírus e mentiras sobre processos eleitorais. Entram aí temas cuja linha entre a adequação e potenciais prejuízos é tênue. Em documento com propostas para regulação de plataformas de diversas entidades latino-americanas, entre elas o Intervozes (OBSERVACOM et al., 2019), é assumida como alternativa de exclusão de publicações os riscos de danos irreparáveis.

A gravidade das notícias falsas acerca da Covid-19 – como as que minam medidas de prevenção, a exemplo do uso de máscaras ou o distanciamento social – evidenciaram exemplos de riscos de danos irreparáveis. Por outro lado, o mesmo contexto também indica a dificuldade de lidar com a moderação de conteúdos. A própria Organização Mundial da Saúde, cujas orientações são referências para a análise e moderação de conteúdos, já mudou de posição sobre determinados temas (como a obrigação de uso de máscaras pela população). Assim, se mensagens mais fantasiosas (como as que prometem a cura da Covid-19) pelo consumo de alho ou ingestão de água sanitária podem ser facilmente enquadradas, medicamentos de eficácia disputada (como a utilização de

243. SOLON, Olivia. Sensitive to claims of bias, Facebook relaxed misinformation rules for conservative pages. NBC News. Publicada em 7 ago. 2020. Disponível em: https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182.

hidroxicloroquina ou mesmo a efetividade de vacinas anunciadas) implicam em uma análise mais nuançada que não pode ser reduzida aos polos de verdadeiro ou falso.

Outro tipo de remoção admitida pelo Facebook foi em relação a anúncios. Tal decisão parece acertada, uma vez que a plataforma lucra com desinformação e, no caso de mensagens publicitárias pagas, a permissão de alegações inverídicas seria um reforço da lógica dessa prática como negócio potencializado pelo aporte de recursos de anunciantes na plataforma. O desafio é a fiscalização desses anúncios e o tratamento dos conteúdos impulsionados, além da geração de receitas pelas mensagens orgânicas viralizadas. Uma vez que a empresa não disponibiliza dados sobre o assunto, não foi possível aferir o funcionamento dessa vedação nem o nível de questionamento acerca de análises incorretas.

E aí está um problema relevante na falta de transparência da firma. Um dos pilares da desinformação é a promoção visando a viralização e a monetização decorrente dela. O Facebook alega cortar os incentivos financeiros de páginas e grupos, mas não disponibiliza informações sobre quantos perfis, páginas e publicações foram objeto desse tipo de sanção. Assim, não é possível saber se as ações anunciadas são efetivas ou não para combater a dimensão econômica da desinformação, incentivo chave a sua disseminação.

Ainda no campo da moderação de conteúdos, um tema controverso é a exceção a discursos de autoridades públicas, inclusive em anúncios políticos. O tema gerou pressão sobre o Facebook em 2019 e 2020[244]. Somente em 2020 postagens do presidente Donald Trump foram retiradas, mas por violar as regras da empresa

244. Trabalhadores da empresa divulgaram posições contrárias à permissão de alegações falsas em anúncios políticos em carta a Mark Zuckerberg em outubro de 2019. PAUL, Karin. Facebook employees 'strongly object' to policy allowing false claims in political ads. The Guardian. Publicada em 28 out. 2019. Disponível em: https://www.theguardian.com/technology/2019/oct/28/facebook-employees-strongly-object-to-policy-allowing-false-claims-in-political-ads.

para a Covid-19. Aí reside outra linha tênue entre a promoção da liberdade de expressão e a coibição de seus abusos. Se, por um lado, é importante preservar o debate político de ideias, inclusive daquelas de conteúdo mais crítico ou polêmico, por outro, as figuras públicas por sua visibilidade se tornaram grandes fontes de desinformação.

Uma vez que essa prática não é apenas um desvio, mas também um negócio e uma estratégia política (INTERVOZES, 2020), adotada sobretudo por grupos de extrema-direita, a exceção cria um ambiente propício para que agentes políticos e governantes possam aproveitar de suas redes de seguidores para propagar informações falsas como forma de defesa de suas ações e combate aos seus adversários políticos. No Brasil, essa tática vem sendo adotada de forma recorrente pelo presidente Jair Bolsonaro.

Por mais que a remoção se dê em situações específicas, a sanção da redução de alcance não pode ser vista como algo menor. Segundo o Facebook, a diminuição ocorre no patamar de 80%. Isso significa que a punição opera como limitador efetivo da liberdade de expressão não em seu direito de falar, mas de ser ouvido. Por isso, o modelo de sanções da plataforma não deve ser visto como um atenuador dos erros e abusos nos processos de verificação, uma vez que geram uma invisibilização considerável sobre contas e conteúdos marcados como desinformativos.

Uma preocupação grande presente nos referenciais de direitos humanos acerca da moderação de conteúdos são as regras de devido processo. O Facebook dispõe de poucos mecanismos para assegurar aos usuários informações adequadas, recursos e reparações. Quando uma medida é aplicada, o usuário é notificado de que houve uma violação das regras da plataforma, mas não explica o porquê. O acesso à empresa para obter essa informação é bastante dificultado, bem como a possibilidade de recurso. O usuário pode contestar, mas no caso de uma verificação, por exemplo, ele deve ir atrás da agência responsável pela checagem diretamente. Nos

materiais analisados da empresa, não ficou claro o processo de revisão nem se ele é feito necessariamente por pessoa natural.

3.1.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Na promoção de informações, a abordagem do Facebook envolve três frentes: identificar os conteúdos verificados, oferecer "mais contexto" e supostamente promover uma maior pluralidade de vozes e fortalecer o jornalismo como maneira de combater as notícias falsas no seu interior. Em relação à identificação das verificações, elas inicialmente foram rotulada de maneiras menos visíveis até a implementação de um filtro sobre a imagem da publicação, o que contribui para destacar de forma mais adequada esses conteúdos.

O tema é complexo, uma vez que há estudos com conclusões distintas sobre a eficácia da rotulagem de conteúdos. Pesquisa de acadêmicos publicada em 2017 (PENNYCOOK et al., 2017) apontou que os alertas geraram pouco impacto aos usuários e tiveram um "efeito colateral" de gerar um "efeito legitimador da verdade" sobre os demais conteúdos não rotulados, inclusive aqueles que continham desinformação mas não foram identificados enquanto tais. Já outras investigações ressaltaram os efeitos desse mecanismo. Mena (2020) entrevistou 500 pessoas e concluiu que a intenção de compartilhamento de conteúdos é menor no caso daqueles rotulados.

A despeito das discussões acerca da efetividade do recurso, parece-nos adequado não somente a adoção dos alertas como seu destaque como forma de avisar aos usuários da contestação da notícia. Considerando a prática de muitos usuários de compartilharem links sem a leitura do seu conteúdo, o recurso traz um elemento importante para evitar a difusão desses conteúdos. Além disso, a verificação (a despeito das problematizações já apontadas) torna-se um referencial de questionamento de conteúdos quando

seus partidários buscam reforçar a sua validade, tornando-se um reforço na crítica à difusão dessas mensagens nas redes sociais como o Facebook.

A segunda frente de promoção de informações é a oferta de informações de contexto. A inclusão de um botão "i" para obtenção de informações sobre veículos é positiva, mas extremamente tímida. Uma das estratégias utilizadas por promotores de desinformação é criar veículos de nomes semelhantes a mídias conhecidas. Por isso, a visibilidade e destaque da fonte da informação compartilhada são essenciais. Na organização gráfica do feed do Facebook até o momento de conclusão desta pesquisa, a fonte era o elemento com menor visibilidade, o que potencializa esse tipo de estratégia.

Também são ofertadas informações de contexto sobre um determinado fato classificado. Esse é um mecanismo complexo, uma vez que evidencia um papel editorial da plataforma na definição das informações que não só serão utilizadas como contestação como serão propagadas com um "carimbo" de discurso "verdadeiro". Entram aí as limitações do trabalho de checagem e de seu caráter não neutro. Para além disso, há um potencial ainda mais problemático caso esse "contexto adicional" implique uma escolha editorial para além das agências, como na de veículos de mídia tradicionais, que têm suas opções editoriais marcadas também por critérios políticos e que não podem ser considerados como neutros ou inquestionáveis.

Ainda na linha de informações adicionais sobre conteúdos desinformativos, as iniciativas de sensibilização e educação midiática são extremamente tímidas. A plataforma possui condições de incluir esse tipo de recurso nos feeds dos usuários de maneira muito mais recorrente do que faz, a baixíssimo custo. Se é fato que isso foi feito durante a pandemia para dar visibilidade às orientações da Organização Mundial de Saúde, tendo lançado um centro sobre o tema da pandemia, é estranho que a plataforma não lance

mão desta ferramenta como algo frequente. Ou pode ser apenas uma decisão política, como tem sido até o momento.

A terceira frente de promoção de informações envolve o que a plataforma chama de "fortalecimento do jornalismo de qualidade". O projeto de jornalismo desenvolve uma série de ações, como exemplificado no capítulo anterior. Contudo, essas são muito pouco efetivas diante dos impactos das mudanças nos critérios do *newsfeed* sobre os veículos de mídia. Em 2018, por exemplo, uma nova mudança no algoritmo trouxe alguns novos critérios de valorização das publicações (MOSSERI, 2018)[245].

Diante dos escândalos da Cambridge Analytica, uma das reações foi reduzir os conteúdos de notícias e política em favor de "interações significativas" de pessoas próximas, como família e amigos. Entre elas estavam o favorecimento de fontes consideradas "confiáveis", notícias que as pessoas tomassem como "informativas" e conteúdos de teor local. Com isso, a empresa, por um lado, favoreceu veículos locais tradicionais de mídia e, por outro, prejudicou aqueles de abrangência nacional. No Brasil, a *Folha de S. Paulo*, um dos mais importantes jornais, anunciou após a mudança que deixaria de publicar na plataforma[246].

Outra alteração estrutural que causou impacto forte sobre a circulação de jornalismo profissional na plataforma foi a decisão de derrubar o alcance orgânico para pressionar os usuários e páginas a adotar o impulsionamento de mensagens. Ao comercializar o debate público no seu interior, o Facebook dificultou a difusão orgânica de jornalismo ao encarecer o alcance pelos veículos de mídia. Essa nova realidade é especialmente problemática para

245. MOSSERI, A. News Feed Ranking in Three Minutes Flat. Facebook Newsroom, 22 mai. 2018. Disponível em:<https://newsroom.fb.com/news/2018/05/inside-feed-news-feed-ranking/>. Acesso em: 30 jul. 2018.

246. FOLHA DE S. PAULO. Folha deixa de publicar conteúdo no Facebook. Folha de S. Paulo, 8 fev. 2018. Disponível em: <https://www1.folha.uol.com.br/poder/2018/02/folha-deixa-depublicar-conteudo-no-facebook.shtml>. Acesso em: 09 jan. 2019.

mídias locais e de menor estrutura, como a comunicação pública, alternativa e comunitária.

Sobre as críticas às alterações no *newsfeed*, o argumento da empresa é o de disponibilizar ferramentas de configuração do fluxo de conteúdos que permitiriam ao usuário a seleção de fontes (com recursos como definir "ver primeiro" o conteúdo de uma determinada página)[247]. Contudo, esses recursos são limitados. Um dos mais importantes é a possibilidade de configurar o *newsfeed* para um fluxo cronológico (mais recentes), em vez da apresentação de conteúdos ser definida pela plataforma (principais histórias). Testes conduzidos pela equipe de pesquisa revelaram que o padrão (*default*) é o de principais históricas; quando o modo é alterado para as publicações mais recentes, ele é revertido pouco tempo depois automaticamente para manter a forma editada pela plataforma.

Uma quarta frente de promoção de informações que deveria compor a estratégia do Facebook, mas não o faz, é a de transparência sobre suas ações. Conforme já indicado, a política não é claramente apresentada aos usuários, pontuada de forma genérica nos parâmetros da comunidade e dispersa em uma série de comunicados institucionais. O trabalho da presente pesquisa em sistematizar essas diretrizes e medidas é um indicador da dificuldade de encontrá-la de forma clara e organizada em algum canal institucional.

É sintomático, por exemplo, que as medidas de moderação empregadas pela rede social não estejam presentes no Relatório de Fiscalização dos Padrões da Comunidade[248]. Assim, como indicado anteriormente, não há disponibilização de informações sobre a quantidade de conteúdos verificados como falsos (e como as

247. Informações sobre o *newsfeed* são disponibilizadas pela empresa aqui: https://www.facebook.com/help/www/964154640320617/?helpref=hc_fnav.

248. O relatório está disponível em: https://transparency.facebook.com/community-standards-enforcement.

demais categorias), as medidas adotadas e seus quantitativos, os pedidos de recurso, os procedimentos de revisão e eventuais providências reparatórias implementadas pela plataforma.

A insuficiência e o caráter precário são perceptíveis quando comparados, por exemplo, com as políticas de privacidade e de uso de dados, uma vez que a plataforma anunciou esforços para tornar as configurações de privacidade mais simples e fáceis de encontrar[249]. O mesmo não ocorre com as políticas relacionadas à desinformação. Em primeiro lugar, pela não existência de uma política própria. Em segundo lugar, pela inexistência de um local que a reúna.

3.1.4 MEDIDAS CORRELATAS

Como visto, o Facebook elenca em seu discurso medidas correlatas como parte de sua estratégia contra a desinformação. A remoção de contas é assinalada como um dos pilares da abordagem da plataforma. Como descrito, foram diversas as iniciativas de exclusão de perfis e páginas. Em julho de 2020, por exemplo, o Facebook derrubou contas vinculadas ao PSL, ex-partido do presidente Bolsonaro[250].

Entretanto, essas ações continuam ocorrendo a partir de conceituações problemáticas. A base delas é o chamado "comportamento inautêntico" e "comportamento inautêntico coordenado". Esse conceito é deveras vasto e genérico, podendo comportar avaliações diversas e dar à equipe da plataforma um guarda-chuva para a remoção de contas de forma pouco transparente.

249. EGAN, Erin. BERINGER, Ashlie. It's Time to Make Our Privacy Tools Easier to Find. Facebook blog. Publicada em: 28 mar. 2018. Disponível em: https://about.fb.com/news/2018/03/privacy-shortcuts/. Acesso: 14 dez. 2020.

250. GOMES, Helton. Facebook derruba rede de contas ligadas ao clã Bolsonaro e deputados do PSL. UOL. Publicada em: 8 jul. 2020. Disponível em: https://www.bol.uol.com.br/noticias/2020/07/08/facebook-derruba-rede-de-contas-ligadas-ao-cla-bolsonaro.htm. Acesso: 14 dez. 2020.

O Intervozes defende que a retirada se dê pelo comportamento e não pelo conteúdo. O que se problematiza aqui é a apresentação da derrubada de contas por comportamento inautêntico coordenado como estratégia de combate à desinformação, já que não há análise de conteúdo, portanto não é possível verificar a verdade ou não da postagem.

Em 2019, a Central Única dos Trabalhadores (CUT) teve oito páginas ligadas a ela e a sindicatos filiados tiradas do ar no mesmo dia[251]. A entidade denunciou o fato e entrou em contato com a empresa buscando uma explicação, não tendo obtido êxito. As páginas voltaram ao ar mas os responsáveis não souberam o porquê da suspensão.

O fato de que a maioria das exclusões anunciadas nos últimos anos ter sido relacionada a partidos e grupos políticos conservadores e notoriamente difusores de desinformação não retira o caráter problemático da categoria utilizada para a suspensão de perfis. Somada à ausência de mecanismos democráticos de devido processo, o uso do comportamento inautêntico coordenado é uma abertura para arbítrios perigosos e possíveis impactos negativos à liberdade de expressão.

A mesma preocupação pode ser aplicada à possibilidade de declaração de "*personas non gratas*" e sanções definitivas de exclusões de determinados grupos. Se tal atribuição é acertada no caso de organizações supremacistas brancas cuja atuação é fortemente voltada à difusão de ódio ou a grupos organizados para contestar métodos eficazes de prevenção à Covid-19, é preciso olhar atento para verificar como isso pode ser usado para organizações que venham a ser vistas como um problema para a plataforma. Um extenso relatório publicado pelo jornal *New York Times* (FRENKEL

251. ARAGÃO, Érica. Facebook tira do ar cinco páginas de entidades ligadas à CUT no mesmo dia. CUT. Publicada em 16 out. 2019. Disponível em: https://www.cut.org.br/noticias/facebook-tira-do-ar-cinco-paginas-de-entidades-ligadas-a-cut-no-mesmo-dia-65f3. Acesso: 14 dez. 2020.

et al., 2018) denunciou estratégias da companhia de ataque e tentativa de desacreditar críticos ao buscar associá-los a atitudes antissemitas.

Para além das derrubadas das contas, é por meio das atitudes vedadas nos padrões da comunidade (cuja prática implica a derrubada de mensagens) que podem ocorrer erros e abusos. Conforme pontuado antes, há uma linha tênue entre conteúdos de fato problemáticos cuja remoção é justificada (como atos de violência explícita ou incitação a ela) e outros em que essa classificação pode ensejar distintas interpretações.

Se, por um lado, o discurso de ódio pode e deve ser coibido, por outro, ao demandar uma análise de conteúdo e contexto das mensagens, pode haver diferentes interpretações, contestações e necessidade de revisão das medidas. São recorrentes as denúncias da comunidade LGBTQI+ de que o Facebook derruba postagens da própria comunidade, por automaticamente identificar termos que podem ser usadas por elas ou por grupos LGBTfóbicos como relacionados a discurso de ódio. É exatamente por essa dificuldade em alguns casos que os procedimentos transparentes e democráticos são necessários, o que, como já discutido, não ocorre no Facebook. Esses riscos são ainda maiores diante da dependência crescente de sistemas automatizados para o processo de análise de conteúdo da plataforma. Se, por um lado, há um volume considerável de publicações a serem analisadas (um desafio de fato), por outro, o aprofundamento do papel dos sistemas automatizados utilizando aprendizagem por máquina cria maior espaço para equívocos.

O relatório de transparência do Facebook indicou que, no 1º trimestre de 2020, 88% dos conteúdos enquadrados como discurso de ódio foram marcados pelos sistemas da própria plataforma antes das denúncias de usuários. Há dois anos esse índice era de 38%. Naquele trimestre, 10 milhões de posts sofreram medidas de moderação, dos quais 1,3 milhão foi objeto de recursos, uma taxa

de 13%. Mais de 1 milhão de conteúdos questionados não é uma taxa pequena, a despeito do percentual.

Outro ponto de preocupação acerca da insuficiência das respostas são os processos eleitorais. Enquanto o Facebook tentava mostrar que atuava para impedir práticas abusivas e nocivas em eleições, novas denúncias colocaram em dúvida as intenções da companhia. Matéria do *BuzzFeed* revelou um memorando interno da cientista de dados Sophie Zhang em que ela relatava ter tomado decisões com impactos em processos políticos em diversos países. "Tenho sangue nas mãos. Eu pessoalmente, sem supervisão, tomei decisões que afetaram presidentes e agi contra tantos políticos proeminentes globalmente que perdi a conta. O Facebook projeta uma imagem de força e competência para o mundo exterior, mas a realidade é que muitas de nossas ações são acidentes imprudentes e descuidados", disse[252].

No pleito eleitoral dos Estados Unidos de 2020, apesar das medidas, a velocidade da atuação na plataforma seguiu sendo um problema grave. Veículos noticiaram que, no processo de apuração, um grupo foi criado com o intuito de difusão de teorias da conspiração sobre uma suposta fraude na votação, que nunca teve evidências e foi rejeitada pelas autoridades eleitorais do país. O grupo arregimentou 365.000 integrantes em dois dias, antes de ser excluído pela plataforma[253].

252. WAKENFIELD, Jane. 'Tenho sangue nas mãos': a ex-funcionária do Facebook que denuncia responsabilidade da rede em campanhas de manipulação. BBC News. Publicada 16 set. 2020. Disponível em: <https://www.bbc.com/portuguese/internacional-54184735.amp>. Acesso em: 8 dezembro 2020.

253. LESKIN, Paige. Facebook takes down group for making false election claims and 'worrying calls for violence' after it amassed over 365,000 members in 2 days. Business Insider. Publicada em: 5 nov. 2020. Disponível em: <https://www.businessinsider.com/facebook-group-stop-the-steal-removed-election-fraud-calls-violence-2020-11?r=ts-sub>. Acesso em 8 dezembro 2020.

3.2 INSTAGRAM

Apesar de já ter ganhado um número considerável de usuários e ter ascendido como uma das maiores redes sociais do mundo, o Instagram recebe muito menos atenção como lócus de desinformação do que outras plataformas, especialmente quando comparado com a plataforma original da sua empresa controladora, o Facebook. Mesmo assim, há estudos que apontam a presença de mensagens enganosas neste ambiente on-line. Mak (2017)[254] indicou que 20 milhões de usuários receberam material de contas falsas vinculadas a fontes russas nas eleições de 2016 nos Estados Unidos.

Outro estudo (SELTZER et al., 2017) sobre a difusão de posts sobre o Zika vírus encontrou que 60% dessas informações eram enganosas ou confundiam os usuários. As mensagens fabricadas eram mais populares do que aquelas com alegações verdadeiras. Com o crescimento do número de usuários, a plataforma assumiu maior importância, também se tornando canal de difusão de conteúdo enganoso.

Investigação da Fundação Oswaldo Cruz[255] identificou que o Instagram estava entre as três principais plataformas de difusão de materiais enganosos sobre a Covid-19. A rede social ficou atrás apenas de Facebook e WhatsApp como canal de difusão de desinformação sobre o tema, como curas milagrosas, descredenciamento da seriedade da pandemia e contestação de métodos de prevenção, como o distanciamento social.

254. MAK, A. How Russia used Instagram to influence the 2016 presidential election. Slate. Publicada em: 1º nov. 2017. Disponível em: https://slate.com/technology/2017/11/how-russia-used-instagram-to-influence-the-2016-presidential-election.html.

