REDES SOCIAIS E MODERAÇÃO DE CONTEÚDO:

Criando regras para o debate público a partir da esfera privada.

AUTORES
Marina Estarque
João Victor Archegas

REVISÃO
Celina Bottino
Christian Perrone

ITS
Instituto de Tecnologia & Sociedade do Rio

SUMÁRIO

RESUMO EXECUTIVO

O presente relatório tem como objetivo apresentar um arco histórico da moderação de conteúdo em redes sociais, com foco no Facebook. Plataformas digitais são, via de regra, criadas e administradas por empresas privadas, muitas delas com sede na Califórnia. Nada obstante, grandes corporações como o Facebook e o Twitter criam regras que impõem limites ao exercício da liberdade de expressão na arena digital, atingindo, assim, o debate público e não apenas as interações privadas. Ademais, essas empresas ofertam seus serviços ao redor do mundo, o que significa que essas regras são aplicadas em centenas de países, englobando diferentes culturas, línguas e doutrinas de liberdade de expressão.

Com base nesse contexto, o ITS Rio apresenta um panorama geral da moderação de conteúdo, passando pela análise dos padrões da comunidade do Facebook, o regime jurídico por trás da regulação e autorregulação das plataformas, a tensão entre a dimensão global e local das redes sociais, a criação do Oversight Board e de diferentes diretrizes internacionais e, por fim, o redesenho do Facebook em direção à uma rede social voltada aos grupos e interações privadas. Esse panorama servirá de base para outras pesquisas na área, oferecendo, assim, um importante repositório de fontes e discussões iniciais sobre o tema.

Ao final do relatório, o leitor terá uma melhor compreensão das diferentes dimensões envolvidas no debate sobre moderação de conteúdo em redes sociais. O ITS Rio também oferece uma linha do tempo com os principais eventos desde 1996 com a criação da *Section* 230 como parte do *Communications Decency Act* nos EUA até o banimento de Donald J. Trump em janeiro de 2021. Um segundo relatório será publicado com os resultados de uma pesquisa que o ITS Rio vem desenvolvendo a respeito da moderação de conteúdo em grupos do Facebook no Brasil.

LINHA DO TEMPO

FEVEREIRO
1996

A *Section* 230 é aprovada pelo Congresso dos EUA e passa a compor o *Communications Decency Act* (CDA).

FEVEREIRO
2004

Lançamento do Facebook.

NOVEMBRO
2009

Facebook cria equipe de cerca de 12 pessoas responsável pela moderação de conteúdo. Na época, regras de moderação eram apresentadas em apenas uma página.

FEVEREIRO
2013

O relatório *Rabat Plan for Action*, criado no âmbito de reuniões do Alto Comissário das Nações Unidas para Direitos Humanos, é lançado, com recomendações para combater discurso de ódio.

ABRIL
2014

É aprovado o Marco Civil da Internet pelo Congresso Nacional.

MARÇO
2015

Organizações internacionais, por ocasião da conferência RightsCon, publicam os *Manila Principles on Internet Liability,* uma lista de seis princípios para proteger a liberdade de expressão no espaço digital.

NOVEMBRO
2016

Donald J. Trump é eleito presidente dos EUA depois de uma campanha baseada primordialmente no uso de redes sociais.

MAIO
2017

Materiais internos de treinamento para moderadores de conteúdo do Facebook são vazados e publicados pelo The Guardian.

JUNHO
2017

O Bundestag aprova a NetzDG, uma lei que responsabiliza redes sociais que hospedam conteúdos considerados "manifestamente ilegais".

FEVEREIRO
2018

O Departamento de Justiça dos EUA acusa formalmente 13 cidadãos e três entidades russas de interferir nas eleições presidenciais de 2016. A investigação apontou que o Facebook foi usado pelos agentes russos, atingindo 126 milhões de usuários.

MARÇO
2018

O Facebook é surpreendido pela publicação de uma série de reportagens no Guardian e no New York Times sobre o escândalo envolvendo o uso de dados de milhões de usuários dos EUA pela empresa Cambridge Analytica durante as eleições de 2016.

ABRIL
2018

Facebook publica pela primeira vez as guidelines internas que a empresa usa para aplicar os Padrões da Comunidade.

MAIO
2018

Em conferência da Universidade de Santa Clara (EUA), organizações e especialistas lançam os *Santa Clara Principles*, princípios para uma maior transparência e prestação de contas por parte das plataformas sobre moderação de conteúdo.

SETEMBRO
2018

Relatório de missão da ONU diz que Facebook foi "instrumento útil" para disseminar discurso de ódio em Mianmar, em um contexto de genocídio contra a minoria muçulmana no país.

OUTUBRO
2018

Reportagens apontam que o WhatsApp, aplicativo que pertence ao Facebook, foi usado para disparo em massa de mensagens durante as eleições presidenciais brasileiras.

NOVEMBRO
2018

Em comunicado, o fundador e CEO do Facebook, Mark Zuckerberg, fala sobre criar um órgão independente para avaliar a moderação de conteúdo, o que no futuro se tornaria o Comitê de Supervisão.

DEZEMBRO
2018

Reportagem do The New York Times revela mais de 1.400 páginas de manual interno usado pelo Facebook para moderar conteúdo.

MARÇO
2019

Atirador mata 51 pessoas em mesquitas em Christchurch, na Nova Zelândia, e transmite atentado ao vivo pelo Facebook | No mesmo mês, Zuckerberg publica um manifesto no qual defende uma reformulação do Facebook para focar mais na privacidade dos usuários e na criação de comunidades, se afastando da imagem de arena pública.

ABRIL
2019

O governo do Reino Unido publica o Online Harms White Paper no qual propõe um modelo inovador de co-regulação de plataformas digitais e a estipulação de um "dever de cuidado" na Internet.

MAIO
2019

Chefes de Estado e empresas se comprometem a adotar ações para combater o terrorismo e o extremismo violento na internet, no *Christchurch Call to Action* | No mesmo mês, o governo francês publica um relatório propondo um novo marco regulatório das redes sociais no país.

SETEMBRO
2019

Nick Clegg, Vice-presidente global de Políticas Públicas e Comunicação do Facebook, anuncia que a plataforma não iria encaminhar publicações de políticos para verificação de checadores independentes, como faz com outros conteúdos.

JUNHO
2020

Funcionários do Facebook fazem greve virtual e se manifestam contra decisão da empresa de manter publicações controversas do então presidente dos EUA, Donald Trump.

MAIO
2020

Facebook anuncia os primeiros 20 integrantes do Comitê de Supervisão.

MARÇO
2020

Facebook, Instagram e Twitter retiraram vídeo publicado pelo Presidente brasileiro Jair Bolsonaro, em que ele provocava aglomerações e se posicionava contra o isolamento social.

FEVEREIRO
2020

Facebook publica o relatório Charting a Way Forward, no qual identifica os principais desafios da regulação na área e elenca princípios para informar futuras regulações.

Facebook remove contas e perfis falsos ou com "comportamento inautêntico" que, segundo a plataforma, eram ligadas ao Partido Social Liberal (PSL) e a gabinetes de membros da família do presidente brasileiro Jair Bolsonaro.

JULHO
2020

Facebook retira pela primeira vez uma publicação de Trump por desinformação sobre o coronavírus.

AGOSTO
2020

Comitê de Supervisão seleciona e começa a analisar os primeiros casos.

DEZEMBRO
2020

Facebook anuncia que vai encaminhar caso de suspensão do ex-presidente Trump para avaliação do Comitê de Supervisão.

Comitê de Supervisão anuncia as suas primeiras decisões – o Comitê optou por reverter as decisões originais do Facebook em quatro dos cinco casos avaliados.

Facebook, assim como outras plataformas, remove publicações de Trump e suspende por tempo indeterminado as contas do então presidente. A reação da empresa veio após a invasão do Capitólio por apoiadores de Trump.

JANEIRO
2021

MAPA DO RELATÓRIO

O primeiro capítulo, a introdução, destaca a relevância do Facebook como um grande moderador de conteúdo e a sua capacidade, ainda que como empresa privada, de influenciar o debate público de mais de 2 bilhões de usuários. O capítulo também apresenta as principais controvérsias e alguns casos polêmicos sobre moderação de conteúdo na plataforma em diversos países.

No segundo capítulo, os chamados "*community standards*" ou padrões da comunidade do Facebook são explicados em profundidade, apresentando o que é ou não permitido na plataforma. Nesse ponto, o relatório analisa a apresentação e comunicação dessas regras, em português, para o usuário final – há, por exemplo, partes que não foram traduzidas ou que podem ser de difícil compreensão. Já o terceiro capítulo conta como funciona o processo de construção e atualização dos padrões da comunidade dentro da empresa.

O quarto capítulo apresenta o desenvolvimento histórico das regras que permitem às empresas monitorar conteúdo online. O foco é a *section* 230 do CDA nos EUA, uma regra que imuniza plataformas digitais de responsabilização por conteúdos gerados por terceiros. Além disso, o capítulo trata da tensão existente entre a autorregulação das plataformas – o que a regra estadunidense não apenas permite, mas incentiva – e as tentativas de regulação por parte de governos nacionais.

No quinto capítulo analisa-se a criação do *Facebook Oversight Board* ou "comitê de supervisão", criado pela empresa para atuar como órgão independente, com a capacidade de rever decisões sobre a moderação de conteúdo. O Comitê serve para lidar com os questionamentos sobre a legitimidade das decisões da empresa nessa área.

O sexto capítulo avalia a tensão entre a dimensão global da plataforma e a dimensão local, esta última representada pelas diferentes visões dos países onde o Facebook opera e os efeitos sobre a liberdade de expressão. O sétimo capítulo é reservado para a discussão de iniciativas internacionais na área de moderação de conteúdo, a exemplo de documentos como *Santa Clara Principles e Manila Principles*, que buscam estabelecer parâmetros globais sobre moderação de conteúdo a serem seguidos por empresas e governos.

Por fim, o oitavo e último capítulo explora a mudança de foco do Facebook, que passou a se voltar mais para grupos. Esse enfoque apresenta, assim, novos desafios no contexto de moderação de conteúdo.

1. INTRODUÇÃO

Em qualquer sociedade, os seres humanos estabelecem normas básicas para a convivência: desde simples regras de etiqueta para se comportar na mesa a legislações complexas sobre liberdade de expressão. Da mesma forma, o Facebook, que surgiu em 2004 e se tornou a maior rede social do mundo, com mais de 2 bilhões de usuários, se transformou em uma gigantesca sociedade virtual, que também necessita de regras de convivência.

