RODRIGO VIDAL NITRINI

Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas

Tese de Doutorado

Orientador: Prof. Dr. Conrado Hübner Mendes

UNIVERSIDADE DE SÃO PAULO

FACULDADE DE DIREITO

São Paulo – SP

2020

RODRIGO VIDAL NITRINI

Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas

Tese apresentada à Banca Examinadora do Programa de Pós-Graduação em Direito da Faculdade de Direito da Universidade de São Paulo, como exigência parcial para obtenção do título de Doutor em Direito, na área de concentração Direito do Estado, sob a orientação do Prof. Dr. Conrado Hübner Mendes.

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE DIREITO
São Paulo – SP
2020

Catalogação da Publicação
Serviço de Biblioteca e Documentação
Faculdade de Direito da Universidade de São Paulo

NITRINI, Rodrigo Vidal.

Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas. 187 páginas.

Tese (Doutorado - Programa de Pós-Graduação em Direito do Estado) - Faculdade de Direito, Universidade de São Paulo, 2020.
Orientador: Conrado Hübner Mendes

1. Liberdade de expressão. 2. Direitos fundamentais.
3. Internet. 4. Redes sociais. 5. Constitucionalismo digital.

RODRIGO VIDAL NITRINI

Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas

Tese apresentada à Banca Examinadora do Programa de Pós-Graduação em Direito da Faculdade de Direito da Universidade de São Paulo, como exigência parcial para obtenção do título de Doutor em Direito, na área de concentração Direito do Estado, sob a orientação do Prof. Dr. Conrado Hübner Mendes.

BANCA EXAMINADORA

Presidente:_____

Professor Doutor Conrado Hübner Mendes

1º Examinador

(a):_____

2º Examinador

(a):_____

3º Examinador

(a):_____

4º Examinador

(a):_____

5º Examinador

(a):_____

Agradecimentos

Aos professores Conrado Hübner Mendes e Virgílio Afonso da Silva, sou muito grato pela confiança e pela enriquecedora oportunidade de integrar por um novo período o grupo "constituição, política e instituições", desta vez durante o ciclo do doutorado. A convivência entre pesquisadores comprometidos, em turmas que se renovam, mantém uma singular vivacidade que faz deste grupo um espaço privilegiado de reflexão e produção acadêmica na Faculdade de Direito da USP. Tão ou mais importante que a confecção individual da tese nesse período foram os inúmeros seminários de pesquisa e eventos de debates internacionais ("Dialogues") que marcaram a trajetória. Agradeço também aos professores pelas conversas, incentivos e reflexões ao longo de pouco mais de três anos.

Pelos mesmos motivos, agradeço imensamente aos colegas que integraram o grupo "constituição, política e instituições" no período pelos debates, trocas de ideias e suporte mútuo sempre presente. Faço isso com uma menção especial a Artur Péricles Lima Monteiro (que por tantas vezes se dispôs a debater reflexões sobre a pesquisa comigo, nunca negando um bom café) e a Natalia Langenegger (também muito gentil ao compartilhar sua experiência e perspicácia na área de direitos digitais).

A Dennys Antonialli e a Ronaldo Porto Macedo Júnior, agradeço pela composição da minha banca de qualificação, em um momento no qual minhas ideias ainda eram muito exploratórias. As sugestões e as críticas foram fundamentais para um amadurecimento da pesquisa e de seus argumentos. Dennys, em especial, como uma referência na área de direito & internet, mostrou ser, dentro e fora da banca, uma pessoa que sabe aliar os méritos do seu conhecimento com uma capacidade para fornecer encorajamento, observações e críticas construtivas.

Carlos Eduardo Ramos, colega no programa de pós-graduação, também foi um exemplo de generosidade. Muito me auxiliou com conversas e referências bibliográficas, movido apenas pelo intuito de colaborar e fomentar uma discussão honesta.

Ao professor Daniel Wang, da Escola de Direito da FGV-SP, agradeço não apenas pelas excelentes conversas em torno da regulação da liberdade de expressão pelas grandes redes sociais, mas também pela chance de debater esse tema em uma de suas aulas de graduação, com uma turma engajada e interessada de alunos.