255. Fiocruz. Estudo identifica principais fake news relacionadas à Covid-19. Portal Fiocruz. Publicada em: 21 mai. 2020. Disponível em: https://portal.fiocruz.br/noticia/estudo-identifica-principais-fake-news-relacionadas-covid-19.

3.2.1 ABORDAGEM DO FENÔMENO

A abordagem do Instagram é a desenvolvida pelo Facebook discutida na seção anterior. Há alguns pontos específicos, que serão abordados aqui, mas para evitar repetição, remetemos ao debate realizado como o exame crítico da política da companhia para o combate à desinformação na rede social originária de mesmo nome e no Instagram. O WhatsApp, também controlado pelo conglomerado, possui uma estratégia própria, que será analisada adiante.

No Instagram, também vigora o modelo calcado na análise por verificadores de fora da plataforma, medidas de moderação e oferta de mais informações aos usuários. Contudo, as medidas foram implementadas posteriormente ao Facebook, o que indica uma lentidão na companhia em adoção das políticas de forma conjunta, deixando o Instagram mais vulnerável à presença dessa prática no seu interior.

Assim como no Facebook, há uma série de vedações de conteúdos e condutas que podem estar relacionados ou conter desinformação, como contas falsas, comportamento inautêntico e discurso de ódio. Essas, a despeito de constituírem o que chamamos aqui de medidas correlatas, são apresentadas como complementos das iniciativas das duas plataformas.

3.2.2 MODERAÇÃO DE CONTEÚDO

Dentro do seu modelo de verificação e moderação, a possibilidade de denúncia de um conteúdo pelo usuário foi implementada no Instagram apenas no início de 2019. Já a parceria com verificadores independentes teve início em maio do mesmo ano nos Estados Unidos, tendo sido estendido para os demais países em dezembro.

Ou seja, se considerada a série de providências iniciada no fim de 2016 diante dos questionamentos em razão das acusações de

impacto sobre as eleições presidenciais dos Estados Unidos, o início da checagem ocorreu apenas três anos depois no Instagram, mesmo sendo a rede social de crescimento mais promissor nos últimos anos, alcançando 800 milhões de usuários em 2017 e batendo a marca do 1 bilhão em 2018[256].

Em relação às providências de moderação, houve uma adaptação à arquitetura informacional do Instagram interessante. A redução de circulação passou a ser complementada com a não indexação das hashtags inseridas na publicação e com a exclusão da recomendação na página "Explorar".

Tais medidas são importantes, uma vez que, para além da restrição de alcance das mensagens falsas no *feed* dos seguidores do perfil, contribuem para que a circulação seja menor em outros espaços. Mais do que um benevolência, a limitação seria esperada uma vez que seria estranho a empresa manter em seus sistemas de recomendação conteúdos verificados.

O Instagram informa que a prática recorrente pode implicar em penas maiores. Essa progressão é interessante, uma vez que coíbe de forma mais incisiva usuários cujo comportamento não é apenas habitual, mas constante, sinalizando uma tática de presença na plataforma. Mas a empresa não divulga de forma transparente quais são as gradações das sanções nem o número de violações. Seria mais adequado que essas regras fossem expostas aos usuários.

No tocante às regras de devido processo, a plataforma afirma que as verificações podem ser questionadas pelos usuários. Entretanto, não há dados acerca do número de apelações e do percentual de recursos aceitos pelas equipes de revisão da firma. Também não fica claro no material institucional se toda revisão é realizada por ao menos uma pessoa natural ou se podem ser procedidas por sistemas automatizados.

256. Statista. Number of monthly active Instagram users from January 2013 to June 2018. Statista. Publicada em 3 dez. 2019. Disponível em: https://www.statista.com/statistics/253577/number-of-monthly-active-instagram-users/.

Assim como no Facebook, o grande volume de conteúdo vem sendo tratado pelo emprego em proporção cada vez maior deste tipo de tecnologia. As limitações dessas ferramentas - especialmente na análise de conteúdo e balanceamento de contextos diversos – colocam a questão dos riscos à liberdade de expressão nos erros e abusos dessas medidas de moderação. A dificuldade de mensurá-los ainda é um problema relevante no Instagram.

3.2.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

A rotulagem de conteúdos é uma das ações de promoção de informações da abordagem do Instagram. Esse processo se torna ainda mais importante na rede social, dado sua arquitetura calcada em imagens como elemento central das publicações. Os modos de consumo das publicações são fundamentalmente imagéticos, tendo as descrições dos posts posição bem menos destacada.

Desta forma, os rótulos nas imagens assumem a condição de uma ferramenta necessária para chamar a devida atenção dos usuários. Estudo (VRAGA et al., 2020) confirmou como as correções mostradas na plataforma reduzem a credibilidade de posts enganosos e contribuem para uma postura mais cuidadosa diante do conteúdo, como as escolhas pelo engajamento e/ou pelo compartilhamento deste.

Contudo, a despeito desse procedimento de rotulagem, a arquitetura da rede social traz um obstáculo muito relevante à checagem e a uma postura mais crítica diante dos conteúdos recebidos no *feed*. A impossibilidade de inserção de links nas publicações dificulta o acesso ao conteúdo original, exigindo um esforço de incluir o "link na bio" do perfil e gerando um trabalho adicional para quem quer encontrar a origem da informação.

Assim, a regulação tecnológica do Instagram não somente reforça o elemento imagético em detrimento do texto (onde a contextua-

lização e as indicações de referências podem ser operadas de forma mais completa) como cria essa dificuldade visando manter o usuário no seu interior, promovendo uma arquitetura de "jardim murado".

Tal restrição é relevante uma vez que a verificação e a correção não ocorrem somente pelas iniciativas das plataformas, mas pelas ações dos próprios usuários de buscar mais referência sobre uma determinada notícia. Ou pelo acompanhamento por conta própria de outras agências de checagem, uma vez que as parcerias firmadas pelo Facebook contemplam apenas algumas delas.

Ainda no tocante à promoção de informações, é preciso ressaltar a importância da disponibilização de informações da Organização Mundial de Saúde sobre a pandemia. Assim como no Facebook, orientações de autoridades de saúde foram mostradas aos usuários.

Contudo, também a exemplo do Facebook é estranho que a plataforma não tenha uma prática de envio de mensagens educativas sobre como evitar e não compartilhar conteúdos desinformativos aos usuários, podendo inclusive assegurar o alcance a todos os participantes do ambiente e visibilidade desse tipo de material no topo do feed.

Também seguindo a linha do Facebook, não há uma "política sobre desinformação", como há sobre outros temas, como a "política de dados". As medidas estão dispersas em diversos locais do site institucional, bem como em comunicados da empresa reunidos na apresentação das medidas do capítulo anterior. Na Central de Ajuda, não há links específicos para o tema, assim como há, por exemplo, sobre "distúrbios alimentares" ou "dicas para pais".

Assim, é necessário ressaltar uma ausência fundamental para o combate à desinformação: a existência de uma abordagem exposta enquanto tal de forma simplificada aos usuários, para que eles possam saber os limites de atuação na plataforma e as consequências de promover tal prática na rede social. A falta de uma comunicação mais clara enfraquece o efeito inibidor que a divulgação dessa estratégia pode ter.

Por fim, outra vulnerabilidade da atuação do Instagram é a falta de informações transparentes acerca da quantidade e da efetividade das medidas. Elas não constam no relatório de transparência dos padrões da comunidade[257], que aborda resultados das providências de moderação sobre oito categorias de violações. Dessa forma, as medidas ficam nas promessas sem a condição de que a sociedade possa aferir se estão ou não funcionando.

3.2.4 MEDIDAS CORRELATAS

As medidas correlatas do Instagram envolvem remoção de contas violadoras dos padrões da comunidade e de conteúdos que podem estar relacionado à desinformação, como discurso de ódio. A ponderação aqui é semelhante à colocada no item anterior. A sanção de exclusão de publicações é importante no caso de discurso de ódio.

Mas a complexidade da análise de discursos, as diferentes perspectivas sobre fatos e elementos e as diferenças que contextos podem trazer implicam cuidados na avaliação e punição de posts. O risco de abusos já é real pela própria natureza dos discursos, mas é ainda maior frente ao já destacado crescimento do emprego de sistemas automatizados.

A organização Coding Rights teve uma publicação no dia da visibilidade lésbica removido dos stories de sua conta no Instagram em 2019. O post trazia a imagem de uma sapinha, em referência ao termo sapatão, utilizado para designar lésbicas. O uso da palavra sapatão foi classificado pelo Instagram como discurso de ódio[258]. O caso exemplifica as nuances nesse tipo de avaliação.

257. Disponível em: https://transparency.facebook.com/community-standards-enforcement.
258. Post sobre o caso disponível em: https://www.instagram.com/p/B17E4olJ7A0/?igshid=iebynxuonzhy.

3.3 WHATSAPP

A primeira questão que salta aos olhos na análise das medidas adotadas pelo WhatsApp é o não reconhecimento da ampla utilização do aplicativo para a promoção de desinformação. Enquanto diferentes pesquisas acadêmicas e sondagens de opinião comprovam a circulação de conteúdos desse tipo e apontam forte preocupação da sociedade com essa situação, a corporação afirma frequentemente ser uma plataforma de mensagens privadas criptografadas de ponta a ponta e, como tal, menos permeável à desinformação. Essa postura contrasta, inclusive, com as manifestações de preocupações dos diferentes órgãos e relatorias vinculadas aos direitos humanos, que alertam sobre violações de diferentes direitos associadas às estratégias de desinformação e instam as plataformas e outros agentes a atuarem contra elas.

Apesar dessa perspectiva negacionista, a plataforma foi pressionada a adotar medidas concretas contra a desinformação, especialmente após casos na Índia e no Brasil. Lá, o WhatsApp se tornou uma das principais fontes de desinformação política antes das eleições de 2019, com registro de circulação de muitos conteúdos que tratavam, por exemplo, de sequestros de crianças e supostos coletores de órgãos, que teriam contribuído para a morte de cerca de uma dúzia de pessoas, no ano anterior. Executivos do WhatsApp chegaram a declarar que partidos políticos estavam violando as regras de uso da plataforma ao usar sistemas automatizados para fazer disparos em massa de mensagens ou para espalhar fake news. As autoridades eleitorais tentaram impor uma série de restrições, tendo em vista que o problema também havia sido registrado em pleitos anteriores[259].

No Brasil, conforme já mencionado nesta pesquisa, a utilização do WhatsApp para promoção da desinformação foi bastante evi-

259. Disponível em: <https://www1.folha.uol.com.br/mundo/2019/04/india-tem-maior-eleicao-do-mundo-em-meio-a-temor-de-fake-news.shtml>. Acesso: 18 ago. 2020.

dente nas eleições de 2018. Naquele processo, revelou-se a configuração de estratégias que geraram, inclusive, a captura e a venda de dados pessoais[260], especialmente de contatos telefônicos, para viabilização de disparos em massa. A situação só posteriormente foi reconhecida pela própria companhia[261].

As mudanças aqui consideradas foram aplicadas a partir de março de 2018, data definida para o recorte inicial da pesquisa. De pronto, percebemos que a negação dos fatos acaba gerando falta de transparência e se reflete na existência de poucas medidas anunciadas como relacionadas à desinformação. Pelo que vimos na exposição das ações do WhatsApp, a maior parte do que pode incidir na questão foi rotulada como "medida correlata". No caso das ações propriamente relacionados à desinformações, o maior número foi de ações classificadas como "Promoção de informações e transparência", conforme será detalhado a seguir.

260. Reportagens sobre as eleições no Brasil mostraram, entre outros itens, que o contato telefônico foi vendido por R$ 0,24. O valor do disparo de uma mensagem depende da quantidade. O mais caro, com 1 milhão de disparos, cobrava R$ 0,062 por envio. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2020/07/02/proibidas-pela-justica-maquinas-de-spam-no-whatsapp-continuam-operando.htm?cmpid=copiaecola>. Acesso: 17 ago. 2020.

261. Em 2019, um ano após a eleição que levou Jair Bolsonaro à presidência da República, o WhatsApp admitiu que a eleição brasileira teve uso de envios maciços de mensagens, com sistemas automatizados contratados de empresas. Em 2020, o PT teve nove de suas contas no WhatsApp desativadas por envio automatizado de mensagens, em um processo que envolveu denúncias de spam político. Ainda que não proibido pela legislação nacional, o spam é uma das vedações que constam nas regras do WhatsApp, que desativa quem "enviar mensagens em massa ou automatizadas que violam nossos termos de serviço". Disponível em: <https://www1.folha.uol.com.br/poder/2019/10/whatsapp-admite-envio-massivo-ilegal-de-mensagens-nas-eleicoes-de-2018.shtml?origin=folha#> e <https://www.uol.com.br/tilt/noticias/redacao/2020/07/07/whatsapp-bloqueia-contas-do-pt-flagrado-por-disparo-em-massa-de-mensagens.htm?cmpid=copiaecola>. Acesso: 16 ago. 2020.

3.3.1 ABORDAGEM DO FENÔMENO

As duas características que o WhatsApp alega ter merecem ser discutidas para se compreender a abordagem do fenômeno: o fato de ser criptografado e privado. Primeiro, de fato, entendemos que a adoção da criptografia favorece o combate à desinformação, já que esta é organizada e direcionada a partir da utilização de dados pessoais, inclusive de informações produzidas a partir do que os próprios usuários escrevem em redes que não dispõem de tal mecanismo, caso do Facebook. Nesse sentido, é positivo que o WhatsApp siga defendendo tal configuração, como ocorreu em dezembro de 2019, quando o Facebook, dono do WhatsApp, manifestou-se em documento intitulado *Facebook's public response to open letter on private messaging*[262], dirigido a autoridades do Reino Unido, dos Estados Unidos e da Austrália, no qual defendeu a manutenção da encriptação de ponta a ponta como forma de proteção.

No texto, em referência à possibilidade de estabelecer *backdoor*, uma porta de acesso ao sistema, criada a partir de um programa instalado que não foi autorizado pelo proprietário do sistema e que permite o acesso ao computador por pessoas não autorizadas, escreveu: "Os especialistas em ciber-segurança provaram repetidamente que, quando se enfraquece qualquer parte de um sistema encriptado, enfraquecem-no para todos, em todo o lado. O acesso 'de porta traseira' que exige para a aplicação da lei seria uma prenda para criminosos, hackers e repressivos regimes, criando uma forma de entrar nos nossos sistemas e deixando todas as pessoas nas nossas plataformas mais vulneráveis aos danos da vida real. É simplesmente impossível criar uma tal porta traseira para um propósito e não esperar que outros a tentem abrir".

A corporação detalhou sua visão para o futuro da comunicação, apontando que espera que seja mais permeada por "serviços

262. https://about.fb.com/wp-content/uploads/2019/12/Facebook-Response-to-Barr-Patel-Dutton-Wolf-.pdf

privados, codificados e seguros". Como exemplo do que pretende que seja feito para contribuir com isso, citou o caso do WhatsApp: "centrarmo-nos no caso de utilização mais fundamental e privada - o envio de mensagens – para torná-lo o mais seguro possível, e depois construir mais formas de as pessoas interagirem além disso, incluindo chamadas, conversas de vídeo, grupos, histórias, empresas, pagamentos, comércio e, em última análise, uma plataforma para muitos outros tipos de serviços privados". No artigo, listou princípios que orientam a atuação, conforme seu fundador: interações privadas; criptografia; redução da permanência ("não vamos guardar mensagens ou histórias por mais tempo do que o necessário para prestar o serviço ou por mais tempo do que as pessoas querem", escreveu); segurança; interoperabilidade; e armazenamento seguro de dados.

Ocorre que, se a criptografia protege os dados e as conversas entre dois usuários, outros mecanismos acabam indo de encontro à privacidade das comunicações. Entramos, aqui, no segundo ponto relacionado à forma como o WhatsApp se apresenta. Ainda que desenvolva, nas palavras de Pablo Bello, diretor de políticas públicas da companhia para América Latina, "um esforço sistemático para se manter como um espaço de conversação privada"[263], algumas de suas funcionalidades destoam disso. Uma delas consiste na opção de criação de lista de transmissão, em que um usuário pode enviar conteúdo para até 256 contatos. Até 2016, eram até 100 os destinatários de uma lista. Os atuais 256 podem ser pouco para o que, em geral, se considera "massivo", mas vale notar que não existe um limite para o número de listas de transmissão que se queira criar. Sistemas automatizados podem facilitar a criação e a administração de incontáveis listas, portanto.

Sobre isso, a plataforma argumenta que bloqueia contas denunciadas por uso excessivo. "Suas mensagens enviadas por uma

263. https://www.uol.com.br/tilt/noticias/redacao/2020/06/23/rastrear-mensagem-e-por-tornozeleira-eletronica-em-usuario-diz-whatsapp.htm?cmpid=copiaecola

lista de transmissão só serão recebidas por usuários que tenham adicionado seu número de telefone à lista de contato deles. Observe que a utilização frequente de listas de transmissão pode levar outros usuários a denunciar suas mensagens para nossa equipe", diz o texto disponível no FAQ do blog do WhatsApp. Avaliamos como positivo o fato de haver restrições a essa prática, ainda que seja necessário ponderar que faltam comunicações aos usuários das mudanças desses termos. A efetividade dos bloqueios será discutida mais adiante. Por enquanto, fiquemos com a caracterização da plataforma e de suas ferramentas.

Além das listas, há a possibilidade de criação de grupos – inclusive sem que o usuário necessite aceitar previamente sua inclusão em alguns deles, conforme detalhado na exposição das mudanças no aplicativo anteriormente. Atualmente, os grupos também podem ter, cada um, 256 integrantes. É viável configurá-los, ainda, para que apenas administradores enviem mensagens, o que o transforma em canal de distribuição unilateral de conteúdos. Não há restrições para o número de grupos em que cada usuário pode ser membro[264]. Embora especialistas defendam a redução no número de grupos que cada usuário pode criar, o mensageiro não sinaliza pretender restringi-lo[265]. O máximo hoje é 9.999. Além da redução, é recomendável que a configuração do WhatsApp seja alterada para que, por *default*, a inclusão nos grupos só seja feita após permissão do usuário.

Mais recentemente, foi criado o WhatsApp Business, que facilita a operacionalização desses e de outros mecanismos, que modificam a relação entre o criador de conteúdo e seu receptor, tornando-a mais desigual, pois o criador tem mais poder sobre as

264. https://canaltech.com.br/redes-sociais/whatsapp-diz-ser-impossivel-reduzir-o-numero-de-encaminhamentos-antes-da-eleicao-125172/

265. A Safernet apresentou ao TSE, ainda antes das eleições de 2018, proposta de redução de 9.999 para 499, conforme registrado pelo Intervozes. Disponível: http://www.intervozes.org.br/direitoacomunicacao/?p=30215

trocas instituídas. Todas essas funcionalidades e a transformação do WhatsApp em uma praça de mercado – por meio da qual, inclusive, começa a ser possível efetivar pagamentos – faz com que se gere como subproduto um mercado de dados pessoais, conforme mencionado no início do texto.

Cabe notar ainda que os usos que têm sido feitos por grupos que constroem verdadeiras cadeias de distribuição de conteúdos por meio do aplicativo acabam fazendo com que as características atuais do WhatsApp possuam pontos de contato com a lógica da comunicação de massa, definida por Thompson como "a produção institucionalizada e a difusão generalizada de bens simbólicos através da transmissão e do armazenamento da informação/comunicação" (THOMPSON, 2018, p. 288). Usos que são, em parte, facilitados pelas próprias configurações da plataforma, como as que citamos anteriormente.

É possível sustentar, pelo exposto, que, mais que mensageiro, o WhatsApp deve ser visto como um tipo de rede social. Como tal, é mais suscetível à ocorrência de desinformação em seu interior do que a corporação reconhece. Santos et al (2019, p. 317) assim descrevem esse processo de produção de redes no WhatsApp: "mais do que uma rede de pessoas conectadas através de grupos, o WhatsApp está sujeito às dinâmicas de uma rede bipartite de grupos interconectados por participantes em comum que regulam o intercâmbio de informações, permitem o aumento exponencial de visibilidade e lógicas de difusão viral de notícias falsas mesmo dentro de uma rede fechada".

Não há, devemos ressaltar, uma prática de distribuição a partir do próprio WhatsApp, que não se coloca como produtor de conteúdo, mas permite uma lógica de operacionalização que gera um alcance ampliado de conteúdos, ainda que de forma menos centralizada do que o que ocorre na radiodifusão ou mesmo em plataformas de redes sociais como o Facebook, o Twitter e o YouTube, que não são criptografadas e que se pautam pela lógica de

exposição de qualquer produção de um usuário para um conjunto amplo de pessoas (sejam elas já associadas – "amigas", seguidoras" – ou não a ele).

Esses elementos são desconsiderados pelo WhatsApp na discussão sobre sua utilização para a promoção da desinformação. Por se compreender como uma plataforma de trocas privadas, abstém-se de processos que poderiam ser úteis para enfrentar o problema e dar transparência às medidas tomadas em relação a ele. Não há, por exemplo, uma definição de desinformação que oriente suas ações, ainda que a palavra desinformação apareça em alguns de seus documentos, conforme apontado ao detalhar suas medidas. De acordo com o que informa a corporação, não existem equipes e processos definidos em caso de reconhecimento de conteúdo desinformativo, o que configura ausência de uma política consolidada sobre desinformação dentro da plataforma.

Tal abordagem destoa, portanto, do que recomendam, em geral, as autoridades que atuam na defesa de direitos humanos, que pedem que as plataformas tomem iniciativas e as desenvolvam com transparência e em diálogo com instituições e outras instituições locais.

3.3.2 MODERAÇÃO DE CONTEÚDO

Uma vertente de ação que apontamos como diretamente vinculada à desinformação foi apontada nesta pesquisa como de moderação de conteúdo. Ainda que o WhatsApp negue tal prática, no contexto da pandemia do coronavírus ele divulgou que as "mensagens altamente encaminhadas" passariam a ter o encaminhamento reduzido a um contato de cada vez. Esse tipo de restrição à circulação de mensagens já vinha sendo adotada em momentos, mas não era apresentada como uma reação à desinformação, mesmo em lugares como a Índia, onde a redução para um só contato foi

feita pela primeira vez de forma mais agressiva (limitando a possibilidade de encaminhamento que antes era para 250 chats para 5, ao passo que, em outros países, essa medida começou com o teto fixado de até 20 chats).