Ao longo dos anos, tendo em conta as pressões da sociedade civil e governos em diversos países, a empresa passou a desenvolver e implementar um conjunto de normas para lidar com conteúdo publicado na plataforma que não estava de acordo com normas nacionais ou com a sua visão sobre o ambiente digital – são os chamados Padrões da Comunidade (ou *community standards*). Esse documento, publicado em 2018, contém as principais diretrizes sobre conteúdos e ações que são permitidos ou não na rede social, como, por exemplo, violência, assédio, discurso de ódio, notícia falsa, nudez e terrorismo.

Um ano antes, porém, materiais internos de treinamento para moderadores de conteúdo do Facebook, muito mais detalhados, já tinham vazado e sido publicados pelo jornal britânico The Guardian, gerando uma série de questionamentos e críticas contra a empresa. No ano seguinte, o jornal americano The New York Times publicou uma reportagem após ter acesso a mais de 1.400 páginas de um manual usado pela empresa para moderar conteúdo.

Nessas últimas duas matérias, que discutem o poder do Facebook para controlar e influenciar o debate público em diversos países, a plataforma aparece associada a expressões como: "possivelmente um dos reguladores políticos mais poderosos do mundo" ou "o maior censor do mundo". Há divergências sobre o tamanho desse poder, mas é inegável que o Facebook possui um papel de importante facilitador e mediador do debate público.

Entretanto, os Padrões da Comunidade, assim como estão publicados pelo Facebook ou descritos nessas reportagens, são fruto de esforços recentes, afirmam os pesquisadores Matthias C. Kettemann e Wolfgang Schulz em seu estudo sobre o tema, feito no âmbito do Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI).

A visão é corroborada pela pesquisadora Kate Klonick, em artigo publicado na Harvard Law Review. No estudo, ela cita entrevista realizada com Dave Willner, que foi chefe de política de conteúdo no Facebook. Ele afirma que, ao menos até

2009, as regras de moderação eram muito limitadas e baseadas em um documento interno de apenas uma página.

"A [política de] orientação tinha cerca de uma página; uma lista de coisas que você deveria deletar: então eram coisas como Hitler e pessoas peladas. Nenhuma dessas coisas estavam erradas, mas não havia um enquadramento explícito sobre por que essas coisas estavam na lista", disse Willner.

Charlotte Willner, esposa de Dave e que também trabalhava no Facebook, disse que o cerne do treinamento para moderação de conteúdo, antes de 2008, era basicamente: "Você se sente mal [com o conteúdo]? Tire [da plataforma]".

Desde então, as regras se tornaram bem mais extensas e específicas, mas ainda são objeto de constante questionamento e controvérsia. Nos últimos anos, o Facebook atraiu críticas ferrenhas em casos como o de Mianmar, em que a plataforma foi acusada de ter feito pouco para impedir a disseminação de discurso de ódio.

Em 2018, uma investigação da ONU acusou militares do país de "limpeza étnica" e genocídio contra a minoria Rohingya em Mianmar. Em seu relatório, a missão diz que as mídias sociais tiveram um "papel significativo", e que o Facebook foi "um instrumento útil para aqueles que buscam espalhar o ódio". O documento ainda alertou que a plataforma foi "lenta e ineficiente" ao lidar com a situação. Um dos investigadores da ONU disse que a empresa tinha se transformado em "um monstro" no país. Posteriormente, o Facebook admitiu ter sido "usado para incitar a violência offline" em Mianmar. Há um risco de que uma situação similar, de uso da plataforma em um contexto de genocídio, se repita na Etiópia, segundo reportagem da Vice.

A empresa tem sido acusada de alimentar o ódio e a desinformação em diversos países, como o Sri Lanka, e servir de plataforma para governos autoritários, como nas Filipinas. Segundo investigações do Departamento de Justiça dos EUA, o Facebook foi usado por agentes russos em uma tentativa de influenciar as eleições presidenciais de 2016, atingindo 126 milhões de usuários, segundo dados apresentados pela própria plataforma. E há fortes indícios de que os esforços de interferência russa para favorecer o presidente Donald Trump por meio das mídias sociais continuaram nas eleições de 2020.

O Facebook também tem sido criticado pela forma como lida com discurso antissemita. Segundo um estudo do Institute for Strategic Dialogue, publicado em agosto de 2020, se uma pessoa passa a seguir uma página pública que tem conteúdo

de negação do holocausto, o algoritmo do Facebook "promove ativamente" mais grupos e publicações com esse perfil para o usuário.

No Brasil, em julho de 2020, a empresa anunciou que havia removido uma rede de contas e perfis falsos e com "comportamento inautêntico" da plataforma, que, segundo o Facebook, eram ligadas ao Partido Social Liberal (PSL) e a gabinetes de membros da família do presidente Jair Bolsonaro. Nas eleições presidenciais de 2018, o Facebook, através da sua subsidiária WhatsApp, se viu envolvido numa crise que afetou sua imagem pública. O aplicativo de mensagens foi usado para o disparo em massa de mensagens durante o pleito, algo que, conforme a própria empresa reconheceu, feria seus termos de uso.

Outro comportamento do Facebook que tem sido alvo de reprovação é a política de tratamento diferenciado dado a políticos, em comparação a usuários comuns. Em 2019, Nick Clegg, Vice-presidente global de Políticas Públicas e Comunicação do Facebook, anunciou que a plataforma não iria encaminhar publicações de políticos para verificação de checadores independentes, como faz com outros conteúdos.

"Nós não acreditamos, entretanto, que é um papel apropriado para nós arbitrar debates políticos e impedir que o discurso de um político alcance seu público e seja sujeito a debate e escrutínio públicos. É por isso que o Facebook isenta os políticos de nosso programa de fact-checking de terceiros", argumentou, em texto em inglês publicado no site da empresa e traduzido livremente.

Da mesma forma, Clegg afirmou que o Facebook já trabalhava com uma noção de "digno de notícia" [*newsworthy*] desde 2016. "Isso significa que se alguém fizer uma declaração ou compartilhar uma postagem que viole os nossos padrões da comunidade, ainda assim permitiremos esse conteúdo em nossa plataforma se acreditarmos que o interesse público de vê-lo supera o risco de danos. Hoje, anunciei que, de agora em diante, trataremos as falas dos políticos como conteúdo digno de notícia que deve, regra geral, ser visto e ouvido", disse Clegg.

No texto, o Vice Presidente faz a ressalva de que o Facebook pode remover publicações de políticos quando avaliar que o conteúdo incita a violência e pode representar um risco para a segurança, o que seria mais importante do que o valor de ser digno de notícia.

Na época, a decisão de tratar políticos de forma diferente recebeu muitas críticas e foi até chamada de "covardia" por Dave Willner, que foi chefe de política de conteúdo do Facebook. Segundo a revista Wired, Willner disse, em texto publicado em seu perfil pessoal, que permitir discurso de ódio na plataforma, seja de cidadãos

ou de políticos, pode criar uma atmosfera perigosa.

Ele afirmou ainda que a decisão do Facebook transformava políticos em uma classe privilegiada, que desfruta de direitos negados aos outros usuários da plataforma. "Não só o Facebook está evitando fazer escolhas difíceis, diz Willner, como está traindo a segurança dos seus usuários para aplacar políticos que ameaçaram regular e até dividir a empresa", afirma a matéria.

Outra reação significativa veio dos próprios funcionários do Facebook que, em junho de 2020, fizeram uma greve virtual e se manifestaram publicamente contra a decisão da empresa de manter publicações controversas do presidente dos Estados Unidos, Donald Trump. Em uma das postagens, no contexto dos protestos contra violência policial em Minneapolis, o presidente sugeriu atirar em manifestantes para evitar saques.

"Estes BANDIDOS estão desonrando a memória de George Floyd, e eu não deixarei que isso aconteça. Acabei de conversar com o governador Tim Waltz e disse que o Exército está com ele até o fim. Qualquer dificuldade e nós assumiremos o controle, mas quando começam os saques, começam os tiros", escreveu Trump.

Ao contrário do Twitter, que sinalizou a mensagem como "apologia à violência", o Facebook decidiu manter as publicações inalteradas, sem qualquer indicação ou rótulo que pudesse alertar os usuários sobre o conteúdo ali veiculado.

Após os protestos dos seus funcionários, Mark Zuckerberg disse que ia revisar algumas políticas da plataforma, inclusive sobre como lidar com ameaças de uso da força por parte do Estado. Em publicação no seu perfil, Zuckerberg afirmou que poderia rever a possibilidade de classificar ou "etiquetar" conteúdo que viole os Padrões da Comunidade, para não ficar limitado à escolha binária de manter ou remover.

Pouco tempo depois, em agosto, o Facebook retirou pela primeira vez uma publicação de Trump por desinformação sobre o coronavírus, quando ele afirmou que as crianças são "quase imunes" à COVID-19. Em outubro de 2020, a plataforma novamente retirou uma postagem do presidente, por afirmar falsamente que a COVID-19 era menos mortal do que uma gripe. Nas duas ocasiões, o Twitter, por exemplo, também removeu ou colocou alertas no conteúdo.

Em novembro de 2020, após Joe Biden ser decretado o vencedor das eleições presidenciais, Trump usou suas redes sociais para, sem apresentar evidências, atacar o processo eleitoral, alegar fraude nos votos por correio, pedir a suspensão da contagem de votos e declarar vitória. Na ocasião, o Facebook e o Twitter colocaram alertas nas mensagens publicadas por Trump. O feed do Presidente ficou repleto de

avisos afirmando que as alegações de fraude eram contestadas por diversas fontes e que Biden (e não Trump) havia vencido o pleito.

Em janeiro de 2021, o Facebook, assim como outras plataformas, reagiu de forma mais dura quando houve a invasão do Capitólio por apoiadores de Trump, durante a sessão conjunta entre Deputados e Senadores para a certificação da vitória de Joe Biden. Além de remover publicações, foram bloqueadas as contas de Trump por tempo indeterminado. Também foram anunciadas uma série de medidas, em tempo real, para restringir conteúdo que pudesse estimular a violência.