Uma grata e rica surpresa desse período foi a retomada de contato com um antigo amigo de faculdade na PUC-SP, Luiz Fernando Marrey Moncau – que hoje possui as credenciais de um sólido pesquisador com produção dedicada em larga medida ao tema da liberdade de expressão no ambiente digital. Foram várias conversas presenciais e trocas de mensagens, ocasiões nas quais sua generosidade e bom humor apagaram completamente a distância de tempos recentes.

Ao professor Rubens Glezer, da Escola de Direito da FGV-SP, agradeço primeiro pela interlocução qualificada e profissional sobre minhas ideias e algumas versões do texto, mas principalmente pela amizade que vale muito mais do que ouro com essa pessoa sensacional.

Esse ciclo também coincidiu com um período profissional intenso e marcante na Defensoria Pública de São Paulo, onde trabalhei ao lado de amigos colegas. Rafael Strano, Mariana Delchiraro e Julio Grostein são pessoas nada menos que admiráveis que sempre estiveram ao meu lado para os bons e não tão bons momentos, inclusive para compartilhar as angústias e superar os percalços da vida acadêmica na pós-graduação. Glauber Callegari, Alvimar Virgílio de Almeida, Tiago Buosi e Felipe Hotz são amigos cujas cumplicidade e camaradagem foram esteios para levar a vida de um jeito bem melhor nesses últimos anos. E sob a liderança de Davi Depiné e de Juliana Belloque, aprendi com admiração como aliar conhecimento, sentimento, garra e prática em um grupo que foi mais do que a soma de cada um.

Ao meu sobrinho Toti, por ensinar diariamente lições de risos e sorrisos.

A Fred e Vanessa Alvim-Kling, ao Guido e ao Matias, agradeço por serem uma família que a vida deu, o que sempre será fundamental e precioso.

Aos meus pais, Dácio e Tania, retribuo todo o amor e apoio que sempre tive – e agradeço também pelas faíscas da curiosidade, da leitura (e da teimosia) que me deram e que são tão importantes e instigantes para tudo. Vocês é que são meu orgulho.

Esta tese nasceu de uma conversa que tive com a Mariana Beatriz. Curiosamente, assim também começou minha dissertação de mestrado. Só por isso, ela mereceria um lugar de honra nestes agradecimentos. Mas seria pouco: sua presença mais marcante não esteve nas ideias e conversas de jantares, embora tenham sido imprescindíveis. Como verdadeira companheira, ela dividiu angústias (amenizando-as) e celebrou as conquistas de cada etapa (dando-lhes mais sentido). Ao lado dela, é possível ver e sentir mais e

melhor o que a vida tem a oferecer, muito além do mundo das ideias. Não há como deixar de ser grato, por tudo. Por isso, dedico a ela esta tese, com suas delícias, dores, conquistas, imperfeições, e com amor.

Resumo

NITRINI, Rodrigo Vidal. Liberdade de expressão nas redes sociais: o problema jurídico da remoção de conteúdo pelas plataformas. (Doutorado) – Faculdade de Direito da Universidade de São Paulo, São Paulo, 2020.

As grandes redes sociais globais dominam hoje uma parte significativa da infraestrutura da liberdade de expressão na sociedade e constituem um capítulo singular, disruptivo e especialmente importante no processo pelo qual a rede mundial reconfigurou as possibilidades de exercício daquele direito fundamental. As políticas de moderação de conteúdo dessas empresas – ou seja, as regras estabelecidas por esses entes privados, bem como suas decisões, sobre quais tipos de conteúdos são permitidos ou proibidos em seus ambientes – são ainda pouco analisadas ou debatidas. Esse é um problema jurídico singular que não é abordado diretamente pela atual legislação brasileira, embora possua evidentes implicações à liberdade de expressão. O risco de censura privada com alto impacto em debates públicos convive ao mesmo tempo com a necessidade real de abordar conteúdos problemáticos que surgem nesses ambientes virtuais, tais como discursos de ódio e campanhas de desinformação. Este trabalho pretende iluminar como essas políticas de moderação costumam ser implementadas pelas três maiores redes sociais: Facebook, Youtube e Twitter – tanto por meio de seus aspectos operacionais, quanto por uma análise de regras substantivas. Ao final, a tese apresenta argumentos e critérios a partir do marco do constitucionalismo digital para dar respostas conceituais e normativas às perguntas de pesquisa formuladas em torno daquele problema jurídico. Em especial, são apresentadas linhas de atuação ao judiciário brasileiro e também diretrizes que sirvam para uma atualização legislativa do Marco Civil da Internet.