Ainda que o encaminhamento continue sendo possível – afinal só há maior restrição no caso de mensagens altamente encaminhadas e número máximo de destinatários é válido para cada encaminhamento, mas não há um limite do número de vezes que uma mensagem pode ser compartilhada –, números oficiais divulgados em 2020 apontam que a limitação no encaminhamento de mensagens de 20 para 5 reduziu globalmente o HFM em 25%. A restrição de 5 para 1, por sua vez, reduziu o HFM em 70%.

A redução merece ser analisada com cautela, pois trata-se de uma redução da comunicação de todos os usuários. Sabemos que há muitos usos do encaminhamento para projetar conteúdos e vozes – inclusive campanhas de solidariedade no contexto da pandemia e tantas outras iniciativas do tipo. Limitar esse mecanismo pode significar cerceamento da comunicação de grupos que utilizam o aplicativo para esse tipo de finalidade, ao passo que as campanhas de desinformação, que são organizadas de forma profissional, conseguem driblar esses limites por meio de tecnologias de distribuição massiva. Reconhecendo a diferença entre usuários comuns e agentes promotores de desinformação, pessoas atingidas por desinformação e que não possuem os mesmos mecanismos podem ter sua capacidade de resposta limitada.

O aspecto positivo desse tipo de mecanismo é que ele gera a desaceleração da velocidade na difusão da desinformação. A essa mesma conclusão chegou o pesquisador Fabricio Benevenuto. Ao analisar especificamente o resultado daquela mudança, percebeu a redução da velocidade, mas não a interrupção do processo de desinformação. Uma das possíveis explicações para isso é exatamente a ligação entre os grupos. "(Mas) estes grupos políticos públicos da WhatsApp podem realmente criar uma espinha dorsal para que

a desinformação se espalhe rapidamente", explica[266].

No fundo, a questão é que a duplicidade do WhatsApp – uma parte com natureza fechada e outra de fácil transferência de conteúdo multimídia e compartilhamento de informações para grupos de grande escala – torna-o difícil para a implementação de medidas para combater a desinformação. Ainda que longa, as conclusões da pesquisa de Philipe Melo et al são elucidativas:

> Nossos resultados mostram que um conteúdo pode se espalhar muito rapidamente através da estrutura de rede de grupos públicos no WhatsApp, alcançando posteriormente os grupos privados e os usuários individuais. Nossas observações empíricas sobre a rede de grupos públicos do WhatsApp em três países diferentes fornecem um meio de inferir a velocidade da informação em termos de minutos relacionados a cenários do mundo real. Verificamos que a maioria das imagens (80%) não dura mais que 2 dias no WhatsApp, o que, na Índia, já pode ser suficiente para infectar metade dos usuários em grupos públicos, embora ainda existam 20% das mensagens com tempo suficiente para ser viral nos três países usando qualquer uma de nossas estratégias para estimar o tempo de infecção. […] Enquanto o limite de número de usuários por grupo pode proibir a criação de hubs gigantes para divulgar informações através da rede, este limite, no entanto, não é capaz de impedir que um conteúdo chegue a uma grande parte de toda a plataforma. Mais importante, nossa análise mostra que os baixos limites impostos ao encaminhamento e transmissão de mensagens (por exemplo, até cinco encaminhamentos) oferecem um atraso na propagação da mensagem de até duas ordens de magnitude

266. Disponível em: <https://www.poynter.org/fact-checking/2019/whatsapp-efforts-to-curb-misinformation-arent-entirely-effective-research-shows/ >. Acesso: 17 ago. 2020.

> em comparação com o limite original de 256 utilizado na primeira versão do WhatsApp. Observamos, entretanto, que, dependendo da viralidade do conteúdo, esses limites não são eficazes para impedir que uma mensagem chegue rapidamente a toda a rede. Campanhas de desinformação conduzidas por equipes profissionais com interesse em afetar um cenário político podem tentar criar um conteúdo falso muito alarmante, que tem um alto potencial para se tornar viral. Assim, como contra-medida, o WhatsApp poderia implementar uma abordagem de quarentena para limitar os usuários infectados a espalhar informações errôneas. Isto poderia ser feito restringindo temporariamente as características de viralidade de usuários suspeitos e conteúdo, especialmente durante as eleições, impedindo campanhas coordenadas para inundar o sistema com desinformação. (PHILIPE MELO et al, 2019, p. 11-12)[267]

267. Tradução própria do original em inglês: "Our results show that a content can spread quite fast through the network structure of public groups in WhatsApp, reaching later the private groups and individual users. Our empirical observations about the network of WhatsApp public groups in three different countries provides a means of inferring the information velocity in terms of minutes related to real-world scenarios. We verified that most of the images (80%) last no more than 2 days in WhatsApp which, in India, can be already enough to infect half of users in public groups, although there are still 20% of messages with a time span sufficient to be viral in the three countries using any of our strategies to estimate time of infection. [...] While the limit on the number of users per groups can prohibit the creation of giant hubs to spread information through the network, this limit, however, is not able to prevent a content to reach a large portion of entire platform. More important, our analysis show that low limits imposed on message forwarding and broadcasting (e.g. up to five forwards) offer a delay in the message propagation of up to two orders of magnitude in comparison with the original limit of 256 used in the first version of WhatsApp. We note, however, that depending on the virality of the content, those limits are not effective in preventing a message to reach the entire network quickly. Misinformation campaigns headed by professional teams with an interest in affecting a political scenario might attempt to create very alarming fake content, that has a high potential to get viral. Thus, as a counter-measurement, WhatsApp could implement a quarantine approach to limit infected users to spread misinformation. This

Vemos, portanto, que esse tipo de limite à comunicação dos usuários deve ser considerado com cuidado, porque pode acabar favorecendo a circulação de conteúdos daqueles que se valem de mecanismos artificiais. É fundamental que a avaliação desses efeitos seja feita com transparência e que se busque medidas que atinjam não o usuário comum, mas, sobretudo, os grupos que operam a desinformação em escala industrial.

3.3.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

É útil lembrar as medidas associadas à promoção de informações e transparência: presença, no FAQ do blog, de texto com "Dicas para ajudar a evitar a disseminação de boatos e notícias falsas"; premiação de pesquisadores interessados em explorar questões relacionadas à desinformação; campanhas de comunicação; parcerias com projetos de checagem de fatos; promoção de informações, especificamente no contexto da pandemia do coronavírus; além da assinatura de compromissos com instituições brasileiras para apoiar a prevenção de práticas dolosas de desinformação, desenvolver projetos de fomento à educação digital e iniciativas que promovam o "jornalismo de qualidade".

Por mais que a lista pareça longa, os resultados concretos dessas ações são questionáveis. Em primeiro lugar, as comunicações sobre desinformação, como o texto com dicas, são pouco acessíveis e, em geral, não são direcionadas efetivamente ao usuário do aplicativo sobre o tema inclusive quando há alterações nos termos de uso. A pessoa que usa o aplicativo não recebe notificação sobre esse tipo de conteúdo. É provável que ela não encontre facilmente informações sobre a existência ou formas de se combater a desinformação.

could be done by temporarily restricting the virality features of suspect users and content, especially during elections, preventing coordinated campaigns to flood the system with misinformation".

Os poucos textos que mencionam a questão não são aparentes no aplicativo, sendo necessário passar por várias etapas para, em uma busca ativa, chegar até eles. Quando comparamos as informações disponíveis em inglês e português, vemos também que há diferenças, sendo mais detalhadas aquelas no idioma do grupo que originalmente produziu o aplicativo. O cenário de informações presentes, sobretudo, como postagens dispersas em espaços do próprio WhastApp ou do Facebook dificulta inclusive a realização de pesquisas, como discutimos ao abordar a metodologia aqui empregada.

Ainda que a corporação tenha se comprometido com a Justiça brasileira no sentido de promover ações, elas são insuficientes e seus resultados deixam a desejar. É o caso da premiação de pesquisas, que não necessariamente tem seus resultados acolhidos pelas plataformas, e também da campanha de comunicações "Compartilhe fatos, não rumores". O WhatsApp aponta que esta alcançou 8 milhões de brasileiros, número bem menor que o de 130 milhões de contas ativas no país. Apenas na Índia ações presenciais de sensibilização junto à população foram realizadas. Mesmo as parcerias com agências de verificação de fatos têm sido pouco expressivas. Ao contrário, quando outras plataformas já adotavam esse tipo de iniciativa, o WhatsApp mantinha-se fechado, o que chegou a gerar controvérsia no Brasil com as agências de checagem[268] e mesmo pedido de melhorias por parte de integrantes da justiça eleitoral[269]. A política de parceria só ganhou mais volume no contexto da pandemia causada pelo coronavírus. Igualmente, a formações de pessoas sobre o tema é irrisória, tendo alcançado pouco mais de 40 mil, de acordo com dados oficiais da empresa. Ademais, muitas dessas iniciativas são tomadas apenas em um

268. Disponível em: <https://www.aosfatos.org/noticias/whatsapp-diz-combater-desinformacao-mas-causa-ruido-em-reuniao-no-tse-com-checadores/>. Acesso: 20 mai. 2020.
269. Disponível em: <https://www.nsctotal.com.br/noticias/checagem-no-whatsapp-pode-melhorar-diz-vice-procurador-geral-eleitoral>. Acesso: 20 mai. 2020.

pequeno número de países, deixando milhões de usuários em outros lugares mais expostos à desinformação. Além disso, devido à escassez de dados disponíveis ao público, é difícil para terceiros independentes (verificadores de fatos, mídia, acadêmicos e outros) avaliar a eficiência dessas respostas.

Devido à pandemia, as plataformas, incluindo o WhatsApp, desenvolveram mais ações de promoção de informações como reação à proliferação de desinformação, por meio de parcerias com autoridades de saúde, tanto em âmbito internacional quanto nacional, a exemplo da OMS e do Ministério da Saúde no Brasil. Essas medidas são importantes, mas são ainda pouco acessíveis para um usuário médio. Afinal, para obter informações, o usuário precisa tomar conhecimento da existência de um número, entrar em contato com ele, para então enviar questionamentos e receber retornos. As informações poderiam ser ofertadas para o próprio usuário, mas não é o que acontece. Até mesmo as "dicas" para evitar a desinformação estão no site e não diretamente no aplicativo, o que contrasta com o fato do acesso à rede se dar, atualmente, mais pelo smartphone e por aplicativos do que pelo computador. Apenas em agosto de 2020, passou a disponibilizar *chatbot* da IFCN que apresenta mensagens dentro do próprio app. Vale notar também que, ainda que seja uma corporação transnacional, o tempo para a disponibilização, quando isso ocorre, de serviços nas línguas dos diferentes países nos quais opera é mais largo do que a disponibilização desses serviços em inglês.

Em relação a essas formas de lidar com a questão, como não há definição do conceito de desinformação nem do processo de verificação e possíveis medidas diante de conteúdos desinformativos, essas ações podem gerar outros problemas, como imprecisões e mesmo restrições à liberdade de expressão, até porque não há devido processo estabelecido, o que também contraria as recomendações de direitos humanos. O usuário que tenha um conteúdo apontado como desinformativo não tem caminhos

nítidos para questionar essa caracterização nem as possíveis consequências dela.

Por fim, seria possível ao WhatsApp enviar notificação para os usuários sobre o *hub* de informações, *bots* e outras estratégias, o que poderia ajudar a garantir maior visibilidade às iniciativas do tipo e caminhar para que os documentos de direitos humanos chamam de resiliência dos cidadãos ou para aquilo que o Intervozes defende como empoderamento digital e educação para a mídia. Sem esclarecer e alertar os usuários, esses processos ficam mais distantes da realidade atual.

3.3.4 MEDIDAS CORRELATAS

Ao adotar a limitação da possibilidade de encaminhamento, houve também a retirada de um botão de encaminhamento rápido que aparece ao lado de mensagens com mídias, o que mostra que o próprio design da plataforma pode incidir ou evitar a circulação de conteúdos no geral. Esse tipo de medida deve ser mais amplamente discutido, contribuindo também para avançarmos na compreensão de que as tecnologias não são neutras, mas, ao contrário, causam impactos desde a forma como são desenhadas.

Tal problemática também aparece quando observamos medidas consideradas como correlatadas, a exemplo da mudança nas configurações dos grupos que determinou que criadores do grupo não podem mais ser removidos dele. Essa medida pode evitar o comportamento de agentes que criam seguidos grupos e se retiram deles após outro administrador assumir a tarefa de repasse da desinformação. Não obstante, ainda é possível adicionar diretamente as pessoas. Ao longo da pesquisa, verificamos que houve testes no sentido de proibir a adição de uma pessoa a um grupo sem o conhecimento dela. No entanto, isso não foi implementado. Questionado, o WhatsApp limitou-se a afirmar que "avalia continuamente

soluções tecnológicas para fortalecer a privacidade de seus usuários e para manter a qualidade das conversas para eles, incluindo ferramentas para que os usuários relatem contas. Algumas destas soluções são implementadas, enquanto algumas delas ainda estão em processo de avaliação, em testes, ou foram descartadas".

O WhatsApp argumenta que as mudanças nas configurações dos grupos têm ampliado "a definição dos usuários e potencialmente a criação artificial de grupos", sustentando que "os usuários podem escolher entre três opções sobre quem pode adicioná-los a um grupo: todos - seus contatos - seus contatos, exceto... (podem ser todos). Nos casos em que alguém não pode adicionar outro usuário a um grupo, ele pode enviar um convite particular através de um bate-papo individual, oferecendo a opção de se juntar ao grupo. O usuário 'convidado' tem três dias para aceitar o convite antes que ele expire".

Mas não é difícil sustentar se tratar de algo efetivamente acessível, pois depende da atuação dos usuários e se pauta pelo critério de exclusão. São os usuários que devem gerenciar esse procedimento nas configurações. É preciso fazer o seguinte caminho: Configurações do WhatsApp> Conta> Privacidade> Grupos. Ao abrir a seção, escolher "Todos", permitindo que todos o coloquem em grupos sem autorização para isso; "Meus contatos", configurando que a adição se dê apenas por quem está na lista de contatos do usuário; e "Meus contatos, exceto", que autoriza que seus conhecidos o coloquem em grupos, mas com exceção de alguns nomes. O site WABetaInfo, especializado no acompanhamento do WhatsApp, informou que havia uma alternativa chamada "Ninguém", mas que o WhatsApp preferiu, ao disponibilizar as versões beta 2.19.110.20 para iOS e 2.19.298 para Android, trocar pela lista de exceções, que só impede certos usuários identificados de fazer a adição em grupos.

Dessa forma, avaliamos que não há efetivamente independência e autonomia, já que continua sendo permitido que as pessoas

sejam incluídas em grupos ou em listas de transmissão sem consentimento e que apenas administradores possam enviar mensagens, configurando uma relação desigual entre eles e os receptores, conforme já pontuado. Mudanças nessa lógica poderiam contribuir para evitar a formação artificial de grupos e redes de grupos, empoderando, ademais, os usuários, para que eles façam escolhas sobre os espaços que querem integrar.

Em vez de criar mecanismos para limitar esse tipo de interação que pode viabilizar uma comunicação mais massiva na plataforma (mudanças nas configurações dos grupos e listas de transmissão, para citar duas possibilidades), passou a identificar as mensagens repassadas e as altamente encaminhadas com uma e duas setas, respectivamente. Esse mecanismo é interessante para alertar as pessoas quanto à origem da mensagem, sem precisar de rastreamento, medida que poderia fomentar uma política vigilantista no âmbito do WhatsApp. Não obstante, ainda não é possível saber a autoria, pesquisa reversa de imagens, a fim de verificar se uma imagem é alterada, informações como data da criação ou mesmo o número de repasses da mensagem, medidas que poderiam tornar essa informação mais completa para o usuário e que são defendidas pelo Intervozes.

Também como parte do combate à desinformação no contexto da pandemia do coronavírus, o WhatsApp introduziu a possibilidade de pesquisar suas mensagens na web. Uma lupa localizada ao lado de mensagem encaminhada foi inserida para facilitar a possibilidade de pesquisar mais informações sobre o conteúdo replicado. A medida é, por um lado, interessante para que as pessoas busquem informações, por outro lado se baseia no acesso a partir do próprio WhatsApp. De acordo com a companhia, a lupa é apenas um mecanismo direto para facilitar a pesquisa, o WhatsApp não acessa o conteúdo da consulta, faz o *hide* no dispositivo do usuário e abre uma janela do Google. Lamentavelmente, não tem avaliação zero.

Além das medidas correlatas de caráter preventivo e informativo, o WhatsApp também tem banido usuários ou mesmo evitado a criação de contas, tendo como base, em geral, os seus Termos de Serviço, que proíbem, como vimos, envio massivo e automatizado de mensagens. A companhia informa que, por ser criptografada, não pode acessar o conteúdo das mensagens de seus usuários nem pode moderar ou geobloquear o conteúdo. De acordo com o WhatsApp, as únicas informações não criptografadas que pode acessar são imagem do perfil do usuário, nome do usuário, "sobre" o usuário, nome do grupo, foto do perfil do grupo e descrição do grupo. "Assim, para combater usos ilegítimos e/ou ilegais de sua plataforma, ela concentra seus esforços em detectar padrões de uso que não podem corresponder ou combinar com os do comportamento humano", nos informou.

A disseminação artificial, efetivamente, é algo que precisa ser enfrentada na busca por saídas para as fake news. Uma rede de disseminação artificial que coordena e articula comportamentos por intermédio de contas automatizadas não identificadas ou por tecnologia não fornecida pelo provedor de aplicação de Internet pode impactar de forma artificial a distribuição de conteúdos e mesmo o debate público. Mas nem todo comportamento inautêntico é problemático, sendo útil em serviços públicos ou mesmo na entrega de informações sobre Covid-19, como fizeram, aliás, WhatsApp e IFCN.

Para que essa verificação seja feita considerando essas questões e também a possibilidade de questionamentos em relação à categorização da própria plataforma, é preciso que sejam formulados, adotados e divulgados Termos de Serviços mais claros e detalhados. Mais uma fez, sublinhamos que transparência nas práticas e garantia do devido processo também se mostram fundamentais. Seria possível, por exemplo, informar à plataforma sobre a utilização de robôs, o que permitiria o acompanhamento do cadastro deles e a sinalização para o público de que aquela conta é artificial.

Com devido processo – além de procedimentos claros e informados sobre o tratamento de possível desinformação, o que não há, os usuários poderiam pedir revisão humana das decisões, reparação etc. Como isso não ocorre, o WhatsApp acaba exercendo o poder de determinar o que pode ou não permanecer na sua plataforma, a partir de critérios formulados apenas pela corporação, sem debate ou viabilidade de escrutínio público.

Por fim, há iniciativas que, embora testadas, não foram implementadas, como a já mencionada alteração na adição a grupos. Além dela, em julho de 2018, segundo informou o site WABetaInfo[270], especializado na cobertura do WhatsApp, com destaque para o desenvolvimento de ferramentas e para a atualização tecnológica do mensageiro, estava sendo estudada a detecção de links suspeitos. Passaria a ser emitido um alerta em vermelho aos usuários sempre que um link suspeito é enviado, o que seria utilizado para evitar que as pessoas acessassem páginas com conteúdos maliciosos ou fraudes on-line. Até hoje o mecanismo não foi disponibilizado.

O site também divulgou que a plataforma analisa incluir de fato a possibilidade de visualização do número total de encaminhamentos e também de o administrador de um grupo impedir o envio frequente de mensagem para o grupo[271]. Outra medida cuja informação sobre desenvolvimento foi divulgado pela imprensa[272] e que ainda não foi implementada consiste em recurso para validar informações compartilhadas. Quando a medida de registrar com setas duplas as mensagens encaminhadas muitas vezes estava sendo testada, foi anunciado que os usuários poderiam bloquear mensagens que fossem encaminhadas mais de quatro vezes e que ficariam "marcadas" pelo app, o que não ocorre.

270. https://wabetainfo.com/whatsapp-beta-for-android-2-18-204-whats-new/
271. https://wabetainfo.com/whatsapp-beta-for-android-2-19-97-whats-new/
272. https://canaltech.com.br/smartphone/whatsapp-testa-recurso-para-validar-informacoes-compartilhadas-162197/

Por todo o exposto até aqui, vemos que não há medidas voltadas ao combate a quem produz efetivamente a desinformação. Anúncios de possíveis processos de uso da marca ou incidências em relação a empresas que anunciam a venda de serviços de disparos massivos são insuficientes. Essa insuficiência decorre, em primeiro lugar, da própria ausência de reconhecimento da transformação das características do WhatsApp, que deve ser visto como um tipo de rede social e não como uma mensageiro de conversas privadas apenas. Derivados disso, outros problemas como ausência de comunicação ampla com a sociedade e com o usuários sobre o tema emergem. Por fim, na efetivação das medidas, faltam transparência e devido processo, o que acaba fazendo com que elas sejam unilateralmente definidas e implementadas pelo WhatsApp, ignorando o fato de ser um espaço com cada vez maior relevância pública.

3.4 YOUTUBE

No final de 2019, a pesquisa *Video Viewers*, encomendada pelo Google e conduzida pela Provokers, revelou que o consumo de vídeo na Internet no Brasil cresceu 165% nos últimos cinco anos[273]. Para efeitos comparativos, o consumo de programação de TV aumentou apenas 24% no mesmo período. Cerca de 95% dos entrevistados declararam assistir vídeos on-line tendo no YouTube seu principal canal. Os interesses são os mais variados possíveis, de melhorias na carreira profissional a entretenimento e compra de produtos. O *Digital News Report 2020*, do Instituto Reuters, já mencionado nesta pesquisa, confirma o crescimento da plataforma virtual diante de seu principal concorrente, a televisão[274]. Pela primeira vez

273. https://olhardigital.com.br/noticia/consumo-de-video-online-cresce-165-no-brasil-nos-ultimos-5-anos/90726

274. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

desde o início do levantamento, as redes sociais superaram a TV em termos de consumo de notícias. De 2013 para 2020, saltaram de 47% para 67% dos entrevistados; já a televisão caiu de 75% para 66%. Os impressos, antes com 50% da preferência, agora são meio para consumo de notícias para apenas 23% dos entrevistados.