"Acreditamos que os riscos de permitir que o Presidente Trump continue a usar nosso serviço durante este período (da transição presidencial nos EUA) são simplesmente grandes demais", afirmou, por exemplo, a empresa em comunicado publicado no seu site. Sobre as publicações do Presidente retiradas do ar, a nota dizia: "Tomamos essa decisão considerando que, no geral, esses posts contribuem, em vez de diminuir, o risco de violência contínua."

Na ocasião, o Twitter apagou posts do Presidente e baniu a conta de Trump por 12h. Depois do prazo permitiu que ele retomasse as publicações, mas anunciou, no dia 8 de janeiro, que a conta dele tinha sido permanentemente suspensa.

No Brasil, em março de 2020, o Facebook, Instagram e Twitter retiraram do ar um vídeo publicado pelo Presidente Bolsonaro, em que ele provocava aglomerações e se posicionava contra o isolamento social, indo contra as evidências científicas apresentadas por autoridades da área.

Para além da política, outro tema costuma causar polêmica na plataforma: nudez. Em 2018, o Facebook admitiu ter errado ao suspender a conta da Funai (Fundação Nacional do Índio) por sete dias por uma foto de mulheres indígenas com os seios à mostra. Alguns anos antes, em 2016, a plataforma também tinha voltado atrás após a decisão controversa de apagar uma foto icônica da Guerra do Vietnã, de uma menina nua correndo dos bombardeios de napalm.

Em suma, redes sociais desempenham um importante papel de moderação de conteúdo, estipulando limites para o que é ou não permitido em suas plataformas. Embora diversos usuários estejam cientes disso, poucos de fato conhecem as regras criadas por essas plataformas para guiar o processo de moderação. Na próxima seção, então, será feita uma breve análise dos Padrões da Comunidade do Facebook de forma a melhor ilustrar esse debate.

2. O QUE PODE E O QUE NÃO PODE?

Os Padrões da Comunidade são o conjunto de regras do Facebook que explicam o que é permitido ou não na plataforma. Para este relatório, analisamos a versão em português, publicada no site do Facebook, e acompanhamos as atualizações até 14 de janeiro de 2021. Ou seja, mudanças feitas a partir dessa data não estão incluídas nesta análise.

Os Padrões da Comunidade estão divididos em seis partes: comportamento violento e criminoso, segurança, conteúdo questionável, integridade e autenticidade, respeito à propriedade intelectual, e solicitações e decisões relativas a conteúdo.

A parte de comportamento violento e criminoso congrega cinco itens. Um deles trata sobre a proibição de conteúdo que incite a violência, como ameaças, defesa do uso da violência ou declarações de intenção de cometer um ato violento. Há desde diretrizes mais amplas, até regras mais específicas, como por exemplo: tentar contratar um pistoleiro ou um assassino para um homicídio é proibido; mas fornecer instruções sobre como fazer ou usar explosivos é permitido quando tiver um evidente "propósito não violento". O Facebook cita, como exemplo, explosivos no contexto de videogames, shows pirotécnicos e pesca ou, então, explosivos com "um propósito claramente educacional/científico".

O item dois explica que indivíduos ou organizações envolvidos em terrorismo, ódio organizado, assassinato em massa, tráfico humano, violência organizada ou atividade criminosa não podem estar na plataforma. Em seguida, o Facebook veda a "coordenação de danos e divulgação de crime". Esse item contém uma série de atividades proibidas, muito variadas, como difundir danos contra animais, inclusive luta entre eles, vandalismo e até "fraude sob a condição de eleitor".

Há ainda uma seção que relaciona "Produtos controlados", onde se descreve a proibição da compra e venda de drogas não medicinais, medicamentos controlados, maconha, artefatos históricos, sangue, animais, produtos que prometem perda milagrosa de peso. O mesmo item restringe a comercialização de armas de fogo, permitindo apenas o conteúdo de lojas oficiais ou agências do governo e voltado para maiores de 21 anos, por exemplo. Logo depois, o Facebook detalha também que não é possível usar a plataforma para fraudes, esquemas de pirâmide e golpes em geral.

A segunda parte do documento aborda a segurança. A plataforma define que não é permitido conteúdo que incentive automutilação e suicídio, mas pode aceitar algumas imagens e mensagens sobre o tema se estiverem em um contexto de recu-

peração e conscientização. Há pontos que são pouco claros, como por exemplo: "fotos ou vídeos, julgados como interessantes, que exibam o suicídio de alguém" são permitidos, só que são restritos a maiores de 18 anos e antecedidos de uma tela de sensibilidade, para alertar as pessoas de que o conteúdo pode ser desagradável.

Na versão em inglês, essa regra específica é mais fácil de entender, porque não há o termo "interessante", e sim "newsworthy", o que significa algo como "digno de notícia" – expressão que aparece em pontos da versão em português para substituir "newsworthy", o que pode indicar uma inconsistência na tradução.

O segundo e terceiros itens da parte de segurança explica que é proibido conteúdo relacionado à exploração sexual, abuso ou nudez infantil, além de exploração sexual de adultos. Não é permitido qualquer tipo de pornografia por vingança, extorsão sexual ou publicações compostas de contato sexual não consensual.

O Facebook deixa claro que veda qualquer conteúdo dentro do conceito de exploração humana, que inclui tráfico de pessoas, trabalho escravo, tráfico de órgãos entre outros. A parte de segurança relata ainda a impossibilidade de publicar dados e documentos, que violem a privacidade dos outros, como número de CPF, identidade e informações bancárias.

Outro ponto importante e longo desse segmento é uma tentativa de descrever comportamentos de bullying e assédio, que não poderiam estar na plataforma. Nesse item, o Facebook tem políticas diferentes para figuras públicas ou indivíduos privados, sendo que pessoas comuns têm maior proteção. Da mesma forma, menores de idade, que podem ser figuras públicas ou não, também são mais amparados pelas políticas para evitar assédio e bullying.

São vetados, por exemplo, "apelos à morte, deficiência ou doença grave ou epidêmica", ainda que o Facebook não exemplifique de que forma isso pode ocorrer. Já quando se trata da comparação com animais pejorativos, ela aparentemente é permitida para figuras públicas adultas, mas não para "figuras públicas involuntárias", figuras públicas menores de idade ou indivíduos privados.

A indicação dessas diferenças sobre o que vale ou não para cada um desses grupos, entretanto, nem sempre parece ser clara, tem partes de difícil compreensão e que podem gerar dúvidas. Especialmente porque, nesse ponto, há alguns problemas de tradução. Em inglês, está escrito "Do not [...] Target private individuals or involuntary public figures with", ao que se segue uma série de práticas proibidas. Já a versão em português omite a palavra "involuntária" duas vezes, por isso pode ser complicado entender qual regra se aplica a cada grupo.

Na terceira parte dos Padrões da Comunidade, o Facebook entra no mérito

do "conteúdo questionável", que ele caracteriza como discurso de ódio; violência e conteúdo explícito; nudez adulta e atividades sexuais; abordagem sexual; conteúdo cruel e insensível.

No item sobre violência e conteúdo explícito, a plataforma detalha, por exemplo, que não é permitido "vídeos de pessoas ou cadáveres em instalações não médicas se retratarem" desmembramento, vítimas de canibalismo, corte de garganta, entre outros tipos de agressão. Entretanto, fotos de pessoas feridas ou mortas nessas mesmas situações são autorizadas, mas vêm acompanhadas de uma tela de alerta e são restritas a maiores de 18 anos.

O Facebook também proíbe qualquer conteúdo com nudez ou atividade sexual, mas há algumas exceções, como em contexto educativo, científico, médico, de amamentação, de defesa de certas causas, de obras de arte ou com fins humorísticos e satíricos.

No item sobre abordagem sexual, o documento deixa claro que atividades de prostituição e serviços de acompanhante são proibidos, bem como fazer "proposta de cunho sexual explícita" ou usar "linguagem sexualmente explícita", entre outras atividades.

Em seguida, a plataforma descreve o conteúdo considerado cruel e insensível, que inclui zombar de pessoas por doença, deficiência grave ou fatal, fome ou ferimento físico grave. Também veda qualquer conteúdo que mostre pessoas experimentando "morte prematura, danos físicos sérios ou violência doméstica", além de determinar alguns conteúdos cruéis contra animais.

Um dos pontos mais importantes da parte "conteúdo questionável" trata sobre discurso de ódio, que o Facebook define como um ataque a pessoas com base em suas "características protegidas": raça, etnia, nacionalidade, filiação religiosa, orientação sexual, casta, sexo, gênero, identidade de gênero e doença grave ou deficiência. Segundo o documento, também há certas proteções a imigrantes. Não há, entretanto, qualquer proteção especial para origem geográfica dentro de um mesmo território nacional. Ataques relacionados à idade de uma pessoa podem ser removidos caso sejam combinados com agressões contra alguma característica protegida.

Dentro desse ponto, o Facebook explica que o contexto e a intenção importam, isto é, palavras que são proibidas podem ser mantidas caso a pessoa esteja falando sobre si mesma, denunciando discurso de ódio de outros ou defendendo uma causa.

Importante notar que os ataques são divididos em três níveis de gravidade. O caso mais grave diz respeito a conteúdo que visa um indivíduo ou grupo por suas

características protegidas ou status de imigração com discurso violento, comparações com animais ou criminosos, entre outros. Aqui o Facebook dá exemplos claros: não se pode associar negros a macacos, judeus a ratos, mulheres a objetos etc.

Já o segundo nível inclui associar aspectos físicos, capacidade intelectual e traços morais a grupos ou pessoas segundo suas características protegidas, por exemplo: tal grupo é feio, todos são burros ou covardes. Também entram nesse quesito expressões de superioridade, como "homens são superiores às mulheres", ou manifestações de desprezo, repulsa ou nojo.

O Facebook também enumera vários palavrões e expressões proibidas – este é, provavelmente, um dos pontos em que o documento é mais específico e preciso. Não pode "pau no cu", "cuzão", "escroto", para ficar em poucos exemplos.

Por fim, há uma descrição dos ataques menos graves, dentro do quesito discurso de ódio, que seriam os conteúdos que incitem a segregação ou a exclusão política, econômica ou social de pessoas ou grupos por suas características protegidas. Aqui o Facebook faz a ressalva de que permite críticas a políticas de imigração e argumentos favoráveis à sua restrição, embora não explique quais discursos são aceitos.