Palavras-chave: Liberdade de expressão. Direitos fundamentais. Internet. Redes sociais. Constitucionalismo digital. Moderação de conteúdo. Marco Civil da Internet.

Abstract

NITRINI, Rodrigo Vidal. Freedom of expression on social media: the legal problem of content removal by platforms. (Doctorate) – Faculty of Law, University of São Paulo, São Paulo, 2020.

The big and global social media platforms currently dominate a significant part of the freedom of expression infrastructure in society and constitute a singular, disruptive and especially important chapter to the process by which the web reconfigured the possibilities of exercising that fundamental right. Content moderation policies of those companies – that is, the rules set by those private entities, as well as their decisions, about what kind of content is allowed or forbidden in their environments – are still not much analyzed or debated. That is a singular legal problem that is yet not directly approached by the current Brazilian legislation, even if it has obvious implications to freedom of expression. The risk of private censorship with high impact on public debate coexists with the real necessity of approaching problematic content that arises in those virtual environments, such as hate speech and disinformation campaigns. This work plans to illuminate how these content moderations policies are usually implemented by the three biggest social media companies: Facebook, Youtube and Twitter – by their operational aspects, as much as by an analysis of substantive rules. Lastly, the thesis presents arguments and criteria spawning from the landmark of digital constitutionalism to provide conceptual and normative answers to the research inquiries formulated around that legal problem. In particular, it presents a framework for the Brazilian judiciary as well as guidelines that could serve for a legislative update of the Marco Civil da Internet.

Keywords: Freedom of expression. Fundamental rights. Internet. Social media. Digital Constitutionalism. Content Moderation. Marco Civil da Internet.

Riassunto

NITRINI, Rodrigo Vidal. Libertà di espressione sui social media: il problema legale dela rimozione di contenuti dalle piattaforme. (Dottorato) - Facoltà di Diritto, Università di São Paulo, São Paulo, 2020.

I grandi social network globali ora dominano una parte significativa dell'infrastruttura della libertà di espressione nella società e costituiscono un capitolo unico, dirompente e particolarmente importante nel processo attraverso il quale la rete globale ha riconfigurato la possibilità di esercitare quel diritto fondamentale. Le politiche di moderazione dei contenuti di queste aziende – in altre parole, le regole stabilite da questi soggetti privati, nonché le loro decisioni, su quali tipi di contenuti sono ammessi o vietati nei loro ambienti - sono ancora poco analizzate o dibattute. Questo è un problema legale unico che non è affrontato direttamente dall'attuale legislazione brasiliana, tuttavia abbia chiare implicazioni per la libertà di espressione. Il rischio di censura privata ad alto impatto nei dibattiti pubblici coesiste allo stesso tempo con la reale necessità di affrontare il contenuto problematico che appare in questi ambienti virtuali, come le campagne di disinformazione e discorsi di odio. Questo lavoro ha lo scopo di fare luce su come queste politiche di moderazione sono solitamente implementate dai tre più grandi social media: Facebook, Youtube e Twitter - sia attraverso i loro aspetti operativi sia attraverso un'analisi delle regole sostanziali. Alla fine, la tesi presenta argomenti e criteri dal quadro del costituzionalismo digitale per fornire risposte concettuali e normative alle domande di ricerca formulate attorno a quel problema legale. In particolare, le linee d'azione sono presentate alla magistratura brasiliana e anche le linee guida che servono per un aggiornamento legislativo di Marco Civil da Internet.