Entretanto, diversos estudos também apontam o YouTube como um canal para onde as redes de desinformação migraram, sobretudo após as eleições de 2018, quando o Facebook removeu de sua plataforma centenas de páginas que disseminavam conteúdos falsos. Pesquisadores do Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT.CC) avaliam que o fato de uma grande quantidade de links para vídeos do YouTube também ter passado a circular em grupos de WhatsApp no Brasil contribuiu para o crescimento do alcance da plataforma no país.

3.4.1 ABORDAGEM DO FENÔMENO

Como apresentado no capítulo 2, o YouTube afirmou, em entrevista a esta pesquisa, que não trabalha com uma única definição de desinformação, em função da complexidade de se caracterizar conteúdos em vídeo como comprovadamente verdadeiros ou falsos. Assim, dentro das Diretrizes da Comunidade, onde a empresa explica o que é ou não permitido, o usuário não encontrará uma política específica sobre o tema. Apesar de o YouTube ser, dentre as plataformas analisadas, a que mais detalha suas regras de funcionamento – inclusive fornecendo aos usuários diversos exemplos de como uma diretriz é aplicada pelas equipes de moderação de conteúdo, algo recomendado em todos os documentos internacionais de direitos humanos, e tendo criado mais recentemente uma página sobre o combate à desinformação dentro da plataforma –, o fato de não haver uma política própria para o tema dificulta, sobretudo para os usuários padrão (que apenas assistem vídeos),

a compreensão sobre o que pode ou não circular na rede social.

A política que proíbe mídias manipuladas – editadas com a finalidade de adulterarem seu significado e alterarem sua compreensão e que possam apresentar riscos de danos significativos – é uma das que incide diretamente sobre a questão da desinformação, apesar de não ser apresentada enquanto tal nas Diretrizes da Comunidade. Já as moderações feitas desde o início do ano relacionadas à Covid-19, por exemplo, são realizadas com base na política de risco de dano do YouTube.

3.4.2 MODERAÇÃO DE CONTEÚDO

Elemento central no modelo de negócios e de funcionamento do YouTube, as recomendações de vídeos, que aparecem tanto na página inicial da plataforma quanto nas sugestões automáticas de "próximo vídeo", exibidas quando um usuário termina de assistir a um conteúdo, estão no centro das discussões sobre o papel do YouTube na disseminação de desinformação e também de conteúdos extremistas.

Desde janeiro de 2019 nos Estados Unidos e a partir do final do ano no Brasil, o YouTube informou que passou a reduzir o alcance do que chama de "desinformação danosa" e de "conteúdo limítrofe", deixando de recomendá-los aos usuários. Nos EUA, esses tipos de conteúdo representariam, juntos, 1% do que é assistido, e tiveram seu alcance reduzido significativamente desde que a política foi adotada, comprovando o tamanho do impacto que as recomendações tem na plataforma. Diariamente, o YouTube faz mais de 200 milhões de recomendações a seus usuários naquele país. No Brasil, a redução da circulação teria sido de 40%, mas não há números públicos sobre o quanto a "desinformação danosa" ou os "conteúdos limítrofes" representariam no total de vídeos publicados no país.

O problema de transparência sobre a nova política vai além, já que também não há, disponível na plataforma, nenhuma definição clara do que o YouTube trata dentro desses conceitos, tampouco qual a diferença entre ambos. A empresa diz que conta com avaliadores e especialistas externos para implementar a política e que cada conteúdo recebe até nove avaliações. Mas, considerando a redução significativa dessas visualizações, mesmo que tais conteúdos não sejam totalmente removidos da plataforma, há riscos de cerceamento à liberdade de expressão diante da ausência de critérios públicos para a orientação da moderação.

O outro lado da moeda da política de redução desses tipos de conteúdo é sua duvidosa eficácia. Em janeiro de 2020, estudo[275] produzido pela Avaaz em seis países, entre eles o Brasil, indicou que o YouTube segue recomendando – portanto, impulsionando, ajudando a financiar, por meio de sua política de monetização, e lucrando com – conteúdo desinformativo. O levantamento focou em vídeos sobre mudanças climáticas, aquecimento global e manipulação climática – temas que, diante da ausência de definição, não se pode afirmar se seriam enquadrados em "desinformação danosa" ou "conteúdo limítrofe" pela plataforma.

De acordo com o levantamento, dos cem vídeos mais sugeridos pelo YouTube na busca por "manipulação climática", 21% continham informações comprovadamente falsas ou enganosas; por "aquecimento global", 16%; e, por "mudança climática", 8%. Entre elas, afirmações de que a ação humana não é realmente responsável por esses fenômenos. Juntos, eles somavam mais de 21 milhões de visualizações. Na avaliação do relatório, ao funcionar desta maneira, o algoritmo de recomendações do YouTube está levando desinformação a quem não seria exposto a tais conteúdos de outra forma, além de remunerar financeiramente quem os propaga.

275. https://www.bbc.com/portuguese/geral-51111499

Problema equivalente tem sido constatado em casos de conteúdos relacionados à Covid-19, onde a política de risco de danos da plataforma se aplica. Em maio de 2020, o YouTube emitiu comunicado informando a proibição do envio de conteúdos disseminando informações médicas incorretas que contrariassem as orientações da Organização Mundial da Saúde (OMS) ou de autoridades locais sobre tratamento, prevenção, diagnóstico e transmissão do novo vírus. Também disse que bloqueou a monetização de vídeos que trouxessem filmagens angustiantes, com pessoas sofrendo por causa da Covid-19, e com desinformação médica, como declarações sem embasamento sobre a causa da doença, promoção de medicamentos e remédios perigosos. Até a ocasião deste anúncio, vídeos como o documentário "Plandemic", de 26 minutos, baseado em afirmações enganosas sobre a pandemia, seguiam no ar[276], com milhares de visualizações em todo o mundo, mostrando a fragilidade da aplicação das diretrizes do YouTube para informações relacionadas à saúde que já estavam em vigor.

Em 2018, pesquisadores indianos já haviam alertado para os riscos das pessoas se informarem pela plataforma em momentos de emergências sanitárias globais[277]. Durante a pandemia do Zika vírus, eles analisaram mais de cem vídeos no YouTube e concluíram que 23,8% deles propagavam conteúdos enganosos sobre a doença. O estudo também revelou que tais vídeos eram mais propagados do que aqueles considerados informativos e de qualidade superior. Ao final, os pesquisadores afirmaram a necessidade de melhorias na curadoria e autenticação de informações sobre saúde nas plataformas de vídeo, principalmente em contextos de pandemia.

Chegou a Covid-19 e o desafio está longe de ser superado. Estudo da Universidade de Michigan publicado em julho de 2020 apontou que pelo menos um quarto dos vídeos que circulavam

276. https://carnegieendowment.org/2020/03/03/eu-code-of-practice-on-disinformation--briefing-note-for-new-european-commission-pub-81187

277. https://www.tandfonline.com/doi/full/10.1080/20477724.2018.1507784

sobre o novo coronavírus na plataforma não continham embasamento científico válido[278]. A pesquisa teve o objetivo de analisar como os algoritmos do YouTube contribuem para a disseminação desse tipo de conteúdo, e concluiu que, apesar de todas as declarações do Google sobre aprimoramentos nesse sentido, quanto mais informação duvidosa é consumida, mais conteúdo semelhante segue sendo sugerido. E que, quanto maior a popularidade de um vídeo sobre o tema, menor sua probabilidade de contar com fundamentos médicos comprovados.

No Brasil, estudo do INCT.CC, em parceria com o Centro de Análise da Liberdade e do Autoritarismo (Laut) e o Centro de Estudos e Pesquisas de Direito Sanitário (Cepedisa) da Universidade de São Paulo, encontrou três redes coesas, com vocabulários específicos e campanhas relevantes de desinformação sobre a Covid-19 operando no YouTube. Ao todo, a pesquisa "Ciência Contaminada – analisando o contágio de desinformação sobre o coronavírus via Youtube" (MACHADO et al., 2020) analisou 11.546 vídeos sobre o tema. Juntos, os vídeos somaram 486 milhões de visualizações e 16,7 milhões de curtidas. Entre fevereiro e março de 2020, os pesquisadores analisaram o conteúdo verbal de 1.146 horas de gravação, 145.993 comentários e 10.434 descrições, e detectaram interesses políticos, econômicos e pessoais por trás dos vídeos.

A primeira rede traz narrativas de cunho geopolítico, acusando a China de produção do vírus como arma biológica. A segunda, de cunho religioso, enquadra a pandemia num contexto de simbologia bíblica, caracterizada como uma praga divina ou parte do apocalipse. A terceira rede é centrada na figura de médicos para quem o coronavírus virou oportunidade de negócio, com a monetização de conteúdos sensacionalistas e ofertas de produtos que prometem proteção contra a doença. Todas contam com a presença de lide-

278. https://olhardigital.com.br/coronavirus/noticia/algoritmos-do-youtube-contribuem-para-disseminar-fake-news-sobre-saude/103724

ranças políticas, religiosas e médicas que, na avaliação do estudo, possuem certa autoridade perante o público ao qual se dirigem.

Também em março de 2020, o site *The Intercept Brasil* analisou dez canais de extrema direita que têm tratado da pandemia. Pelo menos seis seguiam ganhando dinheiro com anúncios em vídeos, mesmo depois do YouTube ter decidido reduzir exibição de anúncios em conteúdos relacionados à Covid-19 que não seguissem regras de difusão de informações confiáveis sobre a doença[279]. Segundo o Google, uma das razões para que vídeos estejam conseguindo burlar as regras pode ser o fato de que, durante a pandemia, a empresa está trabalhando com um número reduzido de revisores, os profissionais que fiscalizam se os conteúdos violam os termos da empresa.

Por fim, cabe problematizar a disseminação de conteúdos desinformativos, que também pode resultar da ferramenta de autopreenchimento na busca do YouTube. Em março de 2020, o Digital Forensic Research Lab, do Atlantic Council, especializado na análise do fenômeno, divulgou um episódio em que, ao pesquisar sobre o coronavírus na Colômbia, o mecanismo sugeriu o preenchimento "Coronavirus en Colombia Alex Backman"[280]. Ao clicar, o link levava para o canal *Consciencia Radio*, onde Backman, desde 2008, propaga teorias da conspiração. O canal possuiu mais de um milhão de seguidores e, desde o início da pandemia, tem publicado dezenas de vídeos contrariando informações e recomendações médicas sobre a Covid-19, mesmo depois de ter sido confrontado inúmeras vezes por agências de checagem.

A exceção positiva de avaliação das medidas adotadas pelo YouTube veio de um estudo publicado em novembro de 2020 pelo departamento de ciência da computação da Virginia Tech University, dos Estados Unidos. A pesquisa mostrou que, apesar do histó-

279. https://theintercept.com/2020/03/25/youtube-coronavirus-fake-news-olavo/

280. https://medium.com/dfrlab/conspiracy-theorist-gets-thousands-of-youtube-views-spreading-coronavirus-rumors-e53ca7faa013

rico de usuários que assistem vídeos sobre teorias da conspiração resultar em mais vídeos desse tipo sendo recomendados para os mesmos, há uma exceção em prática: vídeos antivacina[281]. A equipe liderada por Tanushree Mitra pesquisou mais de 56 mil vídeos relacionados a teorias como a da Terra plana, a de que o homem não pisou na lua, sobre os atentados de 11 de Setembro e sobre riscos e falta de eficácia de vacinas e auditou os algoritmos de busca e recomendação a partir deles. E concluiu que o YouTube é mais efetivo ao afastar usuários de teorias antivacina do que de outras conspirações, indicando sempre conteúdos que desmentem essa ideia ou comprovam sua importância – mesmo que o histórico e o comportamento do usuário seja de ver vídeos sobre este tema.

O resultado positivo não elimina, entretanto, o forte indício, resultado de diversos estudos, realizados em diferentes países e baseados em metodologias consistentes, de que as regras de combate à desinformação no YouTube, em geral, não têm sido suficientemente efetivas, mesmo em casos de conteúdos de verificação mais objetiva.

Especialistas começam a temer que, mesmo que o YouTube desabilite totalmente seu sistema de recomendações, já estaria tarde para conter a onda extremista que atinge países como os Estados Unidos e o Brasil, porque seus canais de comunicação já operam em conjunto, e a resposta não poderia vir apenas de uma plataforma. No dia 14 de agosto de 2020, um dos colaboradores do site conspiracionista de extrema-direita *Infowars* publicou um vídeo de 90 minutos no YouTube repleto de alegações acerca de um golpe de Estado contra o presidente dos Estados Unidos, Donald Trump[282]. O material foi linkado em diversas páginas e publicado em redes de direita, compartilhado mais de 53 mil vezes no Facebook e citado em outros canais de extrema-direita no próprio YouTube. A plataforma removeu o conteúdo um dia depois, mas mais de um milhão de pessoas assistiram ao conteúdo neste intervalo.

281. https://www.businessinsider.com/study-youtube-anti-vax-conspiracy-theories-2020-11
282. https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories/

Ao mesmo tempo, a adoção da ideia de se trabalhar com "conteúdos limítrofes", sem definir o que cabe dentro desse guarda-chuva, pode resultar na moderação indevida e abusiva de conteúdos, cuja análise é comprometida sobretudo pela ausência de indicadores sobre isso nos relatórios de transparência da plataforma. Se cabe destacar e reconhecer que o YouTube possui um detalhado procedimento para remoção de conteúdos que violam as Diretrizes da Comunidade e que, para os produtores de conteúdo, elas também são bastante exemplificativas, a nova prática na moderação de conteúdo, da maneira como implementada, amplia os riscos de cerceamento à liberdade de expressão na rede.

3.4.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

No esforço de oferecer mais contexto aos conteúdos pesquisados, o YouTube, a depender das buscas feitas, inclui painéis de informação com dados extra aos vídeos publicados, de fontes como a Wikipedia e a Enciclopédia Britânica. A plataforma diz que os painéis de informação são exibidos quando os usuários pesquisam sobre assuntos históricos e cientificamente comprovados ou ainda sobre teorias da conspiração, que podem ser objeto de práticas de desinformação. E que, a depender do tema, se agências de checagem de fatos também tiverem conteúdo produzido sobre o assunto em questão, um outro painel pode incluir links para tais verificações.

O YouTube divulga os critérios utilizados para credenciar checadores, mas informa que não endossa qualquer verificação exibida nos painéis de informação; que se o usuário não concordar com os dados de um artigo específico, deve entrar em contato com a agência que publicou o conteúdo. A checagem, entretanto, como mencionado no capítulo 2, não é sobre cada vídeo individualmente, mas baseada no termo utilizado pelo usuário para a sua busca.

Na ocasião do estudo sobre desinformações relacionadas a mudanças climáticas, a Avaaz divulgou que 64% dos vídeos que continham desinformação encontrados no levantamento continham um painel com link para a Wikipedia, onde o usuário tinha acesso a informações gerais sobre o tema. Entretanto, os pesquisadores criticaram o fato de não haver um alerta aos usuários de que os vídeos continham informações enganosas – além, claro, deles seguirem sendo recomendados pelo YouTube apesar da própria plataforma já ter identificado algum tipo de problema no conteúdo.

Desde 2017, o YouTube informa que também prioriza o que chama de "vozes autorizadas" ou "de credibilidade", como canais de jornalismo, nas buscas por informações e nas recomendações de "próximos vídeos" da plataforma. Afirma que, em 2019, houve um crescimento de 60% na visualização de canais de notícias confiáveis. Em 2020, o aumento teria sido de 75%. E que, em média, em nível global, 93% dos vídeos mostrados nos 10 principais resultados da busca são de canais de alta credibilidade.

Quando o tema são eventos ou notícias sobre política, a plataforma também prioriza esses canais, na busca e nas recomendações. A prateleira "Notícias", dentro de "Em Alta", destaca o que o YouTube considera canais jornalísticos de qualidade. Em março, uma sessão de notícias sobre a Covid-19 passou a ser disponibilizada na homepage do YouTube, trazendo vídeos de notícias sobre a pandemia produzidos por veículos de informação e autoridades sanitárias locais.

Uma busca feita no âmbito desta pesquisa, utilizando os termos "hidroxicloroquina cura Covid" mostrou um banner com link para "informações sobre Covid do Ministério da Saúde". Entretanto, entre as 10 primeiras respostas mostradas na busca, três traziam notícias de veículos da imprensa com reportagens em que o medicamento era apontado como recomendado para o tratamento do novo coronavírus, mesmo depois das orientações da OMS em sentido contrário. Dois dos vídeos exibidos na busca eram da TV

Record, um deles com a chamada "Idosa é curada da Covid em tratamento com uso de cloroquina". O terceiro era TV BrasilGov, com uma reportagem sobre a visita da ministra Damares Alves a um hospital no interior do Piauí, que utiliza a hidroxicloroquina no tratamento da Covid-19. Ou seja, 30% das buscas indicaram conteúdos que, mesmo vindo de veículos da chamada imprensa tradicional, traziam informações cientificamente contestadas, mostrando que o critério de classificação do YouTube para fontes de credibilidade também apresentas brechas, quando se trata da conteúdos específicos.

A pesquisa *Ciência Contaminada*, acima citada, apontou a existência de uma rede de vídeos relacionada ao tema coronavírus que reúne canais jornalísticos que trazem informações científicas sobre a doença. O nível de desinformação encontrado nessa rede foi substancialmente menor que o encontrado nas demais: 4,8%. De acordo com os responsáveis pelo estudo, a circulação de vídeos advindos dos meios de comunicação tradicionais é relevante e constitui um reduto de informação importante em meio a um complexo ecossistema de desinformação. No entanto, como visto, mesmo grandes veículos de comunicação propagam desinformação. No relatório do INCT.CC, na amostragem dos vídeos de maior impacto, com mais de 100 mil visualizações, três deles continham "informações falsas ou imprecisas" ditas pelo comentarista Alexandre Garcia.

O dado é preocupante, sobretudo num contexto de isolamento, em que o número de horas de vídeos consumida por dia por pessoa, mais que dobrou. E quando se considera que a campanha para as eleições municipais de 2020 seria feita sobretudo por esses canais. Nas eleições de 2018 nos Estados Unidos e para o Parlamento Europeu em 2019, quando usuários pesquisavam por um candidato/a na plataforma, o YouTube oferecia junto aos resultados um painel com informações adicionais sobre a pessoa, como afiliação partidária e distrito pelo qual concorria. Nos

Estados Unidos, o Google também está modificando suas regras de propaganda política, para dar maior visibilidade a quem comprar anúncios eleitorais na plataforma. Infelizmente, não há previsão dessas medidas serem adotadas no Brasil.

Por fim, cabe destacar que as informações disponibilizadas pelo YouTube acerca dos canais e conteúdos que deixaram de ser monetizados por disseminarem conteúdo que violam as políticas da plataforma impedem qualquer análise mais consistente sobre o impacto das medidas adotadas, de modo que não é exagero afirmar que a rede social segue lucrando com a ação das redes de desinformação em seu interior. Como recomendou o Conselho Nacional de Direitos Humanos, "as plataformas devem adotar políticas que garantam transparência sobre o seu funcionamento e as regras das suas comunidades e que ampliem o controle dos usuários sobre os conteúdos que publicam e acessam, incidindo sobre o chamado efeito bolha e a estrutura de monetização que estimula a criação e difusão das chamadas 'notícias falsas'. Além do respeito às normas infralegais e internacionais em matéria de direitos humanos e a adequação das políticas das plataformas às especificidades sociais e culturais dos países onde gerem seus modelos de negócios" (CNDH, p. 4).

3.4.4 MEDIDAS CORRELATAS

Somente em 2017 o YouTube deixou de recomendar vídeos de conteúdo supremacista, o que resultou numa redução de 80% de sua visualização. E apenas em 2019 eles passaram a ser totalmente proibidos, junto com conteúdos contendo ameaças implícitas e comportamento abusivo (como assédio). Vídeos que se aproximam do discurso de ódio, mas não são caracterizados como tal, podem seguir no ar, deixando apenas de serem monetizados.

O resultado de tais mudanças ainda precisará ser avaliado, mas na avaliação de pesquisadores como Zeynep Tufekci, professora

da Escola de Informação da Universidade da Carolina do Norte, as recomendações do YouTube seguem padrões extremistas nos mais diferentes tópicos[283], sendo que, em temas políticos, conduz o usuário para vídeos sobre teorias da conspiração tanto de esquerda quanto de direta. Tufekci chegou a denunciar publicamente como o YouTube estava lhe recomendando conteúdos da extrema direita americana após ela ter visto um único vídeo de Donald Trump.

Guillaume Chaslot, ex-funcionário do YouTube que trabalhava na área de recomendações e foi demitido da empresa, também denunciou a tendência de radicalização nos resultados das recomendações para temas relacionados a teorias da conspiração[284]. Na pesquisa, ao questionar se a Terra era plana ou redonda, 90% das recomendações foram na teoria da conspiração da Terra plana. Ao questionar "quem era o Papa", 85% das recomendações associavam a autoridade católica a termos como diabo, anticristo ou satânico. Ao buscar conteúdo sobre o aquecimento global, vídeos falando que se tratava de boato apareceram em 70% das recomendações. Para Chaslot, a plataforma de recomendações do YouTube foi decisiva nas eleições de Trump, espalhando notícias falsas e teorias da conspiração, e o algoritmo vendido como neutro pelo Google ajudou a garantir audiência para vários vídeos conspiratórios.