A quarta parte dos Padrões da Comunidade trata de "integridade e autenticidade". O primeiro ponto fala sobre identidades falsas e explica a importância da autenticidade para a plataforma: "Acreditamos que a autenticidade ajuda a criar uma comunidade na qual as pessoas, de maneiras significativas, tenham responsabilização umas para com as outras e com o Facebook". Assim, o documento deixa claro que o usuário não pode falsificar sua identidade, seus dados, criar contas para outras pessoas ou tentar se passar por alguém.

Em seguida, os Padrões da Comunidade interditam *o spam*, o que abarca, entre outras coisas, métodos de massa e enganosos para atrair usuários para visualização ou comércio de produtos. O item sobre segurança cibernética, que aparece logo depois, não estava traduzido para o português até janeiro de 2021, e não há uma explicação aparente na página para isso. Em inglês, o trecho determina que não é permitido qualquer *software* ou arquivo que tente ter acesso não autorizado a "informações pessoais sensíveis" ou a um aparelho ou rede, entre outras práticas maliciosas.

A seção sobre "comportamento não autêntico" discorre sobre usos indevidos do Facebook, de forma coordenada ou não, para encobrir a finalidade de páginas, enganar pessoas sobre a origem ou a fonte de um conteúdo, ter contas falsas, entre outros. Também não autoriza a "interferência governamental ou estrangeira",

entendida como "comportamento não autêntico coordenado realizado em nome de um ator estrangeiro ou governamental".

Ainda dentro de "integridade e autenticidade", o Facebook aborda a questão das notícias falsas. Esse ponto é mais uma breve descrição das políticas da empresa para lidar com esse problema do que uma série de regras. "Existe uma linha tênue entre notícias falsas e sátiras ou opiniões. Por esse motivo, não removemos notícias falsas do Facebook, mas, em vez disso, reduzimos significativamente sua distribuição, mostrando-as mais abaixo no Feed de Notícias", afirma a empresa.

O Facebook explica brevemente que bloqueia incentivos econômicos a páginas, pessoas e domínios que disseminem notícias falsas, além de trabalhar com organizações independentes que fazem checagem de dados para diminuir a distribuição desse tipo de conteúdo.

A parte de "integridade e autenticidade" aborda ainda "mídia manipulada", que impede a edição de imagens, áudios e vídeos com o objetivo de enganar as pessoas, técnicas de *deepfake*, entre outras práticas. O Facebook faz a ressalva de que conteúdo editado voltado para paródia ou sátira é permitido. A seção termina com a explicação sobre os perfis memoriais, que são permitidos pelo Facebook e servem para homenagear pessoas que morreram.

As partes finais dos Padrões da Comunidade tratam de direitos de propriedade intelectual, solicitações e decisões relativas a conteúdo, proteção adicional de menores e o Comitê de Supervisão. O último conteúdo, informações adicionais, que trata do "envolvimento das partes interessadas", também não possuía tradução para o português até janeiro de 2021.

3. COMO É O PROCESSO DE CRIAÇÃO E ATUALIZAÇÃO DOS PADRÕES DA COMUNIDADE?

O Facebook faz atualizações constantes nos Padrões da Comunidade. Algumas são mudanças substanciais, outras, edições para tornar o texto mais claro e preciso, bem como reorganizações na apresentação do conteúdo.

O processo de desenvolvimento dos Padrões da Comunidade no Facebook tem várias etapas, mas a principal é uma reunião que acontece cerca de duas vezes por mês, que se chama Fórum de Políticas de Produto (Product Policy Forum). No Fórum, funcionários de 11 escritórios do Facebook pelo mundo debatem possíveis mudanças nos Padrões da Comunidade.

Para se preparar para esses encontros, eles analisam dados e pesquisas, bem como consultam especialistas internos e externos, como acadêmicos e ativistas. O objetivo é chegar a regras que sejam claras e possam ser aplicadas de forma consistente, sem sobrecarregar os moderadores de conteúdo, em diferentes países e contextos culturais.

Segundo os pesquisadores Matthias C. Kettemann e Wolfgang Schulz, que acompanharam o desenvolvimento dessas políticas para uma pesquisa, os temas são ordenados de acordo com métricas. Esses parâmetros avaliam o impacto e a viabilidade da mudança, a prevalência do problema, o quanto a questão é controversa etc.

Um desafio, de acordo com os estudiosos, é que algumas questões são tão novas que nem mesmo os especialistas consultados têm uma resposta. Há também uma dificuldade em trazer maior diversidade para o processo de decisão, principalmente com visões de mundo não ocidentais.

No estudo, Kettemann e Schulz falam da primazia de certos valores do Facebook sobre outros. Em 2019, a empresa apresentou duas mudanças, uma delas foi a criação do Comitê de Supervisão (ver mais detalhes abaixo) e a outra foi uma introdução aos Padrões da Comunidade, que descrevem os valores do Facebook.

No texto, a chefe de Global Policy Management, Monika Bickert, afirma que dar voz às pessoas continua sendo um valor supremo, mas que o Facebook passou a focar em autenticidade, segurança, privacidade e dignidade ao elaborar e implementar as suas políticas. E que eles estão dispostos a limitar a liberdade de expressão quando isso esbarra nos outros valores.

Em alguns casos, o Facebook permite conteúdo que vai contra seus Padrões da Comunidade, quando avalia que é "digno de notícia e de interesse público". "Fazemos isso apenas depois de pesar o valor do interesse público contra o risco de dano, e olhamos para os padrões internacionais de direitos humanos para fazer esses julgamentos", afirma o texto.

Segundo Kettemann e Schulz, a primazia da "voz" mostra que há uma preferência contra a remoção de conteúdo. E esse valor foi posteriormente reforçado pelas noções de "digno de notícia e de interesse público". Da mesma forma, os pesquisadores afirmam que, em suas observações de reuniões no Facebook, "dar voz para as pessoas" tem um peso maior na formulação de regras do que o princípio da inclusão ou de "unir pessoas", que acaba ficando em segundo plano.

4. REGULAÇÃO *VERSUS* AUTORREGULAÇÃO

Nos EUA, onde redes sociais como o Facebook e Twitter foram criadas no início dos anos 2000, plataformas digitais estão legalmente escudadas de serem responsabilizadas pelo conteúdo postado por terceiros. A origem dessa isenção de responsabilidade está em dois casos importantes que foram julgados nos anos 90.

O primeiro caso é conhecido como *Cubby v CompuServe* e foi decidido em 1991 pela Corte do Distrito Sul de NY. No entendimento da Corte, a CompuServe, uma desenvolvedora de plataformas digitais, apenas distribuía o conteúdo postado pelos usuários. Assim, ao contrário de um jornal ou de uma emissora de TV – que contam com um corpo editorial próprio, a CompuServe não tomava decisões de editoração.

Na compreensão da Corte, a empresa apenas disponibilizava um espaço onde terceiros poderiam publicar qualquer tipo de conteúdo – ou seja, uma plataforma – e que essas postagens não seriam de responsabilidade da empresa, uma vez que não supervisionava o conteúdo postado por terceiros. Trata-se de uma distinção feita pela jurisprudência do *common law* (a tradição legal herdada pelos EUA) entre *publisher* (editor) e *distributor* (distribuidor) para fins de responsabilização civil pelo conteúdo hospedado na plataforma. Enquanto o editor participa da formação do conteúdo em si, o distribuidor apenas oferece uma plataforma através da qual o conteúdo é veiculado. Portanto, o distribuidor tem menor ou nenhuma responsabilidade sobre o conteúdo disponibilizado.

Já o segundo caso é conhecido como *Stratton Oakmont v Prodigy Services* e foi decidido em 1995 pela Suprema Corte de NY. No entendimento dos juízes, a Prodigy Services, outra desenvolvedora de plataformas digitais, seria responsável pelo conteúdo postado por seus usuários porque ativamente moderava algumas das postagens feitas por terceiros em seu ambiente digital.

No caso, a Prodigy Services havia deletado alguns comentários considerados ofensivos. Assim, a Corte afirmou que uma empresa que opta por moderar o conteúdo disponibilizado em sua plataforma passa a ser responsável por eventuais ilegalidades ocorridas dentro do seu domínio. Ao moderar o conteúdo, a empresa se assemelha a uma editora (ou *publisher*) e deixa de ser apenas uma distribuidora (ou *distributor*). Na prática, ainda que a moderação fosse pontual (e provavelmente justificada), a Prodigy Services acabava assumindo a responsabilidade por toda e qualquer postagem.

Assim, esses dois casos acabaram criando um incentivo considerado nefasto.

Moderar conteúdos ofensivos passou a ser um risco para a plataforma que poderia ser responsabilizada por eventuais ilegalidades cometidas pelos usuários - inclusive sem seu conhecimento. Por outro lado, não agir mesmo diante de conteúdos ofensivos se tornou uma defesa, uma forma de evitar a responsabilização. Ou seja, melhor não moderar e fazer valer o precedente *CompuServe* de 1991.

De forma a harmonizar esses dois precedentes e excluir o incentivo pernicioso criado, o Congresso dos EUA aprovou a famosa *Section* 230 como parte do *Communications Decency Act* de 1996. De acordo com a redação do artigo, "nenhum provedor ou usuário de serviço interativo de computador será tratado como editor de qualquer informação fornecida por terceiros". Ou seja, plataformas digitais não são editoras e, consequentemente, não podem ser responsabilizadas pelo conteúdo de seus usuários.

O artigo também conta com uma "cláusula do bom samaritano", que prevê a extensão da imunidade mesmo àquelas plataformas que optem por moderar conteúdos que considerem obscenos ou ofensivos. Em outras palavras, se provedores de serviços interativos de computador optam pela via da autorregulação, moderando o conteúdo por iniciativa própria, eles não precisam mais temer processos dessa natureza.

De forma distinta, no Brasil não há algo que se assemelhe à "cláusula do bom samaritano" dos EUA. O Marco Civil da Internet, em seu artigo 19, estipula que "o provedor de aplicações de internet somente poderá ser responsabilizado civilmente por danos decorrentes de conteúdo gerado por terceiros se, após ordem judicial específica, não tomar as providências para [...] tornar indisponível o conteúdo apontado como infringente". Em outras palavras, a lei brasileira condiciona a possibilidade de responsabilização civil à existência de uma ordem judicial específica.