Parole chiave: Libertà di espressione. Diritti fondamentali. Internet. Social media. Costituzionalismo digitale. Moderazione dei contenuti. Marco Civil da Internet.

Sumário

Introdução 11
1. Uma nota introdutória 11
2. A ascensão das grandes plataformas globais de redes sociais: Facebook, Twitter e Youtube 13
3. Problema de pesquisa e objetivos 16
4. Estrutura da tese 22

Capítulo 1 – Liberdade de expressão e internet: a reconfiguração da capacidade de estados nacionais para a regulação de discursos 24
1.A – Cães, gatos e ratos no palco da Cosmópolis 24
1.B – Entre a "velha escola" e "nova escola" de regulação de discursos 34
1.C – Considerações finais do capítulo 40

Capítulo 2 – Como as redes sociais operam a moderação de discursos: entre o permitido, o proibido, o visível e o invisível 42
2.A – Controle prévio à publicação por revisão automatizada de imagens 44
2.B – Análise automatizada de linguagem 51
2.C – Bloqueio geográfico 54
2.D– "Flagging" 58
2.E – Moderadores: a aplicação das regras por revisores humanos 61
2.F – Filtragem algorítmica: entre o visível e o invisível 65
2.G– Considerações finais do capítulo 72

Capítulo 3 – Aspectos substantivos da moderação de conteúdo pelas redes sociais: dos anos iniciais às atuais encruzilhadas valorativas e editoriais 76
3.A – Os anos iniciais: da aplicação de "standards" genéricos à construção de um sistema de regras 76
3.B – A proibição do Facebook a discursos de ódio ("hate speech") 84
3.C – Facebook e a proteção ao debate público: da regra da "figura pública" à regra do "interesse noticioso" ("newsworthy") 97
3.D – A nova governança de discursos como um novo tipo de liberdade editorial 111
3.E – Considerações finais do capítulo 118

Capítulo 4 – Constitucionalismo digital e as perspectivas para políticas de moderação de conteúdo de redes sociais pautadas por direitos fundamentais 120
4.A – Constitucionalismo digital: conjugando os planos nacionais e transnacional a partir da lógica de direitos 120
4.B – Constitucionalismo digital, moderação de conteúdo e a perspectiva transnacional do direito das plataformas 129
4.C – Contextualizando a moderação de conteúdo das redes sociais no direito brasileiro a partir do Marco Civil da Internet e suas regras de responsabilização civil de intermediários 139
4.D – A concorrência entre as decisões autônomas de moderação pelas redes sociais e as decisões judiciais 150
4.E – Constitucionalismo digital, moderação de conteúdo e perspectivas normativas para o judiciário brasileiro 157
4.F – Constitucionalismo digital, moderação de conteúdo e perspectivas normativas para uma atualização do Marco Civil da Internet 161
4.G – Considerações finais do capítulo 171

Considerações finais 173

Bibliografia 175

ANEXO 1 – Imagens de manual de treinamento interno distribuído pelo Facebook a seus moderadores, datado de 2016 184

Introdução

1. Uma nota introdutória

"Hossein Derakhshan entrou na prisão com uma internet – e quando saiu havia outra"[1].

A síntese da frase e da história por detrás dela não poderiam retratar melhor as grandes transformações vistas na última década nos ambientes de discursos públicos na internet. Por isso, a introdução desta tese parafraseia o início de livro recentemente publicado por David Kaye, relator especial da Organizações das Nações Unidas sobre as liberdades de expressão e de opinião.

Derakhshan era considerado o padrinho de blogueiros do Irã ("blogfather"), por ter tido papel de destaque na popularização da cultura de blogs naquele país, incentivando textos em farsi em plataformas como a Blogger. Autor de postagens críticas ao regime governista, chegou a viver em autoexílio no Canadá e Europa. Em 2008, cerca de duas semanas após voltar a seu país, foi preso, acusado de "propaganda contra o sistema islâmico", e condenado a uma sentença de dezenove anos e seis meses de prisão. Ficaria preso até 2014 na prisão Evin, local que abriga dissidentes políticos, jornalistas estrangeiros e condenados por crimes comuns. "Em 2008, o Irã tirou-o de um mundo no qual a internet era relativamente descentralizada, onde blogueiros individuais ainda tinham a capacidade de influenciar o consumo midiático. Em 2014, ele foi solto no mundo das redes sociais"[2].