Na avaliação da brasileira Yasodara Córdova, pesquisadora associada da Digital Harvard Kennedy School e do Berkman Klein Center em Harvard, por aqui o quadro não é diferente. Basta assistir a um vídeo de extrema direita que as recomendações vão garantir que você se aprofunde cada vez mais no ódio. Ela afirma que existe um incentivo para os produtores fazerem vídeos cada

283. TUFECKI, Z. Youtube: the great radicalizer. The New York Times, 10 mar. 2018. Disponível em: <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>. Acesso em: 26 fev. 2019.

284. CHASLOT, G. How Youtube´s AI boosts alternative facts. Guillaume Chaslot Medium, 31 mar. 2017. Disponível em: <https://medium.com/@guillaumechaslot/how-youtubes-a-iboosts-alternative-facts-3cc276f47cf7>. Acesso em: 02 nov. 2018

vez mais extremos e assim prender a audiência o máximo possível. Citando estudo de outro pesquisador do Berkman Klein Center de Harvard, Jonas Kaiser, que analisou mais de 13 mil canais de extrema direita no YouTube, Yasodara explica que as redes extremistas seguem conectadas na plataforma[285].

"É uma rede fértil para circular a ideia de que políticas afirmativas para negros são parte de uma conspiração para acabar com a raça branca ocidental, por exemplo, o delírio de que vacinas são parte de um plano para acabar com determinadas populações em um experimento ou até a história de que as eleições brasileiras estariam em risco por uma suposta fraude nas urnas eletrônicas", afirmou. "Os dados levantados por Kaiser mostram que o esquema de recomendação do YouTube conecta diversos canais que poderiam estar mais isolados sem a influência do algoritmo, ajudando a unir a extrema direita".

Quando afirmou que as empresas devem assumir compromissos políticos de alto nível para manter plataformas onde os usuários possam desenvolver suas opiniões, se expressar livremente e acessar informações de todos os tipos de maneira consistente com os padrões de direitos humanos, o então relator das Nações Unidas para a liberdade de expressão, David Kaye, foi muito enfático diante de um desafio que persiste em todas as plataformas digitais. O YouTube ainda tem muitos passos a dar nesse sentido.

3.5 TWITTER

O Twitter não está no topo das redes sociais em termos de preocupações dos usuários quando o tema é desinformação. De acordo com o *Digital News Report 2020*[286], enquanto cerca de um terço

285. https://theintercept.com/2019/01/09/youtube-direita/

286. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

das pessoas entrevistadas em todo o mundo tem preocupações com a disseminação de notícias falsas no Facebook, no Twitter o percentual é de apenas 5%. No Brasil, o país que mais se preocupa se uma notícia é verdadeira ou não, o foco no Twitter é ainda menor: somente 3% dos usuários de Internet. Parte dessa diferença é fruto da própria relevância da plataforma. Enquanto mais de 140 milhões de brasileiros usam o Facebook, o Twitter é acessado por apenas cerca de 15 milhões pessoas no país.

Mas isso não reduz a importância do problema nesta rede social, marcada por dois fatores significativos quando se trata de disseminação de desinformação: o uso de robôs e o fato de que, também de acordo com o Instituo Reuters, os políticos estão entre as principais fontes de desinformação na Internet. No Brasil, eles ocupam o topo da lista. Se há uma rede social largamente utilizada pelos políticos, aqui e no mundo, esta rede é o Twitter. E se ela conta com uma política de exceção, que pode manter disponíveis postagens de membros do governo e de potenciais e atuais detentores de cargos eletivos se isso for considerado de interesse público, torna-se muito relevante compreender o que a plataforma está fazendo para combater a prática de desinformação em seu interior.

O Twitter entende que a disponibilização de posts considerados de interesse público por trás de um aviso que fornece contexto ao usuário e informa a violação praticada representa "o equilíbrio certo entre permitir a liberdade de expressão, promover a responsabilidade e reduzir o dano potencial causado por esses Tweets" – como declarou em entrevista a esta pesquisa. Num país em que a desinformação se transformou numa das principais armas políticas do governo federal, a validade dessa exceção precisa ser reavaliada de modo permanente.

3.5.1 ABORDAGEM DO FENÔMENO

A empresa reafirma, em todos os seus comunicados e posicionamentos, que o Twitter não se considera um "árbitro da verdade" e que, por isso, não trabalha com verificação de fatos nem com uma definição de desinformação para validar a autenticidade ou a precisão dos conteúdos publicados por seus usuários. O foco da plataforma tem sido o de fornecer contexto para ajudar as pessoas a se decidirem nos casos em que o teor de um tweet é contestado.

Somente em casos de mídias manipuladas tecnicamente, de conteúdos que possam comprometer a integridade de eleições e, mais recentemente, no caso da pandemia da Covid-19, que o Twitter modera conteúdos relacionados com desinformações – como mostraremos no próximo tópico. Mesmo assim, afirma que somente age em função do potencial de dano que tais informações enganosas possam causar aos usuários.

Nesse sentido, o Twitter vai ao encontro de recomendações internacionais, como as analisadas no capítulo 1 deste estudo. No caso dos documentos da Comissão Europeia, por exemplo, que trazem definições para o conceito de desinformação, a ideia do objetivo de causar danos é parte intrínseca do que os especialistas consideraram ser abarcado por este fenômeno. E, para eles, o risco de danos inclui ameaças aos processos e valores políticos democráticos, que podem visar setores como saúde, a ciência e a segurança.

Entretanto, mesmo nesses casos, o fato de não apresentar uma definição pública para o termo e de declarar que consulta especialistas para caracterizar se uma informação é enganosa ou não – sem que tais especialistas sejam públicos ou que haja critérios públicos para sua escolha –, compromete a compreensão dos usuários sobre como a plataforma pode agir nesses casos. Considerando que todas essas informações estão disponibilizadas apenas em comunicados, descentralizados, publicados no blog do Twitter – não organizados dentro de uma política contra desinformação no

âmbito das regras e políticas da plataforma[287], saber o que pode ou não ser publicado sobre o tema se torna uma tarefa ainda mais difícil, comprometendo significativamente o princípio da transparência no funcionamento da plataforma.

Não à toa, o Twitter foi duramente criticado em outubro de 2020 quando, ao lado Facebook, moderou postagens trazendo uma reportagem do jornal *New York Post* que acusava Hunter Biden, filho do então candidato à presidência dos Estados Unidos, Joe Biden, de tráfico de influência. A plataforma considerou a reportagem "insegura" e proibiu seu compartilhamento. Ao mesmo tempo, bloqueou temporariamente as contas do presidente Donald Trump e da secretária de imprensa da Casa Branca, Kayleigh McEnany, alegando que ambos teriam publicado materiais obtidos ilegalmente – o que violaria as políticas da plataforma. Especialistas em desinformação nos Estados Unidos e no Brasil acusaram o Twitter de censura privada, pelo fato da empresa ter tomado tais medidas de maneira preventiva, antes de qualquer verificação independente comprovar que a reportagem tinha informações falsas.

Nos comunicados específicos sobre desinformação, o Twitter também inclui, como parte de seus esforços contra o fenômeno, a proibição do uso de robôs para a manipulação das funcionalidades da plataforma. Nessa medida, considerada em geral correlata dentro dos critérios deste estudo, o Twitter informa que reduz a visibilidade de tweets dessas contas e de contas potencialmente *spammers*, podendo chegar a suspendê-las se houver desrespeito aos termos de uso da plataforma.

Entretanto, são diversos os estudos e iniciativas – como o PegaBot[288], projeto do ITS Rio e do Instituto Tecnologia e Equidade, e o Bot Sentinel[289] – que mostram como o uso dos chamados robôs segue sistemático na plataforma, com altos riscos de interferência no debate público na rede, como veremos adiante.

287. https://help.twitter.com/pt/rules-and-policies
288. https://twitter.com/pegabots
289. https://botsentinel.com/

3.5.2 MODERAÇÃO DE CONTEÚDO

A política contra mídias manipuladas, que está em vigor no Brasil desde o dia 5 de março de 2020, é um mecanismo interessante, já que baseado em alterações técnicas, como edições, que podem ser feitas nas mídias publicadas na plataforma – e, portanto, menos sujeito a análises subjetivas de conteúdo por parte da empresa. Dentro desta política, o Twitter pode remover, por exemplo, conteúdos chamados de "*deepfakes*", além de aplicar marcações aos conteúdos, exibir avisos antes que as pessoas curtam ou compartilhem a postagem, reduzir sua visibilidade e também fornecer link para mais explicações sobre a manipulação aplicada.

Já a política de integridade nas eleições é totalmente baseada na análise do conteúdo para que o Twitter defina o tipo de moderação que será aplicado. Uma vez mais, ela ocorre apenas em caso de desinformação enganosa que possa causar dano. Conteúdos polarizados, com pontos de vista controversos sobre as eleições, ou até mesmo preconceituosos contra candidatos ou partidos, não são passíveis de moderação.

Sobre as eleições no Brasil, as iniciativas para o combate à desinformação seguem tímidas, se limitando à existência de um canal para o recebimento de requisições das autoridades eleitorais brasileiras e à verificação dos perfis oficiais dos principais candidatos e partidos. Para efeito de comparação, nas eleições dos Estados Unidos em 2018 e da União Europeia em 2019, o Twitter identificou todos as candidaturas e criou um selo em cada uma delas, assim como nos tweets e retweets publicados por elas, para informar os usuários da rede. Nos Estados Unidos, também foi colocada no ar uma ferramenta para permitir aos usuários denunciarem informações enganosas sobre como participar da eleição.

As medidas, mais efetivas que as adotadas no Brasil, também foram uma resposta a estudos que chegaram a mostrar, em alguns cenários, que até 25% dos tweets sobre as eleições presidenciais

norte-americanas continham notícias falsas ou fortemente enviesadas. Um deles, conduzido pelo City College de Nova Iorque, em parceria com a Universidade Católica de Louvain, na Bélgica[290], partindo de uma base de 171 milhões de postagens, de 2,2 milhões de usuários, analisou 30 milhões de tweets que continham links para veículos de comunicação. Com base numa classificação desses veículos feita por diferentes agências de verificação de fatos e grupos de pesquisa listados no estudo, a investigação revelou que 10% dos tweets direcionavam os usuários para websites com difusão sistemática de desinformação, incluindo teorias da conspiração, e 15% apontavam para páginas com conteúdo altamente enviesado.

Quando consideradas as postagens originadas por aplicações não desenvolvidas pelo Twitter ou com base na sua API pública, a taxa de posts que direcionavam os usuários para esses sites era quatro vezes maior do que a daqueles direcionando para páginas da chamada imprensa tradicional – o que, para os pesquisadores, sugeriu um grande número de bots na difusão de desinformação.

Na "Declaração Conjunta sobre Liberdade de Expressão e Eleições na Era Digital", mencionada no capítulo 1, os organismos multilaterais signatários recomendaram, entre outros pontos, que as plataformas realizassem esforços suficientes para adotar medidas que permitissem aos usuários acessar várias ideias e perspectivas políticas, e adotassem medidas de transparência em relação ao uso e impacto que as ferramentas automatizadas podem causar no processo eleitoral - incluindo até que ponto essas ferramentas afetam a coleta de dados, publicidade direcionada e sua divulgação (ONU, OSCE e OEA, p. 4). Esta pesquisa não constatou perspectivas de que o Twitter avance ao encontro de tais recomendações.

290. BOVEN, Alexandre; HERMAN, A. Makse. Influence of fake news in Twitter during the 2016 US presidential election. **Nature communications** 10, no. 1 (2019): 1-14. Disponível em <https://www.nature.com/articles/s41467-018-07761-2>. Acesso: 10 mai. 2020.

No campo das eleições, a medida recente mais importante tomada pela plataforma foi a proibição da veiculação de propaganda política na rede social durante os períodos de campanha eleitoral. Tal propaganda, autorizada até o final de 2019 em alguns países, incluía não apenas conteúdos diretamente postados pelas candidaturas, mas também identificava "anúncios temáticos" que identificassem um candidato ou defendessem mudanças legislativas de importância nacional – como aborto, direitos civis, mudanças climáticas, armas, imigração, impostos, comércio e seguridade social. Os conteúdos ficavam disponíveis num Centro de Transparência de Anúncios.

O banimento, fundamental para coibir a influência econômica nos processos eleitorais e a comercialização do debate público, também tem um impacto positivo no enfrentamento à desinformação. Do ponto de vista da transparência, entretanto, seria interessante que o Twitter mantivesse as informações sobre essa política na plataforma, visando uma melhor compreensão de como ela funcionava antes de sua descontinuação.

Sobre a moderação de conteúdo relacionada a desinformações sobre a Covid-19, o Twitter ampliou sua definição de "dano" para incluir conteúdos que contrariem as recomendações das autoridades mundiais e locais de saúde. Ao informar que a prioridade é remover conteúdos que façam um chamado claro a ações que possam colocar diretamente a saúde das pessoas em risco, a plataforma trouxe uma lista de exemplos concretos do que entende por esses chamados, facilitando a compreensão dos usuários. Casos de conteúdos com risco de dano menos severo podem receber um rótulo ou aviso.

Em maio de 2020, entretanto, a Universidade Carnegie Mellon, nos Estado Unidos, divulgou estudo sobre o país apontando que 45% das contas no Twitter que estão postando sobre a Covid-19 provavelmente são robôs, e que elas já produziram cerca de

100 narrativas falsas diferentes sobre o tema[291]. Para determinar as chances de um perfil ser um bot, os autores avaliaram informações como o número de seguidores, a frequência de postagem e as menções recebidas na plataforma. Mais de 200 milhões de tweets sobre novo coronavírus foram analisados. De acordo com a professora de ciência da computação da Carnegie Mellon, Kathleen Carley, responsável pela pequisa, muitos desses perfis foram criados em fevereiro e, desde então, estão divulgando informações falsas como supostos conselhos médicos, possíveis curas, teorias da conspiração sobre a origem do vírus e pressão para acabar a quarentena.

Em análise produzida a pedido da revista *Business Insider*, o criador do Bot Sentinel, Christopher Bouzy, identificou que robôs e trolls nos Estados Unidos estão usando hashtags como #ReOpenNC, #ReopenAmericaNow, #StopTheMadness, #ENDTHESHUTDOWN e #OperationGridlock para disseminar desinformação no Twitter[292]. As narrativas construídas nas postagens incluem afirmações de que os Democratas querem prejudicar a economia do país para atacar o presidente Donald Trump e que também querem impedir as pessoas de votar nas eleições que se aproximam.

O Twitter afirma que, até julho de 2020, removeu mais de 14.900 postagens contendo conteúdo potencialmente danoso sobre o tema da Covid-19 e seus sistemas questionaram 4,5 milhões de contas. O desafio imposto a contas suspeitas pode ser a confirmação de um e-mail ou telefone ou a digitação de um código captcha. A plataforma não divulga, entretanto, números por país, de maneira que é impossível saber se as medidas estão respondendo ao tamanho do desafio da desinformação sobre a pandemia no Brasil. A plataforma admitiu que as análises automatizadas aumentaram significativamente neste contexto e que erros podem

291. https://www.uol.com.br/tilt/noticias/redacao/2020/05/24/estudo-quase-metade-das-contas-postando-sobre-covid-no-twitter-sao-robos.htm

292. https://botsentinel.com/newsroom/press

acontecer com maior frequência. Por isso, declarou que não vai suspender permanentemente nenhuma conta apenas com base na decisão de seus sistemas automatizados, priorizando a revisão humana para casos de maior impacto.

A decisão é fundamental, sobretudo quando se considera que uma das grandes preocupações nos referenciais de direitos humanos acerca da moderação de conteúdo é o chamado devido processo, que deve incluir mecanismos de notificação aos usuários sobre medidas aplicadas a suas postagens e o direito de questioná-las. Em suas regras, o Twitter diz que informa os usuários sobre as ações tomadas contra suas contas, bem como os motivos que a justificam e o direito que todos têm à apelação[293], que pode ser feita mediante o preenchimento de um formulário on-line. No processo de apelação, se o Twitter considera que a medida adotada foi correta, ele responde o usuário com informações sobre a política violada[294]. Não há entretanto, na plataforma, nenhuma informação detalhada sobre o processo de apelação, sobre se a plataforma responde a todas as contestações e denúncias, quanto tempo o usuário deve esperar até receber uma resposta e nenhuma explicação adicional sobre o que ocorre se a plataforma errar em sua moderação.

De acordo com as regras, medidas adotadas contra contas acontecem depois que as regras do Twitter foram violadas de maneira muito séria ou sistemática depois do usuário ter recebido notificações da plataforma. Antes disso, quando se trata de um conteúdo específico, o Twitter o torna temporariamente indisponível e pede que o responsável pela postagem a remova, explicando a regra que foi violada. Contas também podem ser colocadas temporariamente no modo leitura, perdendo sua capacidade de tweetar, retweetar ou dar likes em conteúdos. Tal medida pode durar de 12 horas a 7 dias, dependendo da violação.

293. https://help.twitter.com/pt/rules-and-policies/enforcement-philosophy
294. https://help.twitter.com/en/rules-and-policies/enforcement-options

3.5.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

No relatório "Uma abordagem multidimensional à desinformação", o Grupo de Alto Nível independente sobre notícias falsas e desinformação on-line, formado pela Comissão Europeia, recomenda, como medidas de empoderamento dos usuários, o desenvolvimento, pelas plataformas digitais, de marcadores e selos de qualidade, assim como rótulos de conteúdos verificados e disponibilização da maior variedade possível de notícias relevantes aos cidadãos. A estratégia é "diluir" a desinformação com maior transparência e visibilidade de notícias confiáveis.

Desde que implementou a ferramenta #ConheçaosFatos, para responder às campanhas de desinformação contra vacinas nos Estados Unidos, o Twitter tem ampliado esta linha de atuação, que se tornou bastante significativa no contexto da pandemia da Covid-19. O objetivo é facilitar a oferta de informações factuais e o que chama de "contexto adicional de fontes confiáveis", para que as pessoas possam tomar decisões informadas sobre o novo coronavírus. Diferentes ferramentas foram colocadas no ar neste sentido e a agregação de conteúdos sobre a Covid-19 e sua disponibilização tem sido feita de maneira acessível e amigável.

Em se tratando de conteúdo científico e tendo a Organização Mundial de Saúde como referência primeira, a tomada de decisão sobre o que entra no rol de fontes confiáveis também pode parecer menos complexa. O Twitter, entretanto, também tem incluído em suas recomendações especialistas, jornalistas em saúde e veículos de notícia, evidenciando algo que já apontamos, neste estudo, na análise sobre o Facebook: o papel editorial de curadoria da plataforma na escolha e definição daquilo que será apresentado aos usuários como verdadeiro e confiável.

E se a Comissão Europeia sugeriu a disponibilização da maior variedade possível de notícias relevantes aos cidadãos, ela também recomendou que tal medida fosse baseada em critérios objetivos,

em indicadores de fiabilidade das fontes de conteúdos, aprovados por associações de órgãos de comunicação social, em consonância com os princípios e processos jornalísticos – algo inexistente, ao menos de maneira transparente, nas práticas do Twitter.

Ainda no esforço de transparência, os organismos internacionais recomendam que as redes sociais divulguem informações pormenorizadas sobre o comportamento de algoritmos que estabelecem a prioridade de apresentação de conteúdos. E o estabelecimento de sistemas de marcação e regras claras para a operação de robôs nas plataformas, de modo a assegurar que a sua atividade não seja confundida com interações humanas. Tratam-se de duas medidas também ignoradas, até o momento, pelo Twitter.

3.5.4 MEDIDAS CORRELATAS

Ainda acerca dos robôs, o Twitter tem uma política contra spam e manipulação da plataforma extremamente detalhada. O fato de, em 2018, ter realizado mudanças em sua API que resultaram na exclusão de 142 mil aplicativos de acesso à plataforma e do número de contas suspensas em 24 após uma denúncia ter crescido três vezes de acordo com o últimos Relatório de Transparência da empresa mostram uma busca por enfrentar este que parece ser um dos maiores problemas da rede social.

O desafio, entretanto, parece longe de ser superado e não é possível afirmar se as medidas adotadas estão sendo adequadamente implementadas ou, ainda, efetivas. Durante as eleições presidenciais de 2018, por exemplo, pesquisa realizada pela Diretoria de Análise de Políticas Públicas da Fundação Getúlio Vargas (FGV-Dapp) mostrou que milhares de contas suspeitas de automatização operaram de maneira intensa durante a campanha (RUEDIGER e GRASSO, 2018). Segundo a FGV, somente no segundo turno, 3.198 contas suspeitas trabalharam em prol das

candidaturas de Jair Bolsonaro (43% do total) e Fernando Haddad (28% do total). O ápice das atividades automatizadas em prol dos candidatos foi registrado entre 12 e 18 de setembro de 2018, quando 5,3 milhões de interações no Twitter foram analisadas, envolvendo mais de 712 mil perfis, com 12,9% de retweets apontados como oriundos de contas controladas por máquinas.

Em entrevista concedida em junho de 2020 ao *Sonar*, blog do jornal *O Globo*, o diretor de políticas públicas do Twitter no Brasil, Fernando Gallo, afirmou que, para a empresa, nem toda automação é indevida ou indesejável[295]. "Diversos atores - incluindo veículos de imprensa - utilizam automação para prover diferentes serviços e informações às pessoas no Twitter, como atualizações de tráfego, alertas meteorológicos ou canais de atendimento ao consumidor, por exemplo. Por isso, nosso olhar não se resume a analisar se uma conta é ou não um 'robô', mas sim consiste em uma avaliação holística de seu comportamento quando há suspeita de envolvimento em tentativas de manipular o debate na plataforma", declarou.

Gallo também afirmou que ferramentas de terceiros que utilizam a API pública do Twitter para tentar adivinhar se uma conta é ou não uma automação indevida "costumam valer-se somente de sinais externos dos perfis, como nome, volume e frequência de publicação de tweets, localização descrita na bio ou hashtags utilizadas". Para o diretor, "a análise baseada em sinais externos leva a falsos positivos e, além disso, esses levantamentos não consideram as medidas defensivas do Twitter para garantir que o conteúdo automatizado não influencie as discussões na plataforma, dado que essas iniciativas – como redução da visibilidade das contas em áreas de busca e respostas a Tweets – não são refletidas em tempo real [...] na API pública".