Mas se essas plataformas não correm o risco da responsabilização desde o advento da *Section* 230 em 1996, por qual motivo elas voluntariamente optam por moderar conteúdo online? Vale lembrar que a doutrina estadunidense de liberdade de expressão é altamente permissiva, uma vez que a primeira emenda à Constituição dos EUA é interpretada extensivamente pela Suprema Corte daquele país. Em outras palavras, os EUA é um dos países mais permissivos em relação à liberdade de expressão de seus cidadãos, permitindo, inclusive, a circulação de conteúdos expressamente nazistas.

Era de se esperar, assim, que empresas como o Facebook e o Twitter adotassem uma perspectiva igualmente liberal. Nada obstante, é importante notar que recentemente o Facebook vem se afastando gradualmente dessa posição. Em 2018,

Zuckerberg defendeu a decisão da empresa de permitir a negação do holocausto na plataforma. Já em 2020, como notado anteriormente, o Facebook anunciou que passaria a proibir esse tipo de discurso em razão da conexão entre o crescimento do antissemitismo e a ignorância histórica a respeito do holocausto, principalmente entre os mais jovens.

Levando em consideração o estudo de Kate Klonick, existem pelo menos dois grandes motivos por trás do desenvolvimento de sistemas de autorregulação pelas próprias plataformas. Em primeiro lugar, elas querem evitar possíveis novas regulações por parte do poder público. Ao criar e implementar um processo de moderação, essas empresas passam a impressão de que são responsáveis e capazes de lidar com conteúdo nocivo e, assim, seriam capazes prevenir (ou ao menos retardar) uma possível intervenção estatal.

Em segundo lugar, há um incentivo econômico. Afinal, as plataformas precisam criar um ambiente acolhedor para atrair e manter usuários online além de promover engajamento. Caso contrário, elas podem acabar afastando usuários que não se sentem seguros online. Assim, conteúdos obscenos e ofensivos (ou, em geral, sensíveis) devem sofrer moderação para evitar a alienação de parcelas do público.

Em suma, desde a década de 90, quando os casos *CompuServe* e *Prodigy Services* foram decididos e a *Section* 230 do CDA foi aprovada pelo Congresso dos EUA, as diversas plataformas digitais passaram a atuar dentro de um paradigma de autorregulação. Estabelecem regras internas, como os Padrões da Comunidade, contratam moderadores de conteúdo, desenvolvem algoritmos de moderação e removem publicações que não estão de acordo com suas propostas de comunidade.

Entretanto, pelo menos desde a segunda metade dos anos 2010 os ventos começaram a soprar em outra direção. O paradigma da autorregulação passou a ser questionado. Diversos países passaram a pressionar as empresas donas das redes sociais mais influentes, exigindo que elas fossem mais eficientes no combate à desinformação online. Isto é, elas deveriam atuar de maneira mais intensa e moderar não apenas posts obscenos ou ofensivos, mas também conteúdos mais controversos como publicações deliberadamente falsas ou com o intuito de enganar outros usuários.

Nesse contexto, em junho de 2017, o Congresso Alemão (*Bundestag*) aprovou a NetzDG (ou *Network Enforcement Act*), uma lei que obriga redes sociais ou outros provedores que hospedam informações de terceiros a remover conteúdo "manifestamente ou obviamente ilegal" dentro de do prazo exíguo de 24 horas após serem notificadas. Caso deixem de retirar o conteúdo no prazo legal, as empresas poderão

ser multadas em até 50 milhões de euros.

A lei alemã também faz uma distinção entre conteúdo "ilegal" e "manifestamente ilegal". Enquanto o prazo para a retirada do segundo tipo de conteúdo é de 24 horas, para o primeiro, é de sete dias. Além disso, a lei exige que empresas como o Facebook criem um procedimento próprio para receber e processar esse tipo de notificação, tomando uma decisão dentro do prazo estipulado.

Uma série de dificuldades podem ser apontadas quanto a NetzDG. A lei cria um incentivo para que plataformas digitais "pequem pelo excesso", retirando conteúdos que potencialmente se enquadram na categoria de "ilegal" ou "manifestamente ilegal" para cumprir o prazo e evitar a multa. Após uma análise detida, se o conteúdo na verdade for considerado legal, a empresa poderá restaurá-lo. Ou seja, um dos efeitos colaterais da lei alemã é uma maior limitação à liberdade de expressão em redes sociais.

Ademais, a solução alemã terceiriza para o Facebook a tarefa de interpretar o conteúdo postado por seus usuários à luz do direito alemão. Essa tarefa, que usualmente recau sobre o poder estatal, acaba sendo delegada para uma empresa privada, que deverá decidir o que o direito alemão significa, traçando o limite entre o legal e o ilegal naquele país.

A solução alemã pode ser contrastada com as soluções elaboradas pela França e pelo Reino Unido, que buscam estabelecer obrigações procedimentais e lidam menos com a legalidade de um determinado conteúdo. Em maio de 2019, a França publicou um relatório expondo os contornos de um possível marco regulatório para as redes sociais no país. Segundo o documento, o objetivo é estabelecer um equilíbrio entre a abordagem punitivista (como a NetzDG) e a preventiva, tornando o processo de moderação de conteúdo das plataformas mais transparente e responsável.

De acordo com relatório francês, a assimetria de informação entre órgãos governamentais e as redes sociais justifica a intervenção do poder público na esfera da moderação de conteúdo. Empresas como o Facebook desenvolveram sistemas de autorregulação que, segundo as autoridades francesas, são pouco transparentes e cujo objetivo primordial é evitar novas regulações por parte do poder público. Assim, autoridades governamentais devem adotar as medidas necessárias para que o processo de moderação de conteúdo seja informado também pelo interesse público e não apenas pelo interesse privado dessas empresas.

O relatório francês reconhece que medidas punitivas podem levar a uma maior censura na internet, terceirizando para empresas privadas a responsabilidade

pela interpretação da lei nacional. Assim, aposta na criação de um marco regulatório menos focado na punição cujo objetivo seria estabelecer uma "obrigação de transparência" associada a uma "obrigação de defender a integridade dos usuários".

Para atingir esse objetivo, o documento propõe a criação de um órgão independente responsável por monitorar a implementação dessas duas obrigações e formado por representantes do governo francês. Daí a ideia que a proposta das autoridades francesas se enquadra numa solução procedimental, ou seja, muda o processo de moderação de conteúdo em si e aposta na combinação de autorregulação das empresas com regras mínimas estabelecidas pelo Estado para preservar o interesse público.

No mesmo sentido, em abril de 2019 o governo do Reino Unido publicou o *Online Harms White Paper,* propondo um novo sistema de regulação das redes sociais a ser implementado por um órgão regulador público e independente. Segundo o documento, esse órgão seria responsável pela implementação de padrões para garantir a segurança dos usuários nas redes sociais e, ao mesmo tempo, deveria se preocupar com a proteção da liberdade de expressão no ambiente digital.

Assim, o relatório inglês aposta na criação de um "dever de cuidado" por parte das redes sociais e na promoção de uma "cultura de transparência, confiança e prestação de contas". Dentre outras atividades, o órgão regulador produziria "códigos de boas práticas" para as redes sociais, monitoraria a implementação do "dever de cuidado", prepararia relatórios a respeito do processo de moderação de conteúdo e, por fim, promoveria campanhas de educação e conscientização do público sobre os desafios impostos pela liberdade de expressão online.

A abordagem "procedimental", a exemplo do que foi proposto pelos relatórios da França e do Reino Unido, está mais alinhada com os interesses de plataformas globais como o Facebook. Até porque requisitos como o de tornar o processo de moderação mais transparente, publicar periodicamente relatórios de moderação de conteúdo e respeitar um "dever de cuidado" ou uma "obrigação de defender a integridade dos usuários" podem ser aplicados em escala global.

Já as propostas apresentadas pela NetzDG, que requerem que as plataformas interpretem a lei do país, exigem soluções locais. Afinal, o que é "manifestamente ilegal" na Alemanha pode não ser na França, embora seja possível dialogar e atingir um denominador comum a respeito de como tornar o processo de moderação mais transparente em ambos os países.

Até mesmo nos EUA – onde desde os anos 90 impera o paradigma da autorregulação associado à imunidade de plataformas digitais – começaram a surgir propos-

tas de reforma da *Section* 230 do CDA. Uma das mais conhecidas foi defendida por Danielle Citron e Benjamin Wittes em um estudo publicado em 2017. Segundo os autores, ao invés de garantir imunidade às plataformas digitais, a proteção ofertada pela *Section* 230 deveria ser condicionada à adoção de "medidas razoáveis de prevenção ou remediação de usos ilegais de seus serviços".

Outra medida ventilada recentemente por Danielle Citron e Mary Anne Franks seria negar a imunidade da *Section* 230 às empresas que deliberadamente mantêm em suas plataformas conteúdo "inequivocamente ilegal" ou capaz de gerar "sérios danos à terceiros". Ainda é cedo para dizer se essas propostas irão receber o apoio necessário no Congresso dos EUA, mas elas já cumpriram o papel de provocar um debate a respeito do tema.

A *Section* 230 virou alvo não apenas de acadêmicos, mas também do legislativo norte-americano. O Senador republicano Josh Hawley apresentou ao Congresso dos EUA o *Ending Support for Internet Censorship Act* em 2019. Segundo o projeto, a imunidade da *Section* 230 passaria a ser condicionada à apresentação de evidências "claras e convincentes" de que os algoritmos e as práticas de moderação de conteúdo das redes sociais são "politicamente neutras".

Ainda em 2019, o Senador republicano Jon Kyl publicou o *Covington Interim Report*, no qual sugere que o Facebook está apresentando um viés anti-conservador em suas práticas de moderação de conteúdo, além de apontar possíveis mudanças institucionais com as quais a empresa deveria se comprometer. Desde então, o Facebook se encontra numa encruzilhada. Enquanto há democratas que parecem acusar a empresa de favorecer republicanos, também há republicanos que reclamam que as práticas de moderação são injustas quanto a vozes conservadoras.

A discussão apresentada até aqui ilustra o arco atravessado pelo processo de moderação de conteúdo através dos anos. Em artigo publicado na Harvard Misinformation Review, os professores John Bowers e Jonathan Zittrain descrevem as três fases desse processo histórico. A primeira era, que se estende desde os anos 1990 até aproximadamente 2010, é caracterizada pela criação de novos direitos e tem como marco a *Section* 230 do CDA. A ideia durante os primórdios da regulação da internet era proteger a liberdade de expressão nas redes, afastando tentativas de interferência por parte do poder estatal e promovendo inovação.