O período de segregação do encarceramento deu a Derakhshan a possibilidade de contrastar abruptamente muitas das mudanças que ocorreram enquanto ele tinha cumprido sua pena em razão das postagens em seu blog. "Seis anos foi um tempo longo na prisão, mas foi toda uma era online", resumiu. Para ele, a cultura de blogs era construída pelas possibilidades de exploração em aberto a partir de hyperlinks, pelos quais um texto poderia fornecer um caminho ou referência a um outro; o público poderia

[1] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 10.

[2] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 10.

estar lendo um texto e, em meio a ele, perseguir seu interesse por uma referência externa, sem caminhos pré-definidos no ambiente da internet. "O hyperlink representava o espírito aberto e interconectado da rede mundial de computadores (...) era uma maneira de abandonar a centralização – todos os links, linhas e hierarquias – e substituí-la com algo mais distribuído, um sistema de nós e redes ('nodes and networks')". Derakhshan disse ter um público de cerca de vinte mil leitores diários quando foi preso[3].

Quando solto no mundo das redes sociais, ele percebeu que para muitas pessoas o uso da internet para consumo de textos, informações e opiniões começava a se confundir com o uso de redes sociais. "Escrever na internet em si não havia mudado, mas *a leitura* – ou, pelo menos, fazer com que algo fosse lido – tinha sido alterada dramaticamente". Desde seus primeiros dias em liberdade, já tinha ouvido que teria que se valer dessas redes para manter um público relevante. Mas ao postar um link para uma postagem externa feita em seu blog, viu que ele parecia "um anúncio sem graça", que conseguiu "apenas três likes". Ele aprendeu logo que as redes sociais davam melhor visibilidade e apresentação a conteúdos nativos que eram nela postados, em oposição a *hyperlinks* para ambientes externos. O objetivo era manter as pessoas dentro daquela plataforma, pelo maior tempo possível; toda a lógica da cultura dos blogs havia sido abandonada naqueles *aplicativos*. Para ele, passava a vigorar a lógica da corrente ("stream"), que tendia a tornar a internet mais parecida com a televisão – linear, passiva, programada e insular[4].

A nostalgia de Derakshan ao que lhe parecia uma época de ouro dos blogs traz consigo esse tom crítico e ácido sobre o movimento de consolidação das grandes plataformas globais de redes sociais e as mudanças que trouxeram à esfera pública online.

Essas redes seriam grandes beneficiárias de um novo modelo de internet comercial que passou a prevalecer em meados dos anos 2000: a *Web 2.0*, na qual plataformas operam a partir de conteúdos criados por usuários[5]. Com uma cada vez mais

[3] Todas as citações do parágrafo provenientes de: Hossein Derakhshan, "The Web we have to save", artigo publicado por Matter, em 14/07/2015.

[4] Igualmente, todas as citações do parágrafo provenientes de: Hossein Derakhshan, "The Web we have to save", artigo publicado por Matter, em 14/07/2015. Para o autor, "a Corrente significa que você não precisa mais abrir tantas páginas na internet. Você não precisa de tantas abas. Você sequer precisa de um navegador. Você abre o Twitter e o Facebook no seu smartphone e mergulha dentro. A montanha vem até você. Algoritmos selecionaram tudo para você. Conforme o que você ou seus amigos tenham visto ou lido anteriormente, eles predizem o que você provavelmente vai gostar de ver. É muito boa a sensação de não ter que gastar tanto tempo achando coisas interessantes em tantas páginas. Mas estamos perdendo algo? O que estamos dando em troca dessa eficiência?".