295. https://blogs.oglobo.globo.com/sonar-a-escuta-das-redes/post/entrevista-termo-robo-acaba-sendo-usado-de-forma-generalizada-diz-executivo-do-twitter.html

Ainda no âmbito das medidas correlatas à desinformação, é importante falar das políticas do Twitter contra propagação de ódio e comportamentos abusivos, que proíbem assédios, intimidações, insultos excessivamente agressivos e o silenciamento de vozes. Como explicado no capítulo descritivo, até 2018, o Twitter não tomava medidas próprias para combater conteúdos que violassem suas políticas. A plataforma apenas revisava postagens denunciadas por outros usuários. Essa prática, bastante permissiva, acabou transformando a rede social num dos espaços favoritos para a prática de violências contra grupos minorizados.

Em 2018, a Anistia Internacional lançou o relatório "#TwitterTóxico: Violência e abuso contra mulheres na internet"[296], que mostrou as falhas da empresa em respeitar os direitos das mulheres por conta de sua ineficácia no combate a abusos e agressões motivadas por gênero. Entre elas, ameaças diretas ou indiretas de violência física e sexual, discriminação contra um ou mais aspectos da identidade de uma mulher, assédio, doxing e compartilhamentos de imagens íntimas sem consentimento. A experiência de mulheres negras, de grupos étnicos ou religiosos minoritários, lésbicas, bissexuais ou transgêneras, assim como de mulheres com deficiência, expostas à natureza interseccional de abuso na plataforma, foi ainda mais violenta.

Entre as críticas feitas pelo estudo, bastante aprofundado e que ouviu mulheres de oito países diferentes, estão o fato dos usuários são saberem como as políticas da plataforma são aplicadas e dos moderadores de conteúdo não serem treinados para responderem de maneira adequada às denúncias recebidas. Por isso, a organização internacional recomendou ao Twitter compartilhar exemplos específicos de violência e abuso que não são tolerados; compartilhar dados sobre tempo de resposta a denúncias de abuso;

296. AMNESTY INTERNATIONAL. Toxic Twitter - A Toxic Place For Women. 2018. Disponível em: <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/>. Acesso: 14 dez. 2020.

e garantir que decisões de restringir conteúdo sejam compatíveis com as leis e normas de direitos humanos internacionais. Uma das conclusões da Anistia sobre o Twitter é que "as situações de abuso prevalecem de maneira consistente e às vezes nem são sequer respondidas, ou seja, esse conteúdo permanece na plataforma mesmo violando as regras".

Mais de três quartos (76%) das mulheres entrevistadas declarou ter vivenciado abusos ou assédio na Internet, o que mudou sua maneira de usar as redes sociais. Cerca de 32% das mulheres pararam, por exemplo, de postar opiniões sobre certos temas.

A partir de 2018, com uma nova abordagem pró-ativa, por meio de tecnologias de automação, 20% dos tweets que demandaram alguma medida contra abusos passaram a ser detectados antes de qualquer denúncia. O número cresceu depois para 38% e, de acordo com o último Relatório de Transparência do Twitter, lançado em outubro de 2019[297], mais de 50% dos tweets abusivos publicados entre janeiro e junho de 2019 foram detectados tecnologicamente. Os casos de propagação de ódio tiveram um aumento de 48% de denúncias e a plataforma acionou 133% mais contas por esse motivo em relação ao relatório anterior. No caso de conteúdo abusivo, as contas reportadas foram em número 22% maior, mas o Twitter agiu contra 68% a mais de casos.

As dúvidas sobre se as medidas estão sendo suficientes, entretanto, permanecem. No dia 14 de agosto de 2020, em entrevista ao jornal *O Estado de S. Paulo*, a filósofa Djamila Ribeiro disse que encerrou sua conta no Twitter porque a rede social se tornou um ambiente tóxico para a comunidade negra e, principalmente, para as mulheres negras. Em parceria com a Coordenação Nacional de Articulação das Comunidades Negras Rurais Quilombola (Conaq) e a União de Negras e Negros Pela Igualdade na Cidade de São Paulo, Djamila entrou com uma representação junto

297. https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html

ao Ministério Público Federal para obrigar o Twitter a criar um fundo direcionado ao combate à discriminação e a capacitar sua equipe na questão racial para desenvolver um novo conjunto de regras e condutas para a rede social. O documento cita a tese de doutorado do sociólogo Luiz Valério Trindade, que aponta que as mulheres negras têm 84% mais chances de receber comentários desrespeitosos nas redes sociais em função de seu protagonismo nas plataformas.

"Há tempos que o Twitter me incomoda como uma rede social em que esses ataques acontecem e nada é feito", disse Djamila Ribeiro ao *Estadão*. "Achamos que era importante representar o Twitter aqui no Brasil por fazer essa exploração econômica do racismo e da misoginia, e de não melhorar suas políticas. Não é uma questão do indivíduo, é uma questão da plataforma que permite que esse tipo de coisa aconteça, lucre com isso, e não tome medidas", afirmou a filósofa.

No final de 2020, novas denúncias de racismo algorítmico atingiram o Twitter. Dessa vez, os usuários apontaram que a ferramenta de cortes automáticos nas imagens publicadas na rede social – que tem o objetivo de destacar objetos e pessoas mais importantes da composição – priorizava a exibição de pessoas brancas em relação às negras, que eram removidas do quadro. O Twitter declarou que testes anteriores à adoção da ferramenta não detectaram tal viés e que está corrigindo o problema.

3.6 ANÁLISE COMPARATIVA DAS MEDIDAS DE ENFRENTAMENTO À DESINFORMAÇÃO

Nesta conclusão, oferecemos um olhar global das estratégias antes descritas e analisadas e mapeamos também as especificidades dos métodos empregados. Tal exame é promovido sempre considerando como referenciais valorativos os parâmetros de direitos

humanos discutidos no primeiro capítulo e as formulações acumuladas pelo Intervozes em seus documentos propositivos sobre o tema. Acrescemos no âmbito desta análise ponderações acerca da insuficiência das providências das plataformas e de desafios postos diante da magnitude do problema.

3.6.1 ABORDAGEM DO FENÔMENO

Ao longo da pesquisa, foi possível identificar diferentes abordagens das plataformas estudadas sobre desinformação, embora em nenhum caso tenhamos percebido uma política estruturada de combate ao fenômeno. As ações são recentes e impulsionadas por escândalos ou críticas sobre o papel desses agentes na disseminação do fenômeno. Exemplo dessa postura é o Facebook, que minimizou os questionamentos atribuindo responsabilidade à rede social à desinformação nas eleições de 2016 nos Estados Unidos, mas teve de dar resposta à medida que evidências passaram a aparecer, como o escândalo do uso de dados pela empresa Cambridge Analytica evidenciou.

O WhatsApp também atuou dessa maneira a partir das denúncias do uso do serviço como promotor de ódio levando a linchamentos na Índia desde 2018. O YouTube também foi bastante criticado pela difusão de conteúdos extremos e desinformação sobre temas como mudanças climáticas e pandemia causada pelo novo coronavírus. Já o Twitter entrou na linha de fogo, além das críticas sobre notícias falsas, em razão da profusão de contas automatizadas (bots) e de mídias manipuladas no seu interior. O Instagram ganhou mais visibilidade como espaço de práticas preocupantes como assédio e *bullying*.

Ou seja, enquanto os comunicados institucionais buscam promover a ideia de empresas preocupadas com o problema, percebemos um *modus operandi* muito mais reativo do que proativo das

plataformas, com ações adotadas principalmente a partir de 2019, relegando inclusive orientações de autoridades de direitos humanos que, pelo menos desde 2017, produzem recomendações com sugestões de práticas para combater a desinformação. As pressões se multiplicaram em diversas arenas, seja na profusão de projetos legislativos visando regular direta ou indiretamente a desinformação em diversos países, seja na ação de órgãos reguladores (como tribunais eleitorais) ou em campanhas de boicote, como a Stop Hate For Profit[298]. Mais do que danos potenciais à reputação das empresas, tais movimentos podem ensejar prejuízos financeiros concretos, uma vez que podem impactar nos valores de suas ações na bolsa.

Diante deste cenário, todas elas passaram a lidar de alguma forma com o problema. Algumas (como Facebook, Instagram e YouTube) o fazem de forma mais direta; outros, de maneira menos direta (Twitter e WhatsApp). Enquanto o Facebook (e o Instagram, uma vez que segue a mesma abordagem) se viu pressionado a dar respostas públicas mais claras diante da chuva de críticas, outras plataformas tiveram questionamentos específicos que muitas vezes não se resumiam à desinformação. É o caso do conteúdo extremo e mídias manipuladas no YouTube, dos robôs e contas inautênticas no Twitter e dos disparos em massa no WhatsApp.

A despeito de um repertório variado de medidas, as abordagens das empresas analisadas evidenciaram os distintos "focos" de ação a partir desses questionamentos. O Facebook, principal alvo deles, viu a necessidade de oferecer providências mais explícitas, que resultaram em seu modelo de verificação por agências, sinalização e sanções (como redução de alcance, remoção em casos pontuais e desmonetização). No Instagram, as mesmas regras foram implementadas, embora com atraso. O YouTube buscou dar resposta às afirmações de promotor de conteúdo extremo buscando oferecer

298. https://www.stophateforprofit.org/

mais informação de contexto sobre os temas pesquisados na plataformas e trazendo disposições inclusive do que denominou "conteúdo limítrofe", que deixou de ser recomendado para os usuários. O Twitter deu mais atenção ao monitoramento e fiscalização de contas falsas, de bots e de mídias manipuladas.

O WhatsApp foi a empresa que mais se diferenciou ao apresentar maior resistência a reconhecer e oferecer respostas ao problema, sob a alegação de ser uma plataforma "privada" e pelo compromisso de não acessar os conteúdos das mensagens criptografadas, embora esta e outras pesquisas mencionadas ao longo do texto mostrem que isso não impede a adoção de outras ações. A plataforma trata do tema apenas a título de exemplo de comportamentos ilícitos vedados, como "declarações falsas, incorretas ou enganosas". O foco de sua ação está no combate a disparos em massa, inclusive com mecanismos automatizados, prática proibida por sua política anti-spam.

Em geral, as empresas evitam a análise de conteúdo desinformativo, mas acabam fazendo-o. No caso do Twitter, por exemplo, a empresa afirma que não deseja ser um "árbitro da verdade" e que a profusão de informações serve de "antídoto" à desinformação. Contudo, apesar do modelo de parceria com agências externas (ver próximo ponto) adotado por algumas empresas, a triagem por sistemas informatizados e pelas próprias equipes indica que as plataformas analisam, sim, tais conteúdos. A diferença é se adotam ou não medidas específicas sobre eles, o que fica mais evidente com a aplicação de medidas correlatas, especialmente para casos de conteúdos proibidos em seus padrões da comunidade.

Nenhuma das empresas relatou trabalhar com uma conceituação unificada sobre desinformação, embora tenha sido constatado o emprego de termos diversos, como notícias falsas (Facebook) e informações enganosas (Twitter). Em que pese a polissemia do termo e o debate intenso tanto na sociedade quanto em círculos acadêmicos sobre ele, a explicitação de um conceito seria um

avanço como forma de dar transparência sobre qual tipo de conteúdo pode ser impactado pelas medidas de análise, sinalização ou sanções. O que as plataformas adotam são categorizações.

No Facebook, por exemplo, elas incluem "falso", "parcialmente falso" ou "sátira e paródia", entre outras; no Twitter, "informações enganosas", "afirmações disputadas" e "declarações não verificadas". A matização de diversos "graus" de desinformação seria possível e iria ao encontro do reconhecimento da multiplicidade do fenômeno e da dificuldade de identificar a "verdade". Não abordá-lo diretamente não ameniza o problema e abre margem para outros, como decisões unilaterais e pouco claras.

Por fim, nenhuma das empresas relatou ter uma estrutura específica para abordar a questão da desinformação, o que pode dificultar a coordenação das iniciativas no âmbito de cada organização. Não ficaram claros nas explicações dadas a essa pesquisa os critérios de uso de pessoas naturais na análise e decisão sobre medidas de moderações de conteúdos e eventuais sanções. A análise e o monitoramento dos conteúdos dependem sempre de sistemas automatizados, mesmo quando em última instância a verificação possa ocorrer por agências de verificação de fatos consideradas independentes. O desafio da escala de tratamento dessas mensagens e a solução calcada em decisões automatizadas implica riscos altos, seja pela incapacidade desses sistemas compreenderem contexto (algo fundamental para examinar um caso de desinformação) seja pela alta possibilidade de erros. A falta de estruturas dificulta também o contato com as empresas em busca de respostas, seja por cidadãos comuns, pesquisadores ou até autoridades, como será detalhado a seguir.

As críticas diante da insuficiência das iniciativas por parte das plataformas levou parte delas a instituir colegiados com participação de especialistas e representantes da sociedade civil para auxílio nas ações sobre moderação de conteúdo em geral, que, embora não exclusivamente se dediquem à desinformação, abarcam decisões

sobre o tema. É o caso do Conselho de Supervisão do Facebook e do Conselho de Confiança e Segurança do Twitter. Enquanto o primeiro fica responsável pela análise de conteúdos, o segundo é voltado mais para assessoria na formulação de regras.

3.6.2 MODERAÇÃO DE CONTEÚDO

Em geral, os recursos mais disseminados entre as plataformas para eventuais moderações de conteúdo desinformativos são as ferramentas de denúncias que podem ser feitas pelos usuários. A denúncia, contudo, é apenas um insumo, não vinculado em todas as plataformas à necessidade de análise da reclamação. No caso do WhatsApp, por exemplo, a empresa informa essa alternativa mas não explica que tipo de denúncia pode ser feita nem como ela será encaminhada, estando disponível no aplicativo na seção genérica "Fale conosco".

Outro processo presente em parte das plataformas é a verificação de conteúdos, principalmente por agências externas, consideradas por elas independentes, mas que, além de serem grupos privados, em muitos casos são financiadas pelas próprias plataformas. O Twitter, por exemplo, aportou recursos no Projeto Comprova, assim como o Google (controlador do YouTube). O Facebook destinou verbas à Agência Lupa. A defesa da independência tem entre seus argumentos critérios como os empregados pela Rede Internacional de Checadores de Fatos (IFCN). Esses, contudo, não afastam os riscos da dependência do financiamento dessas plataformas, condição que é reforçada pelo fato de a parceria com determinadas agências ser prerrogativa de cada plataforma.

Quando ocorrem, as verificações são realizadas conforme categorizações da plataforma, e não da agência (como mencionado anteriormente), diretriz necessária para evitar tratamentos diferentes por cada checadora. Mesmo assim, naturalmente a complexidade

de analisar os "tons de cinza" entre um e outro extremo enseja riscos de avaliações questionáveis, razão pela qual a verificação deveria contar com mecanismos de devido processo efetivos para a mitigação de abusos e erros (ver adiante).

O WhatsApp facilitou serviços de checagem cuja operação é baseada no contato pelos próprios usuários, mas não institucionalizou a marcação dos conteúdos por afirmar não trabalhar com a análise das mensagens criptografadas. À pesquisa, informou o firmamento de parcerias com o IFCN, o projeto Comprova no Brasil e com o Verificado no México, mas, além de se tratar de uma prática recente e pontual, como no contexto da pandemia, ela não se aproxima do modelo em que a verificação é veiculada como alerta aos usuários juntamente ao conteúdo. Essa limitação não deve justificar uma defesa da flexibilização da criptografia na rede social, fundamental de ser mantida. Mas revela que suas respostas ao problema demandam um arranjo mais complexo, que considere o uso da plataforma para comunicação massiva.

Nas plataformas onde o conteúdo é verificado (Facebook e Instagram), este é "marcado" (*flagged*). Agentes políticos e acadêmicos apontaram a ineficácia desse mecanismo. Estudos mostraram que os alertas inclusive chamam a atenção para o conteúdo. Mesmo com esse "efeito colateral", parece-nos uma ferramenta necessária. Algumas plataformas passaram a mudar os alertas para que eles ganhassem mais destaque visual. O Instagram alterou os rótulos para as eleições de 2020 dos EUA, a fim de destacar mais essas publicações. Esse tipo de iniciativa é relevante e deve ser parâmetro a ser multiplicado, com destaque visual que auxilie o usuário a diferenciar os conteúdos marcados.

No caso do Facebook, Instagram e Twitter, a moderação abrange também anúncios e conteúdos impulsionados. Nesses casos, há restrições a conteúdos já considerados desinformativos. A extensão das medidas sobre desinformação aos anúncios nos parece fundamental. Isso porque aí reside a dimensão econômica da

prática de produção e difusão de notícias falsas como negócio, o que é explorado pelas "fabricas de conteúdos enganosos" e também pelas próprias plataformas, que se beneficiam com a cobrança de impulsionamentos. Quando as empresas auferem receitas com conteúdos os parâmetros podem e devem ser mais rígidos. No caso dos anúncios com discursos de políticos, as medidas esbarram nas exceções para o discurso desses agentes adotadas por algumas companhias, como Facebook e Instagram. No caso do Facebook, no entanto, com a aproximação das eleições dos EUA de 2020, a companhia anunciou em setembro que alteraria essa regra, proibindo novos anúncios na semana anterior à votação.

No Twitter, o foco está nas chamadas "mídias sintéticas ou manipuladas", que incluem, no termo utilizado em inglês, as chamadas "*deepfakes*". A plataforma analisa se houve edição ou fabricação da imagem ou vídeo e o seu contexto, disponibiliza informações adicionais sobre o tema e aplica sanções. O Facebook também agregou em suas medidas métodos de checagem relacionados a esse tipo de mensagem. A demanda por técnicas de verificação mais sofisticadas para esses conteúdos coloca desafios a todas as plataformas, uma vez que há formas mais sutis, e mesmo verossímeis, de alteração das mensagens (colocando, por exemplo, discursos falsos em vídeos), com potencial ainda mais danoso aos usuários.

Durante a maior parte do período analisado pela pesquisa (2018 a 2020), as plataformas resistiram em remover conteúdos desinformativos como o fazem em outras categorias, a partir do que é unilateralmente definido em suas diretrizes. Mas essa postura começou a mudar, principalmente no contexto da pandemia do novo coronavírus, com a profusão de desinformação sobre a doença e sobre "formas milagrosas" de cura. Os riscos graves à saúde da população pressionaram as plataformas a responderem de forma mais ágil e dura sobre isso. Não por acaso, as companhias passaram a remover publicações relacionadas ao tema, em espe-

cial aquelas que questionam a seriedade da pandemia, oferecem tratamentos não comprovados e que criticam medidas sanitárias recomendadas por autoridades.

Para além deste contexto, o Facebook passou a admitir a exclusão de uma ou outra categoria de conteúdo desinformativo, como sobre vacinas. Tanto no caso do Facebook e Instagram como no do Twitter, há restrições também para publicações que visam confundir eleitores sobre o exercício do voto. Essa constitui talvez a polêmica mais delicada entre as providências sobre desinformação nas plataformas, uma vez que demanda a fixação de uma linha tênue entre a garantia da liberdade de expressão e a coibição do seu abuso. Nesse contexto, a admissão de situações excepcionais de retirada em casos de risco evidente de danos graves parece uma possibilidade razoável, desde que conectada a regras de devido processo que permitam a contestação, a avaliação dos recursos por pessoas e a reparação em caso de erro na moderação aplicada, o que não é garantido em nenhuma das plataformas analisadas.

Tanto o Facebook e o Instagram quanto o Twitter adotam uma exceção para os discursos de líderes políticos. As plataformas entendem que, mesmo diante de mensagens que seriam vetadas a partir de suas regras gerais, a manutenção de alguns desses conteúdos pode se justificar pelo interesse público, inclusive para que seja possível sua crítica. O Twitter impôs limites, como a impossibilidade de interagir com a mensagem e avisos sobre verificações procedidas sobre o conteúdo. Mas diante da mensuração de possíveis danos, mesmo chefes de Estado tiveram conteúdos removidos. Foi o que ocorreu, por exemplo, diante da violação de normas relacionadas à pandemia do novo coronavírus. Postagens no Twitter e no Facebook dos presidentes dos Estados Unidos, Donald Trump, e do Brasil, Jair Bolsonaro, foram derrubadas por conterem afirmações contrárias às orientações de autoridades de saúde.

É fato que a desinformação não se propaga apenas pela dinâmica das plataformas, mas pela ação articulada de grupos

políticos e de representantes eleitos, inclusive presidentes da República, ministros, senadores e deputados. As plataformas, mais do que palanques, toram-se canais de difusão de figuras populares, amplificando essas estratégias de desagregação do debate público. Assim, o tratamento de tais discursos também se constitui em um problema de difícil equacionamento. Manter as publicações com a sinalização de seus problemas e impedindo sua amplificação, como faz o Twitter, pode ser um "caminho do meio" interessante, cuja efetividade ainda deve ser melhor analisada.

Se a remoção do conteúdo desinformativo não é regra em termos de sanção por parte das plataformas analisadas, a redução de alcance é uma providência mais presente. O Facebook diminui o alcance nos feeds tanto na rede social quanto no Instagram. Neste último caso, o conteúdo é retirado de determinadas seções, como a "Explorar". A prática fica, assim, em uma linha cinza entre uma precaução específica para conter a desinformação e uma medida correlata. A redução de alcance acaba assumindo quase uma condição de exclusão, dada a lógica de derrubada da visibilidade dos conteúdos nas plataformas. Assim, aparece como medida de alto impacto sobre os conteúdos, mas que não tem sua eficiência totalmente comprovada. Estudos mencionados na pesquisa apontam uma redução na velocidade na propagação de desinformações, mas a redução do alcance não impede que os conteúdos desinformativos sigam circulando. É necessário aprofundar investigações sobre o tema, sobretudo aquelas que possam efetivamente incidir sobre a operação de grupos profissionais que usam as plataformas com sistemas que permitem burlar essas limitações.