A partir de 2010, a discussão ganha um novo centro gravitacional: as consequências das novas tecnologias para as relações humanas, principalmente na esfera política. Esse giro paradigmático inaugura o que Bowers e Zittrain chamam de era da saúde pública. Plataformas como Facebook e Twitter são pressionadas a adota-

rem medidas para mitigar os efeitos nefastos de campanhas de desinformação, crimes cibernéticos, tentativas de supressão de voto através de anúncios políticos, comportamentos inautênticos nas redes (através do uso de robôs e contas falsas), entre outras atividades.

Nessa segunda era, a *Section* 230 do CDA é o principal alvo. O argumento prevalecente é que plataformas digitais não devem gozar de uma imunidade absoluta justamente porque isso gera um incentivo para que elas ignorem as práticas descritas acima. O problema é que o conflito intergeracional entre a era dos direitos e a era da saúde pública acabou gerando um impasse que aparentemente não será fácil de ser superado organicamente. Há dúvidas sobre os limites da atividade regulatória frente à liberdade de expressão. Até porque não há um consenso - particularmente não internacional - sobre o tipo de discurso objeto de proteção da liberdade de expressão e de como na prática efetuar essa distinção.

Isso de alguma maneira está refletido na discussão norte-americana. De um lado, há o descontentamento dos conservadores americanos que acreditam que o Facebook não é receptivo aos ideais de direita, como oposição ao aborto, criminalização do uso de drogas e promoção ampla do livre mercado. Do outro lado, democratas pressionam o Facebook para que a empresa seja mais rigorosa no processo de revisão de postagens, a exemplo daquelas do ex-Presidente Donald Trump que consideram antidemocráticas, racistas e/ou homofóbicas. Para vencer esse impasse, Bowers e Zittrain apostam numa emergente era do processo. O foco deve ser em promover transparência e *accountability*, deixando o debate de valores para trás.

As propostas de regulamentação na França e no Reino Unido parecem se enquadrar nesse contexto. Na lógica de Bowers e Zittrain, plataformas digitais devem investir na formulação de regras claras, na estruturação de mecanismos recursais, na apresentação de razões e justificativas para as decisões tomadas por funcionários da empresa e, principalmente, na apresentação transparente ao público de dados e relatórios envolvendo o processo de moderação de conteúdo.

Há diversos caminhos que podem ser explorados. Dentro desse quadro procedimental, por exemplo, parece estar a lógica da criação do Facebook Oversight Board, um órgão independente da empresa com capacidade de promover mais previsibilidade e transparência quanto à moderação de conteúdo pela rede social.

5. COMITÊ DE SUPERVISÃO

No final de 2018, o fundador e CEO do Facebook, Mark Zuckerberg, escreveu pela primeira vez sobre a intenção de criar um órgão independente para avaliar a moderação de conteúdo, que foi chamado de Comitê de Supervisão. Na ocasião, ele reconheceu: "o Facebook não deveria tomar tantas decisões importantes sobre liberdade de expressão e segurança por conta própria". Em uma entrevista anterior, no mesmo ano, ele já havia manifestado o desejo de criar uma instância independente, que ele descreveu como "um tribunal superior".

Segundo o Facebook, o Comitê foi criado para fazer recomendações e para exercer um julgamento independente sobre o conteúdo que aparece na plataforma. Os usuários podem recorrer ao Comitê quanto a decisões de conteúdo tomadas pelo Facebook e pelo Instagram, mas o grupo não é obrigado a avaliar todos os casos. O objetivo é julgar os mais emblemáticos e desafiadores, que possam embasar políticas e definir precedentes.

A plataforma se comprometeu a seguir as decisões do grupo –que serão públicas– a menos que a implementação delas configure uma violação de alguma lei. Quando estiver completo, o Comitê vai ser composto por cerca de 40 membros de diversos países e diferentes formações. Em maio de 2020, o Facebook anunciou os primeiros 20 integrantes, contando com um brasileiro.

O Comitê começou a operar no final de 2020. O órgão é financiado por um trust, estabelecido pela empresa, de caráter "irrevogável e independente", de acordo com o estatuto do Comitê.

O anúncio do "tribunal supremo" foi recebido, ao mesmo tempo, com esperança e desconfiança. Alguns especialistas mais otimistas veem o Comitê como um marco da regulação na Internet, como é o caso da pesquisadora Evelyn Douek, doutoranda na Harvard Law School. "Isso representa um momento fundamental quando novas formas constitucionais podem emergir que vão moldar o futuro do discurso online", afirmou ela, em artigo.

Por outro lado, especialistas mais céticos afirmam que o Comitê é uma jogada de marketing do Facebook e uma tentativa da empresa de se proteger de regulações e críticas externas. Siva Vaidhyanathan, professor de estudos de mídia da Universidade da Virgínia e autor de um livro sobre o Facebook, disse ao The Guardian que o novo Comitê é uma tentativa de "greenwashing" – termo usado quando empresas querem criar uma imagem ambientalmente responsável, enquanto ocultam ou desviam atenção de suas próprias ações negativas para o meio ambiente.

Da mesma forma, o Facebook foi acusado de usar o Comitê como um escudo para evitar qualquer tipo de controle externo. O acadêmico e professor John Naughton disse, em coluna no The Guardian, que o Facebook sofre de um "delírio de que é um Estado-nação". Para ele, o Comitê é a prova de que a empresa quer se autorregular e é parte do que ele chama de "uma mais ampla e perturbadora tendência" de ceder poder dos soberanos territoriais, como tribunais e legisladores, para "soberanos funcionais", como Amazon, Google e Facebook.

Outra crítica recorrente é que não faria sentido criar um "tribunal superior", sem uma Constituição, ou seja, não basta criar um Judiciário, sem um Poder Legislativo. E o mais produtivo seria que um Comitê externo, diverso, transparente e independente fosse criado não para julgar, mas para estabelecer as normas e diretrizes de conteúdo que, seguindo a analogia do Estado-nação, formariam a Constituição, afirma David Morar, pesquisador visitante do Digital Trade and Data Governance Hub da The George Washington University. "O que é mais importante, ter um 'Tribunal Supremo' e as suas decisões, ou escrever a 'Constituição', as regras fundamentais que devem guiar as decisões do tribunal?", questiona o acadêmico, em seu artigo, em que ele aponta os riscos de ter um órgão para julgar sobre direitos fundamentais com base em regras "arbitrárias e em constante mudança".

Morar, assim como muitos estudiosos, se mostra preocupado com os perigos de usar mecanismos de estados-nação no âmbito privado, porque os deveres de um governo em relação aos seus cidadãos são muito diferentes dos que regem a interação entre uma empresa e seus consumidores.

Os pesquisadores Matthias C. Kettemann e Wolfgang Schulz, em seu estudo sobre os Padrões da Comunidade, explicam que o Facebook está estabelecendo normas privadas para controlar o discurso público, sem se basear, para isso, em legislações nacionais ou internacionais ou acordos internacionais de direitos humanos. De fato, os estudiosos afirmam que essas normas, que são desenvolvidas pelo Product Policy Team [Time de Políticas de Produto], se tornaram, em si mesmas, um dos principais produtos do Facebook.

Ainda de acordo com o artigo, há um debate atualmente em diferentes países sobre a natureza do espaço de comunicação criado por essas plataformas, se é uma esfera pública ou privada. Nos tribunais alemães, por exemplo, é mais comum o entendimento de que essas empresas, provedores de espaços de comunicação privada, têm uma função de "mercado público".

Embora o Comitê tenha sido originalmente pensado como um ímã de legitimidade para o processo de moderação de conteúdo do Facebook, parece ter tido

seus objetivos questionados desde antes do início dos seus trabalhos.

Um grupo de acadêmicos, jornalistas e ativistas anunciaram, em setembro de 2020, a criação do que eles chamam de "The Real Oversight Board". O grupo indicou que monitorará a atividade dos usuários na plataforma e deve pressionar a empresa quando identificar conteúdos que acredita violar os Padrões da Comunidade. Apresentando os motivos para a criação do grupo, seus membros citaram, dentre outros, a demora na criação do comitê e uma urgência naquele momento, meses antes das eleições presidenciais nos EUA em novembro de 2020.

Em janeiro de 2021 o Comitê publicou suas primeiras cinco decisões. Em quatro delas a decisão interna do Facebook de retirar determinados conteúdos da plataforma foi revertida, demonstrando que o Comitê está disposto a exercer sua independência e se opor ao Facebook quando necessário. Ademais, depois de suspender Trump da plataforma por tempo indeterminado em razão dos acontecimentos de 6 de janeiro, o Facebook solicitou ao Comitê que revisasse sua decisão. Ao encaminhar sua mais importante decisão de moderação de conteúdo ao Oversight Board, o Facebook reforçou seu compromisso com este ambicioso projeto.

6. GLOBAL *VERSUS* LOCAL

Outro aspecto importante da moderação de conteúdo em plataformas digitais é o inevitável atrito entre uma tecnologia global e a regulação local. O Facebook nasceu como uma plataforma para conectar alunos no campus de Harvard e logo se espalhou para outras universidades nos EUA. Quando perceberam que esse poderia ser um projeto viável para além do mundo universitário, Zuckerberg e seus colegas investiram na distribuição de seu produto para consumidores nos EUA e outros países.

Hoje com mais de 2 bilhões de usuários, o Facebook é uma empresa global que não encontra paralelos na história da humanidade, atingindo pessoas de todas as nacionalidades, culturas, etnias e línguas. Não é surpreendente, assim, que a empresa ainda esteja aprendendo a governar o espaço que ela mesma idealizou e fundou, modulando suas regras através de um processo de tentativa e erro. E ainda que o Facebook um dia chegue numa cartilha de moderação de conteúdo relativamente estável, é improvável que essa seja uma solução generalizável a ponto de se tornar universal.

Em outras palavras, regras e valores locais sempre vão estar competindo

com as regras e valores do Facebook. O que se caracteriza como liberdade de expressão em um país pode ser caracterizado como conteúdo proibido em outro. Essa distinção é ainda mais acentuada quando consideramos que o Facebook, por ser uma empresa estadunidense, segue em grande parte uma lógica americana de liberdade de expressão, notoriamente mais permissiva do que outras tradições como a europeia – que, por sua vez, é mais liberal que a tradição asiática e assim por diante.