[5] O maior símbolo inicial da *Web 2.0* talvez seja a Wikipedia. Além dela e de redes sociais, vale mencionar também diversas outras plataformas que operam a partir de conteúdos de usuários, como as resenhas do

acessível banda larga e a popularização de smartphones, tornou-se mais fácil produzir conteúdos – discursos, de fato – na internet, em regra por meio dos serviços gratuitos de plataformas[6]. Mais e mais pessoas podiam falar, local e globalmente; mas essa facilitação de discursos *online* tornou-se possível *dentro dos ambientes dos novos intermediários, sob seus modelos e sob suas regras.* Por ora, nesta nota introdutória, essa certa nostalgia de Derakhshan aponta também para a rapidez com que as condições para a circulação de discursos públicos na internet podem mudar, sublinhando o alto impacto global nos últimos anos decorrente da emergência de grandes redes sociais.

2. A ascensão das grandes plataformas globais de redes sociais: Facebook, Twitter e Youtube

É muito difícil subestimar o impacto que as grandes redes sociais tiveram para o exercício da liberdade de expressão e das discussões públicas na internet, quando paramos para analisar a última década. O espanto de Derakhshan após seu hiato carcerário foi plenamente justificado.

Tome-se o caso da maior plataforma hoje existente: o Facebook sozinho possui 2,41 bilhões de usuários mensais ativos[7] - número que representa mais de um quarto da

Tripavisor ou do Yelp. A própria Amazon consolidou uma plataforma de varejo online que conta com o importante papel das resenhas dos próprios usuários sobre os produtos à venda. A esse respeito: Jeff Kossef, *The Twenty-Six Words That Created the Internet*, Cornell University Press, 2019, capítulo 6.

[6] Essa gratuidade levanta sérias questões a respeito da privacidade de dados pessoais na era digital, pois o acesso sem custos diretos aos serviços é possível a partir do modelo predominante de exploração comercial da internet, construído sobre um "ecossistema de publicidade digital". Esse sistema beneficiou-se do desenvolvimento de tecnologias de coleta e tratamento de dados de usuários, que permitem uma segmentação cada vez mais refinada dos alvos publicitários. Shoshana Zuboff, por exemplo, defende que os rumos tomados pela indústria tecnológica têm consolidado um "capitalismo de vigilância", voltado a minar dados e informações pessoais de indivíduos como aspecto central de seus modelos de negócios, inclusive para influenciar ou determinar comportamentos futuros das pessoas. O caso "Cambrige Analytica" talvez tenha sido o mais emblemático episódio de vazamento de dados a partir do Facebook, que teriam sido usados em campanhas eleitorais de diversos países, inclusive no referendo do "Brexit" na Inglaterra. Trata-se de questão de extrema importância, mas que foge ao escopo deste trabalho – a esse respeito, ver: Dennys Antonialli, *A arquitetura da Internet e o desafio da tutela do direito à privacidade pelos Estados Nacionais*, tese de doutorado apresentada à Faculdade de Direito da Universidade de São Paulo, 2017, pp. 21-33; Soshana Zuboff, *The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power*, Public Affairs, 2019; Roger McNamee, *Zucked: Waking up to the Facebook catastrophe*, Harper Collins, 2019.

[7] Número divulgado em agosto de 2019. A marca de 1 bilhão de usuários havia sido alcançada em outubro de 2012. Um usuário é computado como ativo quando sua conta é acessada durante o mês, seja na plataforma Facebook ou no aplicativo de mensagens Messenger. De acordo com estimativas da própria empresa, contas duplicadas representam cerca de 6% do total. Adicionalmente, os aplicativos Whatsapp e Messenger possuem cada 1.2 bilhões de usuários mensais, enquanto o Instagram possui 700 milhões - todos de propriedade do Facebook; "Facebook hits 2 billion-user mark, doubling in size since 2012", *Reuters*, reportagem publicada em 27/6/2017; "Mark Zuckerberg: 2 billion users means Facebook´s 'Responsability is expanding'", *Forbes*, reportagem publicada em 27/6/2017.

população mundial. Quando Derakshan foi preso, essa marca girava em torno de 100 milhões de usuários. O volume de conteúdo publicado na e gerenciado pela plataforma não possui precedentes – e, como será visto ao longo da tese, esse volume por si só condiciona diversos aspectos de sua operação e capacidade de regulação de discursos. A escala de publicações das plataformas gigantes importa, por si só, para compreender aspectos centrais da moderação desses mercados de ideias.