Nos últimos anos, diante das críticas sobre o papel de catalisador de discurso falso nas eleições, as plataformas passaram a implementar medidas de promoção do que chamam de integridade das eleições. Um exemplo são as iniciativas visando incentivar a participação nesses processos. Para além da remoção de conteúdos que possam enganar eleitores sobre os procedimentos do pleito,

citadas anteriormente, informações corretas sobre processos eleitorais passaram a ser disponibilizadas pelo Facebook, Instagram e Twitter.

Por fim, um aspecto central das medidas de moderação de conteúdo são as regras do chamado "devido processo", termo empregado para designar procedimentos democráticos que permitem a ciência dos usuários sobre a medida adotada pela plataforma, a possibilidade de recurso ou contestação e alternativas de reparação em caso de erros de remoção que tragam danos. Parte das plataformas afirmou sempre notificar seus usuários de que o conteúdo sinalizado violou determinada diretriz ou foi considerado falso por verificadores. Mas somente o YouTube, entre as redes sociais pesquisadas, disponibiliza publicamente informações detalhadas sobre os procedimentos de sanção de seus usuários diante da violação de seus termos de uso. No Facebook, é possível contestar sanções de moderação, mas no caso da verificação de conteúdo desinformativo o usuário deve buscar a própria agência, evidenciando a ausência de fluxos instituídos, de instâncias claras de processamento dos recursos e dos processos de revisão.

Já em relação à reparação, os procedimentos de veiculação de reversões de sanções a conteúdos desinformativos não são claramente estabelecidas nem possuem garantia de proporcionalidade. Os mecanismos de devido processo são fundamentais porque não são apenas garantias democráticas diante da enorme assimetria entre as plataformas e seus usuários. Considerando as complexidades de lidar com uma prática cujas manifestações não são exatas (e podem ocorrer não como "verdade" ou "mentira", mas muitas vezes em expressões intermediárias), a possibilidade de erros é constante, necessitando para isso de estruturas e processos robustos de devido processo como forma de corrigir erros e abusos.

3.6.3 PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Parte das plataformas analisadas disponibiliza mais informações no caso de conteúdos verificados. Uma modalidade de ação desse tipo é a exibição de artigos das próprias agências de checagem com as informações apuradas sobre determinado assunto. Outra é a oferta de informações "de contexto", como chamam o Facebook e o YouTube, ou de "fontes confiáveis", como denomina o Twitter. Sobre determinados temas, como vacinas, as explicações indicadas são oriundas de autoridades de saúde.

Conteúdos de veículos jornalísticos e oficiais podem também ser fornecidos. O WhatsApp, ao não fazer verificação de conteúdo, não apresenta informações de contexto. Mas incluiu uma "lupa" para facilitar a pesquisa sobre o tema. Contudo, na realidade brasileira de oferta de "serviços grátis" (aplicativos cujo uso não gera consumo de dados da franquia), tal ferramenta pode ser inacessível para muitos usuários, que não dispõem de dados para acessar outras fontes, o que reforça a necessidade de garantia de acesso universal à Internet no país como parte do combate à desinformação.

No âmbito desse esforço, uma estratégia das plataformas foi trazer mais informações sobre a fonte das publicações. No Facebook, ao clicar em um botão (i), o usuário pode ter dados sobre a página e seu administrador. No caso de anúncios, o Facebook e o Instagram também passaram a disponibilizar informações adicionais sobre as páginas de candidatos e propaganda eleitoral, tal como a identificação dos financiadores daquela mensagem paga e a "biblioteca de anúncios". Mais transparência sobre os anúncios ou conteúdos pagos no geral, que devem ser identificados de forma explícita, é fundamental.

No já mencionado contexto da pandemia, as plataformas implantaram diversos mecanismos de acesso e recomendação de informações oficiais sobre o tema, especialmente da Organização Mundial da Saúde (OMS) e de autoridades nacionais de saúde.

O Facebook publicou nos feeds dos usuários orientações da OMS sobre a pandemia. O Twitter reforçou sua ferramenta #ConheçaosFatos e criou uma aba dentro da seção "Explorar".

O WhatsApp criou o Coronavírus Information Hub em parceria com OMS, UNICEF, PNUD e IFCN para fornecer orientações e informações de fontes oficiais, além de aportar recursos para expandir os verificadores de fato no aplicativo. A rede social também criou um serviço por meio do qual o usuário pode receber informações de fontes oficiais sobre o coronavírus, bem como um bot para tirar dúvidas sobre a doença. Não obstante, o acesso a esses mecanismos depende, em geral, da busca ativa pelos usuários, que não têm sido sequer notificados diretamente para que tomem conhecimento dessas ferramentas, o que poderia ser feito sem causar prejuízo algum.

Algumas iniciativas já haviam sido promovidas anteriormente na área da saúde no tema de vacinas, diante do crescimento de movimentos contra o tratamento. O Twitter incluiu um aviso quando um usuário realizava busca sobre o assunto, direcionando para sites de autoridades de saúde.

Porém, do ponto de vista da informação sobre como lida com conteúdos desinformativos, chamou atenção a baixa transparência das plataformas. Para além de não disporem de políticas estruturadas de combate à desinformação, as medidas existentes não são apresentadas de forma organizada – à exceção do YouTube, que lançou uma página sobre desinformação na plataforma em outubro de 2020. Boa parte do trabalho da presente pesquisa foi, portanto, exatamente a busca e a organização de informações dispersas em notícias nos sites oficiais e em sites de "ajuda" ou "FAQ" (*Frequently Asked Questions*) de cada uma das empresas. Se a compreensão das estratégias e do conjunto de iniciativas já se mostrou desafiadora para a equipe de pesquisa, imagine para um usuário não dedicado a esse labor. Comparando dois temas, enquanto as plataformas desenvolvem políticas claras e detalhadas

sobre como usam dados pessoais dos usuários e sobre como acessar configurações de privacidade, por exemplo, as diretrizes sobre desinformação estão muito atrasadas nesse quesito. A ausência de reconhecimento da gravidade do problema e da implicação das plataformas acaba significando empecilhos para que a comunicação com os usuários sobre o tema seja efetiva.

Além disso, talvez o elemento mais problemático seja a ausência de balanço das ações propagandeadas. Mais uma vez, as informações são escassas e dispersas. À exceção da Europa, onde o modelo de corregulação demandou a apresentação de relatórios periódicos, em âmbito internacional e no Brasil, em particular, não foi possível identificar qualquer tipo de documento de avaliação deste tipo que indicasse a eficácia ou não das providências adotadas. Nos relatórios e centros de informação de transparência, como ocorre no caso do Facebook e do Twitter, o tema da desinformação não foi incluído. Dessa forma, a aferição da concretude das ações fica restrita às alegações das empresas, aproximando-se mais de promessas não comprováveis.

Por fim, as plataformas vêm, há alguns anos, sinalizando investimentos em iniciativas de educação midiática, campo fundamental para o enfrentamento mais sistemático ao fenômeno das chamadas "fake news". Tais iniciativas podem ser realizadas diretamente pelas empresas, com informações elaboradas pelas próprias plataformas, apresentadas aos usuários ou veiculadas em seus sites, ou feitas por meio de parcerias ou financiamento de projetos de outras organizações.

Na primeira estratégia estão Twitter, WhatsApp e Facebook. O primeiro disponibilizou o manual "Ensinando e aprendendo com o Twitter". O segundo traz em seu FAQ um texto intitulado "Dicas para ajudar a evitar a disseminação de boatos e notícias falsas". Também fez ações de divulgação nas ruas na Índia em 2018 e veiculação de anúncios no Brasil no mesmo ano. O terceiro disponibilizou conteúdos com dicas nos feeds de parte dos usuários em

algumas ocasiões pontuais. Na segunda estão Facebook, Twitter e YouTube. O primeiro financiou no Brasil o projeto "Vaza Falsiane". O segundo apoia iniciativas sobre o tema no Brasil em parceria com o Instituto Palavra Aberta, concede bônus para anúncios de organizações da sociedade civil que trabalham com educação midiática e promove hashtags durante a Semana Global de Educação para Mídia. O terceiro patrocina o EducaMídia, programa de educação midiática em parceria com o TSE. Não obstante, o alcance dessas medidas é questionável, sobretudo se for considerado o público alcançado por elas e o número de usuários que cada plataforma possui.

A realização de cursos e treinamentos é também uma prática, como em atividades promovidas por equipes do Facebook e do Twitter. O WhatsApp informou ter capacitado 40 mil pessoas na América Latina em 2020 em como "permanecer seguros na plataforma". O Google e o Facebook mantêm áreas e projetos específicos (como o Google News Initiative e o Facebook Journalism Project). O Twitter aportou recursos no Comitê de Proteção aos Jornalistas (CPJ) e na International Women´s Media Foundation para financiar ações dessas organizações. Algumas empresas abriram acessos para grupos de pesquisa. O Twitter firmou parcerias para análises por investigadores sobre influência de países externos em processos eleitorais. O WhatsApp lançou um prêmio para pesquisas relacionadas à desinformação, mas não está claro até que ponto os resultados são internalizados pelas companhias.

3.6.4 MEDIDAS CORRELATAS

O combate a contas falsas ou inautênticas apareceu como uma das principais medidas correlatas adotadas pelas plataformas. O Facebook o faz utilizando o conceito de comportamento inautêntico e comportamento inautêntico coordenado, que é passível de remoção.

O Instagram passou a implantar mecanismos de confirmação de identificação caso fossem verificados padrões de possíveis comportamentos inautênticos, o que resulta na redução da circulação de usuários que não forneçam a identificação. O Twitter também remove contas falsas e ampliou as exigências para abertura de contas, como e-mail e número de telefone. O WhatsApp passou a utilizar ferramentas baseadas em inteligência artificial para mapear comportamentos vetados e banir contas. O YouTube remove contas abertas com o intuito de se passar por outro canal ou outra pessoa, de acordo com a política de falsificação de identidade. A falsificação é apontada, entre outros, quando o canal copia o perfil ou a aparência geral de outro ou quando um canal falsifica outro canal de notícias existente, entre outras ações de falsificação.

As companhias divulgam em seus sites institucionais (como exemplificado no capítulo 2) ações de remoção, que em determinadas situações atingem redes de atores nos seus interiores. Outra providência relacionada à coibição de perfis falsos ou enganosos é a verificação de contas, que visa dar uma "chancela" de credibilidade a esses perfis. Uma vez mais, falta, contudo, transparência nesse processo e canais para cobrar ou questionar a certificação.

Outra modalidade de intervenção diz respeito não somente à conta mas a atitudes que "artificializam" o debate público. O Twitter analisa contas de comportamento malicioso. O Instagram ajuizou processo contra uma empresa na Nova Zelândia que comercializava likes, views e seguidores falsos. Dentro da política de envolvimento falso, o YouTube afirma não permitir nenhuma ação que aumente de maneira artificial o número de visualizações.

O combate ao discurso de ódio e conteúdo extremo também apareceu no Facebook, Instagram, YouTube e Twitter. Esse tipo de publicação é proibida pelas regras internas das plataformas. Também há vedações a ameaças e outras incitações à violência contra pessoas ou grupos. Uma prática relacionada, a de assédio, é abordada especialmente no Instagram, YouTube e Twitter. Diante

de questionamentos sobre os impactos do Instagram no bem estar de jovens, a plataforma começou a testar em 2019 a não exibição do número de likes de um determinado post (argumentando que isso diminuiria a pressão e a competição entre usuários por reconhecimento).

A restrição do spam, presente em todas as plataformas, aparece como uma das principais estratégias correlatas do WhatsApp. As medidas da plataforma são outro exemplo de como as alterações se dão como resposta aos problemas e questionamentos dos impactos dessas redes sobre a sociedade. O WhatsApp impôs limitações de reencaminhamento em 2018, em meio a críticas do uso do aplicativo para a difusão de desinformação em países, especialmente na Índia. Neste país, foram registrados episódios em que a difusão de mensagens falsas estava ocasionando linchamentos públicos.

Concluímos que as ações ainda carecem de organização, robustez, transparência e avaliação. O problema da desinformação precisa ser efetivamente reconhecido, comunicado e enfrentado pelas plataformas, o que passa pela revisão da estrutura e do modelo de negócios dessas empresas, sob pena de seguirem oferecendo remédios incapazes de interromper uma forma de comunicação que hoje se converteu em um dos principais desafios às democracias em todo o mundo.

Tabela: Síntese da pesquisa

FACEBOOK

ABORDAGEM DO FENÔMENO	Não possui uma política específica nem trabalha com uma definição própria de desinformação. Apresenta as estratégias adotadas contra o fenômeno de maneira resumida nos Padrões da Comunidade. A elaboração de medidas a serem adotadas sobre o tema é realizada pela equipe de "Políticas Globais de Conteúdo". Em 2020, criou o Conselho de Supervisão de Conteúdo, com membros externos à empresa, responsável pela revisão aprofundada de casos (não apenas sobre desinformação).
MODERAÇÃO DE CONTEÚDO	Recebe denúncias e utiliza sistemas automatizados para analisar as publicações. Encaminha conteúdos para verificação por agências de checagem parceiras e, em caso de classificação como "falso" ou "parcialmente falso", o conteúdo recebe um rótulo específico e passa a ser acompanhado de "artigos relacionados" produzidos por checadores. Declarações e postagens de líderes políticos são excepcionadas e não passam por processos de verificação de agências de checagem parceiras. Vídeos e imagens podem ser rotulados como "manipulados", "tirados de contexto" ou "falsos". Não há remoção da plataforma, mas a circulação de postagens falsas é reduzida nos feeds e nos mecanismos de recomendação. Casos de desinformação que causem violência ou dano, que tragam informação falsa sobre vacinas ou que comprometa processos eleitorais são removidos. Anúncios com mensagens falsas não podem ser veiculados e anunciantes reincidentes podem ser sancionados.
PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA	Oferece ao usuário um ícone para buscar mais informações sobre a fonte da postagem. No contexto da Covid-19, apresenta informações de autoridades de saúde nos resultados da busca (direcionando usuários para o site da OMS) e lançou o Centro de Informações sobre Coronavírus, disponível no alto do News Feed. Pessoas que interagiram com conteúdos "falsos nocivos" recebem mensagens. Financia iniciativas de educação midiática, desenvolveu materiais e campanhas sobre o tema e tem parceria com agências de checagem para disponibilizar dicas sobre verificação de conteúdos. Apoia experiências jornalísticas visando ampliar a circulação de notícias profissionais na plataforma. Disponibiliza dados para 60 pesquisadores em todo o mundo.
MEDIDAS CORRELATAS	Adotou política de proibição do discurso de ódio em 2018 e melhorou sistemas automatizados para identificar vídeos violentos, passando a restringir determinadas transmissões ao vivo. Buscas relacionadas à supremacia branca, como exaltação ao nazismo, recebem indicações de materiais de combate ao ódio. Em 2020 passou a remover conteúdos de negação do holocausto. No mesmo ano também foi banido o grupo QAnon, que difunde teorias da conspiração. A plataforma ainda derruba contas falsas, propagadoras de spam ou consideradas de "comportamento não autêntico coordenado". Procura verificar a autenticidade das contas no momento de sua abertura e mapeia sinais de condutas maliciosas em contas recém-abertas. Também reduz a circulação de contas difusoras de conteúdo de baixa qualidade (caça-cliques ou que levam para sites com anúncios maliciosos e chocantes). Passou a restringir o acesso à coleta de dados por apps de terceiros. Como medida de proteção à integridade eleitoral, disponibiliza um arquivo de anúncios políticos com informações sobre identidade dos anunciantes, pagamento e alcance dos conteúdos, além de rotular essas peças como "propaganda eleitoral". Usuários podem dizer se querem ou não receber esse tipo de anúncio.

INSTAGRAM

ABORDAGEM DO FENÔMENO

Segue a linha geral de abordagem do fenômeno adotada pelo Facebook, seu controlador. Os Termos de Uso determinam que o usuário não pode "fazer algo ilícito, enganoso, fraudulento ou com finalidade ilegal ou não autorizada". Não possui estrutura institucional ou processos específicos para tratar de desinformação. Valem os fluxos e procedimentos adotados pelo Facebook, incluindo as equipes de revisão de conteúdo e o Conselho de Supervisão.

MODERAÇÃO DE CONTEÚDO

Realiza processos de análise de conteúdos a partir de agências de verificação de fatos, que classificam uma publicação em diferentes categorias. Conteúdos são marcados quando são verificados como desinformativos e podem ter sua circulação reduzida (menor visualização no feed e nos stories, remoção da página "explorar" e das páginas de hashtags). Conteúdos relacionados a falsos tratamentos sobre a Covid-19, sobre teorias da conspiração ou falsas alegações registradas como danosas pelas autoridades de saúde são removidos, assim como anúncios e hashtags que possam promover desinformação sobre o tema. Na busca sobre o vírus, prioriza informações de autoridades de saúde. Durante as eleições, se um conteúdo for classificado como falso, um filtro cinza é exibido sobre a imagem e alertas são dados ao usuário antes de compartilhar uma publicação. Ferramentas de transparência para anúncios vêm sendo adotadas em diferentes países.

PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Sistemas informatizados buscam "correspondência de imagens" para rotular conteúdos já classificados como falsos, inclusive quando eles foram inicialmente publicados no Facebook (e vice-versa). Disponibiliza textos de contraponto elaborados por verificadores de fatos, com "desmentidos" ou informações oficiais sobre o tema. Buscas sobre vacinas direcionam usuários para o site da OMS. No contexto da Covid-19, fez o mesmo e passou a disponibilizar informações do órgão no topo do feed dos usuários, além de oferecer imagens sobre medidas de prevenção para compartilhamento no stories.

MEDIDAS CORRELATAS

As Diretrizes da Comunidade condenam o discurso de ódio, mensagens indesejadas enviadas repetidamente e ataques a pessoas com intenção de constrangê-las. Contas falsas são removidas, a maior parte no momento da tentativa de sua abertura. Contas cuja identidade do responsável é verificada podem ganhar sinalização de verificação. Contas ou redes de contas enquadradas na categoria de "comportamento inautêntico" são derrubadas. Contas que registrem determinados números de violações em um período podem ser desativadas, com direito à apelação. A promoção, oferta ou comércio de avaliações falsas de usuários são proibidos; e curtidas, seguidores e comentários não autênticos para "impulsionar a popularidade" são removidos. Tem ampliado medidas contra **bullying** e assédio, dando maior controle sobre interações dos usuários e deixando de exibir o número de likes nos posts.

WHATSAPP

ABORDAGEM DO FENÔMENO

Apresenta-se como plataforma de mensagens privadas criptografadas que, como tal, não acessa, modera, julga, verifica, bloqueia ou retira qualquer conteúdo, o que é contraditório com a existência de grupos e listas de transmissão que acabam fazendo com que tenha funcionamento de rede social. Não assume o uso frequente da plataforma para desinformação. Seguindo tal entendimento, não possui política, processos e/ou estrutura institucional nem conceito de desinformação definidos.

MODERAÇÃO DE CONTEÚDO

Informa oficialmente que não modera conteúdo. Não marca conteúdos como desinformativos, por exemplo. No campo da moderação, apontamos apenas que, desde 2018, tem atuado para reduzir a circulação, por meio do estabelecimento de limites para encaminhamento de mensagens. Primeiro, o limite de envio de uma vez foi fixado em 20 chats. Depois, 5. Desde a pandemia, as mensagens altamente encaminhadas só podem ser encaminhadas para um contato de cada vez. Mesmo essa política não é, em geral, associada ao combate à desinformação. Tal vinculação só foi expressa no contexto da pandemia.

PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

No aplicativo ou no blog oficial, é possível encontrar referências à desinformação, mas as informações são dispersas e pouco acessíveis a um usuário comum, embora a plataforma tenha possibilidade de contato direto com ele, por meio de notificações. A partir das eleições de 2018, foram feitas parcerias institucionais para verificação de conteúdos e realização de pesquisas, mas o alcance e a efetividade delas são questionáveis. Da mesma forma, as ações de sensibilização do público amplo têm sido restritas. As medidas de promoção de informações foram ampliadas na pandemia, com a criação de canais de mensagens. No mesmo sentido, para as eleições de 2020, foi firmado acordo com TSE e criado chatbot para denúncias e figurinhas sobre voto consciente. O usuário tem que adicionar o número da conta e entrar em contato. Falta transparência quanto às práticas adotadas pelo WhatsApp, como proibição de criação ou remoção de contas, além de relatórios sobre suas medidas.

MEDIDAS CORRELATAS

A maior parte das ações desenvolvidas pelo WhatsApp inscrevem-se em "medidas correlatas", a começar pela definição de usos envolvendo declarações falsas, incorretas ou enganosas como violações em seus Termos de Serviço. Além da já mencionada limitação de encaminhamentos de mensagens, tem sinalizado mensagens altamente encaminhadas com uma ou duas setas, indicando maior viralização. Foi retirado também o botão que permitia encaminhamento rápido de conteúdos e inserida uma lupa ao lado das mensagens assinaladas com duas setas. Utiliza ferramentas de tratamento de spam e aprendizagem avançada de máquinas para retirar mensagens automatizadas em massa e banir contas de usuários com comportamentos inadequados, a exemplo do envio de mensagens em massa e da criação de múltiplas contas.

YOUTUBE

ABORDAGEM DO FENÔMENO

Não trabalha com uma única definição de desinformação e não possui uma política específica para o tema. Mas restrições à disseminação de desinformação aparecem em diferentes políticas, como sobre danos e mídias manipuladas. Conteúdos que violam as Diretrizes da Comunidade são removidos e notificados. Casos reincidentes sofrem restrições de publicação e monetização e o canal pode ser excluído. Estimula que a comunidade denuncie conteúdos inapropriados e equipes internas analisam os casos. Lançou recentemente a página "Mitos e Fatos sobre Desinformação", que agrega medidas que adota contra o problema.