Encontrar um denominador comum entre as diversas tradições de liberdade de expressão - e de outros direitos fundamentais – é uma tarefa árdua, e a estipulação de suas fronteiras pelo Facebook sempre será, até certa medida, arbitrária. De acordo com Monika Bickert, diretora de políticas de moderação de conteúdo no Facebook, as fronteiras da liberdade de expressão são historicamente definidas por leis, regulações de entidades privadas e normas sociais informais. O problema é que nenhuma dessas "fontes" pode ser usada irrestritamente por plataformas digitais globais.

Apostar nessas fontes locais para definir os limites da liberdade de expressão poderia colocar em perigo o projeto global do Facebook. Para preservar o diálogo entre seus usuários ao redor do mundo, o Facebook precisa garantir que todos tenham acesso ao mesmo conteúdo e que eles possam interagir em tempo real. Regulações nacionais que impõem limites à liberdade de expressão em plataformas digitais podem representar limitações a essa troca global. Quando um país define o que pode ou não ser dito no Facebook, a comunicação entre usuários em jurisdições diferentes pode vir a ser prejudicada.

Como afirma Bickert, o Facebook deve considerar leis específicas como um ponto de partida para a estipular as suas próprias regras de moderação de conteúdo, mas regulações nacionais não podem representar a base da governança da liberdade de expressão online. A exceção emerge quando há a possibilidade de determinar o bloqueio geográfico do conteúdo controverso. Por exemplo, quando uma publicação é tida como ilegal na Turquia, o Facebook pode bloqueá-la dentro do território turco enquanto mantém o mesmo conteúdo acessível para usuários fora do país, preservando, assim, o dialógo global. Assim, um conteúdo só seria retirado do Facebook globalmente quando violasse os Padrões da Comunidade, esses sim aplicáveis a todos os usuários.

Dois casos recentes ilustram esse debate. O primeiro é conhecido como *Glawischnig-Piesczek v. Facebook* e foi decidido pela Corte de Justiça da União Europeia (CJEU) em 2019. Eva Glawischnig-Piesczek é uma política austríaca e presidente

do partido *Die Grünen*. Em 2016, um usuário anônimo do Facebook compartilhou um artigo de uma revista que afirmava que o partido de Glawischnig-Piesczek apoiava a manutenção da renda básica para refugiados no país. Ao compartilhar a notícia, o usuário chamou Eva de traidora, corrupta e fascista. Ao ter seu pedido de remoção da postagem negado pelo Facebook, Eva entrou com uma ação e venceu em primeira instância. A empresa se viu obrigada a bloquear o conteúdo dentro do território austríaco.

Insatisfeita com a solução adotada, Eva recorreu à Suprema Corte da Áustria argumentando que a decisão de primeira instância deveria ter aplicação global, sendo insuficiente o bloqueio geográfico. Como as normas da União Europeia estavam em jogo, a Suprema Corte consultou a CJEU. A Corte Europeia decidiu que as normas europeias não obrigavam nem impediam um estado-membro de ordenar que uma postagem considerada ilegal no país fosse deletada globalmente desde que a decisão não violasse o direito internacional, especialmente a Declaração Europeia de Direitos Humanos..

O segundo caso é conhecido como *Ramdev v. Facebook* e foi decidido pela High Court of Delhi na Índia em 2019. Ramdev é um guru de yoga e uma figura pública na Índia. Alguns episódios envolvendo sua vida privada foram relatados num livro e Ramdev ajuizou uma ação alegando conteúdo difamatório. Após decisão judicial, o livro precisou ser recolhido e republicado sem os trechos tidos como difamatórios. No entanto, alguns usuários do Facebook compartilharam esses trechos em suas contas, e a empresa se negou a remover as publicações.

Ramdev então entrou com uma segunda ação, agora contra o Facebook, requerendo o bloqueio global das postagens em questão. A empresa apresentou dois principais argumentos defensivos. Em primeiro lugar, disse que o bloqueio global não se coaduna com as leis de outros países nos quais o Facebook opera e, em segundo lugar, afirmou que um bloqueio global geraria incentivos para um "turismo judicial". Ou seja, pessoas de outros países poderiam optar por processar o Facebook na Índia sabendo que o judiciário indiano seria receptivo a esse tipo de bloqueio global. Entretanto, esses argumentos não impediram que a Corte indiana determinasse a suspensão global do conteúdo por considerar que o bloqueio local seria ineficaz.

Por fim, é importante notar que a tensão entre dimensões globais e locais não é exclusiva da área de moderação de conteúdo. Dois casos ajudam a ilustrar esse ponto. O primeiro caso, Equustek I, foi decidido em 2017 pela Suprema Corte do Canadá e versa sobre uma disputa comercial entre duas empresas de tecnolo-

gia, Equusteck e Datalink. A segunda, após o fim de um acordo comercial, violou a propriedade intelectual da primeira. A Equusteck, então, buscou judicialmente a desindexação do website da Datalink da plataforma de buscas Google. A Suprema Corte do Canadá entendeu que a desindexação era devida para cessar a violação à propriedade intelectual da Equusteck e determinou que a Google deveria retirar o website da Datalink de sua plataforma. Embora a Google tenha se posicionado contra a decisão, a Corte decidiu que a remoção deveria ser global e não apenas local.

O segundo caso, Google v. CNIL, foi decidido em 2019 pela Corte de Justiça da União Europeia e versa sobre os limites territoriais do "direito ao esquecimento". O caso surgiu a partir de atritos entre o CNIL, a autoridade de proteção de dados francesa, e a Google. O órgão estatal multou a empresa por não desindexar certos sítios eletrônicos de sua plataforma em escala global. A Google argumentava que a desindexação deveria se limitar ao território da União Europeia e que um eventual precedente de remoção global poderia ser abusado por governos autoritários. A Corte de Justiça concordou com os argumentos da empresa e reconheceu a impossibilidade de impor a lei europeia para além do território europeu, decidindo, assim, que a Google não poderia ser obrigada por autoridades europeias a tomar ações globais de desindexação baseadas no "direito ao esquecimento".

7. INICIATIVAS INTERNACIONAIS

Para além das tensões entre a dimensão global e local da governança das redes sociais, também é importante notar o crescente número de iniciativas internacionais no campo da moderação de conteúdo. Diversos grupos de trabalho e entidades internacionais se reuniram nos últimos anos para discutir quais parâmetros devem guiar o trabalho de plataformas como o Facebook e Twitter.

Um dos primeiros documentos desse gênero é o chamado *Rabat Plan for Action* que acompanha um relatório do Alto Comissário das Nações Unidas para Direitos Humanos de 2013. O texto é fruto de uma série de workshops sobre a problemática da incitação ao ódio nacional, racial e religioso. Embora o plano tenha por objetivo apresentar propostas voltadas à atuação dos Estados nessa área, há algumas propostas direcionadas à mídia e às empresas donas de redes sociais.

Para guiar a atividade de autorregulação dessas empresas, o relatório sugere uma série de medidas a serem implementadas. Dentre elas, defende que é preciso

estar alerta para o perigo da disseminação de estereótipos negativos e evitar menções desnecessárias à raça, religião, gênero ou a outra característica protegida que possam promover intolerância. Também sugere dar espaço aos grupos e comunidades para que tenham a oportunidade de moldar a narrativa sobre suas imagens públicas.

Uma segunda medida notável na área de discurso de ódio se consolidou alguns anos depois. Após o ataque terrorista cometido por um supremacista branco em março de 2019 na cidade de Christchurch, na Nova Zelândia, representantes de governos nacionais e de plataformas digitais se reuniram para elaborar uma carta na qual se comprometiam a adotar novas ações para combater o terrorismo na internet.

O homem que cometeu o ataque e matou mais de 50 pessoas em duas mesquitas, deixando outras 40 feridas, transmitiu ao vivo o massacre através de sua conta no Facebook. A plataforma demorou cerca de uma hora para tirar o vídeo do ar, aparentemente em razão de seu algoritmo não ter identificado aquele material como violento. Embora o Facebook já tivesse experiência com vídeos de terrorismo em terceira pessoa, aquela foi a primeira vez que o atirador filmou a si mesmo, no estilo dos jogos de videogame *"first-person shooters".*

Em maio de 2019, então, liderados pela primeira-ministra neozelandesa Jacinda Ardern, diversos líderes se reuniram em Paris para adotar o chamado *Christchurch Call to Action.* Dentre os compromissos assumidos pelas plataformas digitais estão: (1) adotar medidas transparentes para prevenir o *upload* de materiais de terrorismo e sua disseminação, atuando para remover esse tipo de conteúdo o quanto antes, (2) garantir uma maior transparência na construção de padrões da comunidade e termos de uso, (3) aplicar os padrões da comunidade e os termos de uso de forma consistente com os direitos humanos, (4) implementar medidas para evitar que material terrorista seja transmitido ao vivo, por *livestream* e (6) trabalhar em conjunto com outras entidades (governamentais e não-governamentais) em um esforço robusto e coordenado contra o terrorismo na internet.

Enquanto o *Christchurch Call to Action* requer mais ação para combater o terrorismo e o discurso de ódio, outro documento internacional visa proteger as esferas de liberdade individual no espaço digital. Trata-se do *Manila Principles on Intermediary Liability,* uma lista de seis princípios, elaborada por entidades internacionais, para guiar países em sua atuação legislativa na área da moderação de conteúdo, evitando censura de discurso legal ou mesmo legítimo.

Em primeiro lugar, o documento reforça a noção de que os provedores de serviços na internet (ou intermediários) devem ser protegidos de eventual responsa-

bilização em razão de conteúdo postado por terceiros. Esse primeiro princípio reflete a lógica da *Section* 230 do CDA ao imunizar as plataformas digitais de responsabilidade civil. Em segundo lugar, a remoção ou restrição obrigatória de conteúdo deve ser feita após ordem judicial. Esse é outro marco da governança digital que, dentre outros documentos, é compatível com a lei brasileira sobre o assunto, conhecida como Marco Civil da Internet. Em terceiro lugar, pedidos de restrição de conteúdo devem ser claros e sem ambiguidades, além de respeitarem o devido processo legal.