E, de fato, são pouquíssimas as plataformas globais de redes sociais que consolidaram um domínio sobre a internet, em curto espaço de tempo, com alto impacto em debates públicos online: Facebook, Twitter (300 milhões de usuários mensais ativos[8]) e Youtube (2 bilhões de usuários mensais ativos, contabilizando apenas pessoas que fazem login durante o uso[9]) – a última de propriedade do Google/Alphabet e todas elas provenientes dos Estados Unidos[10].

Assim, é possível dizer que nenhuma outra plataforma de rede social possui um poder global comparável a qualquer uma daquelas três grandes[11] [12]. Os números de

[8] A partir do segundo quadrimestre de 2019, o Twitter passou a contabilizar usuários ativos *diários* – nesse caso, o número mais recente aponta 139 milhões de usuários – "Twitter Q2 earnings: revenue up 18%, daily active users up 14% to 139 million", reportagem publicada por *Fast Company*, em 26/07/2019.

[9] "Youtube now has 2 billion monthly users, who watch 250 million hours on TV screens daily", reportagem publicada por *Variety*, em 03/05/2019.

[10] Não ignoro que existem diversas "internets", como por exemplo o modelo fechado e altamente controlado pelo estado que é vigente na China. O "super aplicativo" WeChat – que além de ser uma rede social, congrega outras funções como mensageria, transações econômicas e compras de serviços – possui mais de 1 bilhão de usuários ativos mensais, alguns deles em países do sudeste asiático com forte presença de chineses ou descendentes. Este trabalho, porém, tem como escopo "a" internet vigente nas democracias liberais do Ocidente e demais países que mantenham relações assemelhadas de abertura de mercado.

[11] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 16. Sobre alegações de que se tratam de companhias privadas, que devem ter a mais ampla liberdade de atuação, Kaye considera que "isso não se aplica mais ao tipo de plataforma que essas três – Facebook, Youtube e Twitter – se tornaram. Suas decisões não têm implicações apenas para suas marcas perante o mercado. Elas influenciam a esfera pública, as conversas públicas, escolhas democráticas, acesso à informação e a percepção da liberdade de expressão. Elas não podem mais se esconder sob a cortina de competitividade corporativa. Elas devem reconhecer seus papéis inusuais, talvez sem precedentes, como monitores do espaço público" (pp. 51-52).

[12] Aqui também merece menção que a rede social russa VKontakte – conhecida como VK – terminou o ano de 2018 com cerca de 60 milhões de usuários mensais ativos. Principal rede social europeia e líder de mercado na Rússia, ela é a mais popular entre as pessoas nativas na língua russa, incluindo forte presença na Bielorrúsia, Cazaquistão e Azerbaijão. Em 2014, o fundador da empresa vendeu sua participação acionária para empresários tido como aliados do Kremlin sob Vladimir Putin, ressaltando temores de um controle cada vez maior do governo russo sobre as informações e os dados mantidos pela rede social. Essa interação entre plataformas gigantes e governos nacionais será abordada novamente ao longo da tese. Em claro exemplo sobre as condicionantes geopolíticas para operações de grandes redes sociais, em maio de 2017 o governo da Ucrânia baniu a VK naquele país (onde também era líder de mercado), nas áreas sob seu controle, em meio ao conflito com a Rússia que perdura há alguns anos. Ver: "How Putin's cronies seized control of Russia's Facebook", reportage publivada por *The Verge*, em 31/01/2014; "Two important

usuários dessas poucas plataformas que dominam a rede só podem ser comparados à soma de populações nacionais inteiras.