MODERAÇÃO DE CONTEÚDO

Tira do ar conteúdos que violam suas diretrizes e reduz o alcance de "desinformação danosa" e "conteúdos limítrofes", deixando de recomendá-los. Vídeos editados e adulterados são proibidos pela política de "mídia manipulada", assim como canais que falseiam sua identidade. Remove conteúdos e canais que desrespeitam a política para eleições, incluindo vídeos que visam minar a confiança nos pleitos. Passou a remover informações enganosas sobre a Covid-19, sobretudo desinformação médica, promoção de medicamentos perigosos e informações sobre a origem do vírus. Adotou forte política contra vídeos antivacina. Recomenda aos produtores checarem fatos e usarem autoridades de saúde como fonte. Nem toda a publicidade em torno da pandemia foi restrita e vídeos que seguem os padrões da comunidade poder receber anúncios. Decisões de cumprimento às diretrizes são tomadas a partir da análise do conteúdo e independem de quem fala ou publica tais materiais, incluindo altas autoridades. Possui um Grupo de Análise de Ameaças para identificar práticas de desinformação patrocinadas por governos.

PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Informa priorizar o que chama de "vozes autorizadas", como canais de jornalismo, nas buscas por informações sobre eventos e política e nos "próximos vídeos" sobre esses temas. Se o usuário pesquisa conteúdo jornalístico, mostra uma estante com "Principais notícias". A estante "Últimas notícias" é exibida para eventos importantes. Exibe painéis de informação com dados de contexto de fontes autorizadas sobre assuntos históricos, cientificamente comprovados e teorias da conspiração. A depender da busca, checagens de fatos por editores independentes também são exibidas. Não há checagem sobre cada vídeo individual. A Covid-19 ganhou sessões especiais em todas essas ferramentas. Nas eleições do EUA, oferece dados adicionais sobre candidatos. No Brasil, fez lives sobre desinformação com o TSE. Via Google News Initiative, apoia o jornalismo digital no combate à desinformação e ações de educação para a mídia.

MEDIDAS CORRELATAS

Possui uma política própria contra discurso de ódio e outra contra assédio e ameaças. Vídeos supremacistas são proibidos e aqueles que não podem ser caracterizados como de ódio mas se aproximam disso têm restrições em engajamento, recomendação e monetização. Um alerta é exibido antes de seu início. Em 2020, passou a remover conteúdos relacionados a teorias da conspiração que resultem em atos de violência. Qualquer pessoa ou canal intimidando, perseguindo, desumanizando e incentivando comportamento violento, inclusive via comentários, pode ser banido da rede. Possui ainda uma política para tratar de temas como privacidade e contra a difamação de pessoas, para recebimento de denúncias. Não autoriza a distribuição de conteúdos perigosos ou nocivos e tem uma política contra spam, falsificação de identidade e envolvimento/engajamento falso. Restringe o acesso ao serviço via qualquer meio automatizado (bots).

TWITTER

ABORDAGEM DO FENÔMENO

Não possui uma política para validar a autenticidade de conteúdos, tampouco trabalha com uma definição do termo desinformação. O foco é oferecer contexto para ajudar pessoas a decidir o teor de um post contestado, agindo diretamente somente em casos de possíveis danos causados por mídias manipuladas, informações enganosas sobre processos eleitorais e, recentemente, relacionadas à Covid-19. Proíbe o uso de robôs manipuladores da rede. Tem um Conselho de Confiança e Segurança que também analisa o tema.

MODERAÇÃO DE CONTEÚDO

Mídias manipuladas, incluindo **deepfakes**, são marcadas, podem ter sua visibilidade reduzida, receber link com mais explicações ou, em caso de risco de danos graves imediatos, removidas. Conteúdos que possam enganar as pessoas sobre quando, onde e como votar, também são proibidos, dentro da política de integridade nas eleições, que nos EUA sinaliza as postagens de candidatos. Informações falsas com intuito de minar confiança do público nas eleições podem ser rotuladas ou removidas. Anúncios políticos e de veículos de comunicação estatais não são mais permitidos. Monitoram tendências e picos nas conversas sobre eleições para detectar atividades de manipulação e derrubam contas falsas. Postagens de líderes globais, mesmo que desinformativas, poder seguir no ar se forem de claro interesse público. O post é exibido por trás de um aviso e não é possível interagir com ele. Pós Covid-19, ampliou a definição de "dano" para incluir conteúdos contrários às recomendações das autoridades de saúde e remove posts com chamado claro a ações de risco. Sob risco de dano menos severo, o post recebe um rótulo ou aviso. A política de anúncio só é autorizada para serviços públicos e privados que mudaram com a Covid-19.

PROMOÇÃO DE INFORMAÇÕES E TRANSPARÊNCIA

Possui a ferramenta #ConheçaosFatos, que é exibida em pesquisas associadas a vacinas e Covid-19. Informações não-críveis sobre saúde não aparecem nas buscas, que priorizam fontes oficiais nas respostas. Adicionou aba especial sobre a pandemia na função #Explorar, com fontes de credibilidade, comunicados públicos e jornalistas em saúde. Aplica rótulo em posts sobre Covid e 5G, encorajando a checagem de fatos e, no recurso "Eventos", também seleciona informações de credibilidade e disponibiliza no topo das timelines. Aumentou o número de contas verificadas nesse sentido. Abriu a plataforma para acompanhamento em tempo real de postagens sobre a Covid por desenvolvedores e pesquisadores. Cria páginas especiais sobre as eleições com informações oficiais e de credibilidade. Apoia iniciativas de jornalismo independente, de checagem de fatos e de educação para a mídia, inclusive por meio de créditos em anúncios.

MEDIDAS CORRELATAS

Sua política contra propagação de ódio proíbe conteúdos com ameaças violentas, que promove ou expressa desejo de morte ou agressão, disseminação de estereótipos, medo e assédio. Comportamentos considerados abusivos também não são permitidos. Infratores podem ter que excluir conteúdo, passar um tempo sem publicar ou interagir ou ser suspensos totalmente. Combate ao spam e à automação mal-intencionada, monitorando e desafiando atividades suspeitas das contas, como por meio da verificação de identidades. O foco é barrar contas falsas, o engajamento não autêntico e atividades coordenadas para influenciar artificialmente conversas. Perfis suspeitos têm a visibilidade reduzida e auditorias são realizadas em contas existentes. Divulga arquivos de posts de atividades coordenadas apoiadas por governos para dar transparência a tentativas de manipulação da rede.

Tabela: Síntese das medidas adotadas pelas plataformas digitais (2018-2020)

	Facebook	Instagram	WhatsApp	YouTube	Twitter
Abordagem do fenômeno					
Política específica e estruturada de combate à desinformação					
Conceito de desinformação adotado e definido					
Menção à desinformação em diretrizes ou políticas			x	x	x
Medidas correlatas que incidem sobre desinformação	x	x	x	x	x
Estrutura institucional específica para o tema					
Moderação de conteúdo*					
Verificação de conteúdos desinformativos por equipe própria				x	x
Verificação de conteúdos desinformativos por terceiros	x	x	x		x
Remoção de conteúdos desinformativos				x	x
Redução de alcance	x	x	x	x	x
Desmonetização e medidas restritivas em anúncios e conteúdos impulsionados	x			x	x
Suspensão de contas por desinformação				x	x
Recurso para denúncia de desinformação				x	x
Notificação de usuários em caso de sanção por desinformação	x	x		x	x
Recurso à plataforma em caso de sanção por desinformação				x	x

Promoção de informações e transparência					
Sinalização de conteúdos desinformativos	x	x			x
Disponibilização da informação sobre verificação do conteúdo	x	x	x		x
Disponibilização de conteúdo de contexto ou fontes oficiais	x	x	x	x	x
Medidas de transparência de anúncios	x	x		x	x
Canais de informação sobre medidas contra desinformação				x	
Balanço das medidas adotadas contra desinformação nos relatórios de transparência					
Iniciativas de educação midiática	x	x	x	x	x
Projetos de apoio ao jornalismo profissional	x	x	x	x	x
Medidas correlatas					
Proibição de contas inautênticas	x	x		x	x
Restrições a contas e ferramentas automatizadas			x	x	x
Vedação de veiculação de discurso de ódio	x	x		x	x
Restrição de mídias manipuladas	x	x		x	x
Restrição à circulação de mensagens			x		

(*) No YouTube e no Twitter, a moderação ocorre quando a plataforma enquadra determinadas desinformações em sua política contra conteúdos que geram dano ou que afetem eleições.

BIBLIOGRAFIA

ABÍLIO, Ludmila Costhek. Uberização do trabalho: subsunção real da viração. *Passa Palavras*, Campinas, 2017.

ALLCOTT, H.; GENTZKOW, M. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, v. 31, n. 2, 2017, p. 211-236.

ALLEN, J. FLORES, N. *The role of government in the Internet. Final report for the Dutch Ministry of Economic Affairs*. Analysys Mason, 2013. Disponível em:<http://www.analysysmason.com/About-Us/Case-Study-Content/Case-study-50-Government-role-Internet-case-study/Government-role-Internet-case-study/>. Acesso em: 24ago. 2017.

ANDERSSON SCHWARCZ, J. *Platform Logic:* An Interdisciplinary Approach to the Platform-Based Economy. Policy; Internet, 9(4), (2017). P. 374-394.

BARDIN, Laurence. *Análise de conteúdo*. Lisboa: Edições 70, 1977.

BARLOW, John Perry. *A declaration of the independence of cyberspace*. 1996. Disponível em: <https://projects.eff.org/~barlow/Declaration-Final.html>. Acesso: 31 mar. 2020.

BOLAÑO, C. R. S.; VIEIRA, E. Economia política da internet e os sites de redes sociais. *Revista Eptic, Aracaju*, v.16, n. 2, p. 75-88, mai./ago, 2014. Disponível em <http://www.eptic.com.br/EPTIC%20I-1.pdf>. Acesso: 10 ago. 2020.

CAMMAERTS, Bart; MANSELL, Robin. Digital Platform Policy and Regulation: Toward a Radical Democratic Turn. *International Journal of Communication*, v. 14, p. 20, 2020.

DANTAS, Marcos. *Internet: praças de mercado sob controle do capital financeiro*. In: XL Congresso Brasileiro de Ciências da Comunicação– INTERCOM, 2017.

DE COCK BUNNING, M. et al. A multi-dimensional approach to disinfor-mation: report of theindependent high level group on fake news and on-line disinformation. *European Comission*. Luxemburgo: Publications Office of European Comission, 2018.

EGAN, Erin. BERINGER, Ashlie. It's Time to Make Our Privacy Tools Easier to Find. *Facebook blog*. Publicada em: 28 mar. 2018. Disponível em: https://about.fb.com/news/2018/03/privacy-shortcuts/.

EVANS, D. S.; SCHMALENSEE, R. Matchmakers: The new economics of mul-tisided platforms. Harvard Business Review Press, 2016.

GAWER, A. Bridging differing perspectives on technological platforms: Toward anintegrative framework. *Research Policy*, [s.l.], v. 43, n. 7, p.1239-1249, set. 2014.

GILLESPIE, Tarleton. The politics of 'platforms'. *New Media & Society*, [s.l.], v. 12, n. 3, p.347-364, 2010.

_____, Tarleton. Governance of and by platforms. *SAGE handbook of social media*. 2017. P. 254-278.

GUNTHER, R.; BECK, P. A.; NISBET, E. C. *Fake news may have contributed to Trump's 2016 victory*. Manuscrito não publicado, 2018. Disponível em: <https://www.documentcloud.org/documents/4429952-Fake-News-May-Have-Contributedto-Trump-s-2016.html>. Acesso em 20 jan. 2019.

HAFIDZAH, N. Anti-fake News Act 2018: A comparative impact of the Act,deterrence or rehabilitation approach? *Malaymail*. Publicada em 20 fev. 2019. Disponível em: https://www.malaymail.com/news/what-you-think/2019/02/20/anti-fake-news-act-2018-a-comparative--impact-of-the-act-deterrence-or-rehab/1724791 Acesso em: 13 setembro de 2019.

HELBERGER, N. PIERSON, J. POELL, T. Governing online platforms: From contested to cooperative responsibility. *The information society*, 34(1), 1-14. 2018.

HERSCOVIVI, A. Deus e o Diabo na terra das plataformas digitais: uma análise a partir da hipótese braudeliena. *Revista Eptic*, v. 21, n. 1, p. 125-141, 2019.

JIN, D. Y. *Digital platforms, imperialism and political culture*. Abigdon: Routledge, Taylor & Francis Group, 2015.

KELLER, Clara Iglesias. Entre exceção e harmonização: o debate teórico sobre a regulação da Internet. *Revista Publicum*. Rio de Janeiro, v. 5, n. 1, p. 137-166, 2019. http://www.e-publicacoes.uerj.br/index.php/publicum.

KHAN, Lina. Sources of Tech Platform Power. *Georgetown Law Technology Review*, 325, 2018.

LESSIG, L. *Code V 2.0*. New York: Basic Books, 2006.

MacKINNON, R.; HICKOK, E.; BAR, A.; Lim, H. Fostering freedom online: The Roles, challenges and obstacles of internet intermediaries. *United Nations Educational*, 2014.

MAK, A. How Russia used Instagram to influence the 2016 presidential election. Slate. Nov, 2017. Disponível em: https://slate.com/technology/2017/11/how-russia-used-instagram-to-influence-the-2016-presidential-election.html. Acesso: 14 dez. 2020.

MARINONI, B.; GALASSI, V. Aspectos da desinformação, capitalismo e crise. In: MARTINS, Helena, Org. *Desinformação:* crise política e saídas democráticas para as fake news. São Paulo: Veneta, 2020, p. 27-62.

MARTINS, Helena. Muito além das fake news: aspectos da desinformação em meio à crise social. In: MARTINS, Helena, Org. *Desinformação:* crise política e saídas democráticas para as fake news. São Paulo: Veneta, 2020, p. 12-26.

MENA, Paul. Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet*, 12, no. 2, 2020.

NELSON, J. Is _fake news‘a fake problem? Columbia Journalism Review, 31 jan. 2017. Disponível em: https://www.cjr.org/analysis/fake-news-facebook-audience-drudge-breitbartstudy.php. Acesso em: 23 fev. 2019.

NOVAK, M. Saudi Arabia Threatens Anyone Spreading 'Fake News' Online with5 Years in Prison, Heavy Fines. *The Verge*. Publicada em: 15 out. 2018. Disponível em https://gizmodo.com/saudi-arabia-threatens-anyone-spreading-fake-news-onlin-1829749930. Acesso em: 15 de setembro de 2020.

PENNYCOOK, Gordon. BEAR, Adam. COLLINS, Evan. RAND, David G. *The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings*. Management Science, 2017.

PHILIPE MELO et al. *Can WhatsApp Counter Misinformation by Limiting Message Forwarding?* In: International Conference on Complex Networks and Their Applications (p. 372-384). Disponível em: <https://arxiv.org/abs/1909.08740 >. Acesso: 10 dez. 2020.

RUSSEL, J. Singapoor passes controversial 'fake news' law which critics fera willstifle free speech. *Techcrunch*. Publicdaa em 9 mai. 2019. Disponível em https://techcrunch.com/2019/05/09/singapore-fake-news-law/. Acesso em: 12 setembro 2020.

SANTOS et al. WhatsApp, política mobile e desinformação: a hidra nas eleições presidenciais de 2018. *C&S*. São Bernardo do Campo, v. 41, n. 2, p. 307-334, maio-ago. 2019. Disponível em: <https://www.metodista.br/revistas/revistas-ims/index.php/CSO/article/view/9410 >. Acesso: 10 ago. 2020.

SELTZER, E. K. HORST-MARTZ, E. LU, M. MERCHANT, R. M. Public sentiment and discourse about Zika virus on Instagram. *Public Health*, 150, 2017.

THOMPSON, J. B. *Ideologia e cultura moderna:* teoria social crítica na era dos meios de comunicação de massa. 7a ed. Petrópolis, RJ: Vozes, 2007.

VALENTE, JCL. **Tecnologia, informação e poder**: das plataformas online aos monopólios digitais. 2019. 400 f., il. Tese (Doutorado em Sociologia) - Universidade de Brasília, Brasília, 2019.

_____, JCL; PITA, Marina. Monopólios digitais: concentração e diversidade na Internet. Intervozes. 2018. Disponível em: https://intervozes.org.br/arquivos/interliv012monodig.pdf

VRAGA, Emily K. SOJUNG, Claire Kim. COOK, John. BODE, Leticia. Testing the Effectiveness of Correction Placement and Type on Instagram. *The International Journal of Press/Politics*, 2020.

WU, Tim. Network Neutrality, Broadband Discrimination. *Journal of Telecommunications and High Technology Law*, Vol. 2, p. 141, 2003.

YIN, Robert K. *Estudo de caso:* planejamento de métodos. 4. ed. Porto Alegre: Bookman, 2010.

DOCUMENTOS INTERNACIONAIS

COMISSÃO EUROPEIA. *Combater a desinformação em linha:* uma estratégia europeia. 2018. Disponível em: https://eur-lex.europa.eu/legal-content/PT/TXT/?uri=CELEX%3A52018DC0236.

CNDH. *Recomendação Nº 4 de 2018*. Disponível em: https://www.in.gov.br/materia/-/asset_publisher/Kujrw0TZC2Mb/content/id/27129495/do1-2018-06-25-recomendacao-n-4-de-11-de-junho-de-2018-27129463.

EUROPEAN UNION. *A multi-dimensional approach to disinformation*. Report of the independent High level Group on fake news and online disinformation. Disponível em: https://ec.europa.eu/digital-sin-

gle-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

ONU. *Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión*. Disponível em: https://www.refworld.org.es/pdfid/5c6b33774.pdf

ONU, OEA, OSCE e CADHP. *Declaración conjunta sobre libertad de expresión y "noticias falsas" ("fake news"), desinformación y propaganda*. 2017. Disponível em: http://www.oas.org/es/cidh/expresion/showarticle.asp?artID=1056&lID=2

ONU, OSCE e OEA. *Declaración conjunta sobre libertad de expresión y elecciones en la Era digital*. 2020. Disponível em: http://www.oas.org/es/cidh/expresion/showarticle.asp?artID=1174&lID=2

OUTROS DOCUMENTOS

APP ANNIE. State of mobile 2020. App Annie. 2020.

WE ARE SOCIAL. Digital em 2020. We Are Social. Disponível em: https://wearesocial.com/digital-2020. Acesso em: 23 setembro 2020.

FACEBOOK. Facebook report on the implementation of the Code of Practice for Disinformation. Annual report, september, 2019.

FACEBOOK. Entrevista aos pesquisadores. 2020

PEW RESEARCH CENTER. Social media update 2016. Pew research center. 2016. Disponível em: https://www.pewresearch.org/internet/2016/11/11/social-media-update-2016/.

DELOITTE. Global Mobile Consumer Survey Brasil 2019. Deloitte. 2019.

AMNESTY INTERNATIONAL. Toxic Twitter - A Toxic Place For Women. 2018. Disponível em: <https://www.amnesty.org/en/latest/research/2018/03/online-violence-against-women-chapter-1/>. Acesso: 14 dez. 2020.

ATKIN, Emily. Fact-check of viral climate misinformation quietly removed from Facebook. *Heated*. Publicada em 20 jul. 2020. Disponível em: https://heated.world/p/fact-check-of-viral-climate-misinformation.

FRENKEL, Sheera. Et. Al. Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis. *The News York Times*. Publicada em: 14 nov. 2018. Disponível em: https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html.

INTERVOZES. *Caminhos para a luta pelo direito à comunicação*. 2015. Disponível em: <http://intervozes.org.br/arquivos/interman-004cldcnb.pdf>. Acesso: 14 dez. 2020.

MACHADO, Caio et. al. *Ciência contaminada:* Analisando o contágio de desinformação sobre coronavírus via YouTube. Relatório 1 de estudo do Centro de Análise da Liberdade e do Autoritarismo (LAUT), Instituto Nacional de Ciência e Tecnologia em Democracia Digital (INCT.DD) e Centro de Estudos e Pesquisas de Direito Sanitário (Cepedisa), maio 2020, Disponível em: https://laut.org.br/ciencia-contaminada.pdf?utm_source=twitter&utm_medium=social&utm_campaign=cincia_contaminada. Acesso em 22/05/2020.

OBSERVACOM et al. *Contribuições para uma regulação democrática das grandes plataformas que garanta a liberdade de expressão na internet*. Disponível em: <https://www.observacom.org/wp-content/

uploads/2019/08/Contribuic%cc%a7o%cc%83es-para-uma-regulac%cc%a7a%cc%83o-democra%cc%81tica-das-grandes-plataformas-que-garanta-a-liberdade-de-expressa%cc%83o-na-internet.pdf>. Acesso: 30 mar. 2020.

REUTERS INSTITUTE. *Digital News Report 2020*. Disponível em: <https://www.digitalnewsreport.org/survey/>. Acesso em: 14 dez. 2020.

RUEDIGER, Marco Aurélio; GRASSI, Amaro. *Desinformação na era digital:* amplificações e panorama das Eleições 2018 Rio de Janeiro: FGV DAPP, 2018. Disponível em: http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/25742/Desinformação%20Policy-Paper-2%20Sala.pdf?sequence=1&isAllowed=y.

SCHERMANN, Daniela. *Panorama Mobile Time/Opinion Box:* Mensageria no Brasil. Opinion Box, 2 mar. 2018. Disponível em: https://blog.opinionbox.com/mensageria-no-brasil-sexta-edicao/. Acesso em: 11 dez. 2020.

STATISTA. *Social Media Usage in Latin America*. 2020. Disponível em: <https://www.statista.com/topics/6394/social-media-usage-in-latin-america/>. Acesso: 14 dez. 2020.

FORMATO **16×23cm**
MIOLO **Papel Offset 75g/m²**
CAPA **Papel Supremo 250g/m²**
TIPOGRAFIA **Sabon LT Std 11/15**
IMPRESSÃO **Gráfica Multifoco**