Em quarto lugar, qualquer legislação que estabeleça parâmetros para a restrição de conteúdo em plataformas digitais deve estar de acordo com testes de necessidade e proporcionalidade, próprios do direito constitucional e do direito internacional público. É possível argumentar que a legislação alemã (NetzDG), abordada no capítulo 4, viola este quarto princípio ao incentivar que plataformas digitais "pequem pelo excesso", removendo mais conteúdo. Em quinto lugar, qualquer legislação que estabeleça parâmetros para a restrição de conteúdo em plataformas digitais deve respeitar o devido processo legal. Assim como ocorre em casos judiciais ou administrativos, o usuário deve ter o direito de se defender e contestar eventuais decisões desfavoráveis. Por fim, em sexto lugar, leis e políticas corporativas na área de moderação de conteúdo devem promover os valores da transparência e *accountability* (prestação de contas).

Outra iniciativa internacional é a da Universidade de Santa Clara que organizou uma conferência sobre o tema e lançou os *Santa Clara Principles on Transparency and Accountability in Content Moderation.* São três ao todo: (1) plataformas devem publicar relatórios com os números de publicações e contas removidas, (2) plataformas devem notificar os usuários cujo conteúdo seja removido ou cuja conta seja suspensa, informando as razões que motivaram a ação, e (3) plataformas devem abrir a possibilidade para que usuários apelem de suas decisões de remoção de conteúdo ou suspensão de conta.

Esses são apenas três exemplos que demonstram como entidades e líderes que atuam no plano internacional estão criando princípios para orientar a moderação de conteúdo em plataformas digitais. Como foi abordado acima, a tensão entre os aspectos global e local da moderação de conteúdo traz consigo a necessidade desse diálogo internacional e a construção conjunta de parâmetros na área. Assim, iniciativas como a *Christchurch Call to Action, Manila Principles e Santa Clara Principles* são essenciais para guiar o debate sobre a moderação de conteúdo online.

Na esteira desses documentos, em fevereiro de 2020 o Facebook publicou o relatório *Charting a Way Forward: Online Content Regulation* no qual identifica os

principais desafios da regulação na área, incluindo, por exemplo, a tensão entre a redução de conteúdos ofensivos e a liberdade de expressão e a imposição de "metas de performance" às plataformas. Ademais, o relatório também elenca o que o Facebook entende por "princípios para futuras regulações", como a proteção da liberdade de expressão, a manutenção da natureza global da Internet, o incentivo à inovação e a aplicação de testes de proporcionalidade e necessidade.

8. REDESENHO DA PLATAFORMA E FOCO NOS GRUPOS

A eleição presidencial dos EUA em 2016 foi um divisor d'águas para a maneira como o Facebook encara suas próprias responsabilidades diante de seus usuários. A eleição de Donald Trump foi marcada por notícias falsas e pela interferência de agentes russos no processo eleitoral, conforme informações coletadas pela investigação do Departamento de Justiça dos EUA. Plataformas foram usadas para amplificar campanhas de desinformação e moldar a opinião de eleitores americanos de acordo com interesses políticos. Nos meses subsequentes à eleição de Trump, as redes sociais sofreram duras críticas por não terem agido a tempo de evitar ou minimizar esses problemas.

No início de 2018, o escândalo envolvendo a empresa britânica de consultoria política Cambridge Analytica desencadeou uma nova onda de escrutínio público. Jornais como o The New York Times e The Guardian revelaram, na época, que a empresa de consultoria política com base em dados, Cambridge Analytica, tinha acessado informações pessoais de milhares de usuários do Facebook, com o objetivo de construir perfis de eleitores e impactar as eleições.

O escândalo impactou particularmente o Facebook, que respondeu implementando uma série de mudanças em seu algoritmo do *news feed* para priorizar conteúdos que, nas palavras da própria empresa, promovessem interações significativas (*meaningful interactions*) entre os usuários. Assim, o Facebook passou a focar em postagens de familiares, amigos e grupos em detrimento de "conteúdos públicos", ou seja, postagens de empresas, instituições governamentais e da imprensa.

A intenção da plataforma é se distanciar de uma imagem de "praça pública" e se aproximar da de uma "rede de comunicações privadas" ou de uma conversa na "sala de casa", como afirmou Zuckerberg em comunicado em março de 2019. Parte central dessa mudança de perspectiva é o foco da empresa nos grupos do Facebook, onde usuários com interesses comuns podem se encontrar para dividir experiências, conversar e estabelecer conexões mais intimistas. O ambiente do

grupo é considerado um espaço de comunicação mais seguro para seus usuários, até porque é possível filtrar membros por afinidade e pertencimento.

Em entrevista ao New York Times, em abril de 2019, Zuckerberg disse que os grupos passariam a ter tratamento prioritário na plataforma e esperava que a mudança tornasse o Facebook mais confiável aos olhos dos usuários, afastando a imagem negativa associada à eleição de 2016 e o caso da Cambridge Analytica. No entanto, o esforço para alterar o desenho da plataforma não é apenas uma tentativa de melhorar a sua reputação, mas está ligado também a uma mudança de comportamento de muitos usuários, que passaram a preferir comunicações mais privadas, de acordo com a reportagem do jornal.

Na mesma entrevista ao NYT, Zuckerberg afirma que as três áreas que mais crescem na comunicação online são: "mensagens privadas, grupos e Stories". Em 2019, mais de 1.4 bilhão de pessoas acessavam um grupo do Facebook todo mês. E cerca de 400 milhões de usuários se consideravam parte de grupos que, para eles, eram "significativos" (*meaningful*) – quatro vezes mais do que o número verificado dois anos antes.

Esse novo desenho da plataforma, entretanto, tem diversas consequências para o campo da moderação de conteúdo. O monitoramento e moderação de conteúdo que antes era realizado pelo Facebook através de seus algoritmos ou de funcionários contratados ou terceirizados, agora, como parte do foco em grupos, foi parcialmente delegado para os administradores e moderadores dessas comunidades. Estes agentes a partir de agora terão um papel cada vez mais importante na plataforma.

Essa transição, no entanto, tem preocupado pesquisadores e especialistas em direitos humanos e desinformação. Segundo eles, grupos secretos ou privados são mais opacos e de difícil escrutínio por observadores externos. Isso pode abrir espaço para o uso dessa ferramenta para disseminação de discurso de ódio, notícias falsas, teorias da conspiração e coordenação de campanhas de assédio e violência.

Um dos casos em que esse tipo de uso ficou evidente foi em um grupo privado no Facebook, exclusivo para policiais da *Border Patrol* (Patrulha da Fronteira), nos Estados Unidos. De acordo com uma investigação de 2019 do veículo americano ProPublica, o grupo fazia piadas com as mortes de imigrantes e publicava conteúdo xenofóbico e sexista. Em outro caso, homens da marinha americana usaram um grupo no Facebook, somente para oficiais, para expor fotos de nudez das suas colegas militares, acompanhadas dos nomes e patentes das mulheres.

De acordo com o jornal americano Washington Post, grupos do Facebook têm

servido para difundir campanhas contra a vacinação e também ajudado supremacistas brancos a organizar marchas, como a de Charlottesville em 2017.

O fato de que grupos secretos sequer aparecerem nas buscas dentro da plataforma e, para comunidades secretas e privadas, é preciso ser aceito como membro para ter acesso às publicações, torna muito complexo um controle externo, até mesmo por parte de pesquisadores. E muitos especialistas alertam que, dentro de um grupo construído por pessoas que pensam de forma parecida, as chances de que um usuário denuncie as publicações de outro são muito menores.

Desde que anunciou que ia priorizar os grupos na plataforma, o Facebook tem implementado uma série de novas políticas e ferramentas de segurança, privacidade e moderação de conteúdo específicas para essas comunidades, para tentar responder a parte dessas críticas.

Por ser ainda muito recente, há poucas pesquisas e textos que se debruçam sobre a moderação dos grupos no Facebook. O que de certa forma é compreensível, mas também um problema. Não há momento mais oportuno para se discutir e pensar essas comunidades do que agora, quando o desenvolvimento das regras para os grupos ainda é incipiente. Ou seja, é um modelo ainda em construção e em intensa atualização.

Assim, é importante se aprofundar nessa nova faceta, não apenas pelo investimento da plataforma nos grupos, mas também pelo crescimento dessas comunidades e como forma de acompanhar essa tendência ampla, que pode afetar outras redes sociais também.

Por isso, o próximo relatório vai explicar, de forma detalhada, como funciona a moderação dentro dos grupos e quais são as novas ferramentas e políticas implementadas pela plataforma.

Por meio de entrevistas individuais com moderadores e administradores, vamos apresentar casos de comunidades que se reúnem em conselhos informais para decidir sobre conteúdo racista e homofóbico, por exemplo, e como essa complexa moderação, que toma tempo e energia, afeta a vida pessoal e profissional desses atores.

Vamos ainda refletir sobre os principais impactos e desafios dessa mudança de rumo para o Facebook que, como "maior censor do mundo", tem o enorme poder de influenciar o debate público e limitar a liberdade de expressão.

SOBRE OS AUTORES

Marina Estarque

Jornalista e pesquisadora, cobre transparência e liberdade de imprensa na América Latina para o Centro Knight de Jornalismo nas Américas, da Universidade do Texas em Austin. Como repórter, trabalhou para veículos como Folha, Estadão, Deutsche Welle, Agência Lupa, Rádio da ONU em Nova York, entre outros. É mestre em edição jornalística pela Universidade da Coruña.

João Victor Archegas

Mestre em direito constitucional comparado pela Universidade de Harvard. Foi Gammon Fellow por mérito acadêmico na Harvard Law School. Ex-aluno do Columbia Summer Program in American Law na Universidade de Leiden. É pesquisador da área de Direito e Tecnologia do Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio).

Celina Bottino

Mestre em direitos humanos pela Universidade de Harvard. Foi pesquisadora da Human Rights Watch em Nova York. Supervisora da Clínica de Direitos Humanos da FGV Direito-Rio. Foi consultora da Clínica de Direitos Humanos de Harvard e pesquisadora do ISER. Diretora de projetos do Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio).

Christian Perrone

Pesquisador Fulbright (Universidade de Georgetown, EUA). Doutorando em Direito Internacional (UERJ); Mestre em Direito Internacional (L.L.M/Universidade de Cambridge, Reino Unido). Ex-Secretário da Comissão Jurídica Interamericana da OEA. Coordenador da área de Direito e Tecnologia no Instituto de Tecnologia e Sociedade do Rio de Janeiro (ITS Rio).

Acesse nossas redes

itsrio.org