Como irá ficar claro ao longo desta pesquisa, essa comparação com estados nacionais não é apenas quantitativa: essas grandes empresas tornaram-se *instituições de governança de discursos na internet*[13], desenvolvendo regras abrangentes e minuciosas sobre a liberdade de expressão (incluindo questões altamente controversas), além de complexos sistemas feitos para aplica-las por meio das mais diversas tecnologias, sempre em constante evolução. Nesse sentido, essas plataformas desempenham funções – de novas maneiras – que remontam a papéis tradicionalmente sob alçada de leis nacionais e de órgãos governamentais.

Claro que essas plataformas não são e nem se confundem com a internet em si[14]. É possível ter uma vida online e participar de debates e discussões fora delas – embora, como Derakhshan descobriu logo que voltou a viver em liberdade, isso signifique abdicar da presença nos locais onde hoje a maior parte das pessoas está e, logo, onde discussões de impacto ocorrem. Ainda assim, essa dominância global de pouquíssimas empresas coloca um problema-chave sobre o papel que *esses intermediários* exercem na governança de discursos, com seus consequentes impactos a direitos fundamentais.

Ainda nesta seção introdutória, é importante fornecer um conceito do que se chama até aqui de "redes sociais". Claro que há um mercado dinâmico com diversas plataformas e produtos na internet – que além de tudo, podem, cada um deles, mudar rapidamente. Por isso, várias definições são possíveis, mas as características a seguir identificam aquelas que interessam a esta pesquisa.

Redes sociais, no sentido aqui empregado, são plataformas interativas da internet que permitem que usuários montem um *perfil pessoal* e, a partir dele e em seu nome, *gerem conteúdos* (tais como textos, postagens, imagens ou vídeos) que não apenas tornam-se visíveis a terceiros, mas que *sirvam de elo para a formação de conexões interpessoais em rede*. Sob esse aspecto, redes sociais são construídas a partir dos

---

results of Ukraine's ban of VKontakte Russian social network", reportagem publicada por *Euromaidan Press*, em 24/03/2019.

[13] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018).

[14] Muito embora o caso de Myanmar demonstre como uma posição de quase completa dominância de mercado possa levar uma plataforma – no caso, o Facebook – a praticamente se confundir com "a" internet em um país. A esse respeito, ver Capítulo 3-B.

conteúdos gerados por usuários, cujos perfis criam redes de conexão para a exposição e o compartilhamento daqueles materiais. Esses conteúdos possuem um grau considerável de publicidade (seja aberta ao público, seja restrita a perfis autorizados) em oposição ao que seriam conversas privadas. Por fim, redes sociais customizam e personalizam a ordenação e a visibilidade de conteúdos aos usuários por meio de algoritmos, de modo que cada perfil tem uma experiência própria de visualização durante seu uso. Facebook, Twitter e Youtube são as plataformas gigantes e mais conhecidas entre as redes sociais, possuindo em comum essas características[15].

3. Problema de pesquisa e objetivos

Esta tese nasceu de uma *inquietação (perplexidade) inicial* relativa ao fato, então pouco conhecido ou debatido, de que o Facebook deliberadamente derrubava de sua plataforma postagens de usuários. Um dos primeiros casos a vir à tona foi de um professor francês, que abriu um processo em seu país contra a empresa em 2011, baseado no direito à liberdade de expressão, depois de sua conta ter sido cancelada "sem aviso ou justificativa", logo após uma postagem com a imagem do quadro "A origem do mundo", de Gustave Coubert – que retrata uma vagina[16].

No início desta pesquisa, praticamente não havia informações públicas abundantes sobre a implementação de políticas de moderação de conteúdo pelas grandes redes sociais. Prevalecia um senso comum de que as redes sociais eram plataformas que por excelência maximizavam a prerrogativa de cada pessoa publicar livremente, em um ambiente que facilitava a formação de redes interpessoais e de engajamentos interativos. Essas características podem ser reais, mas eram amplificadas de modo desproporcional

[15] Essa definição exclui, por exemplo, o Whatsapp: dedicado a conversas diretas (entre duas pessoas ou mesmo entre grupos mais numerosos), ele não opera sob uma ideia de ampla publicidade (não há perfis públicos sob nomes, por exemplo, ou possibilidade de busca por usuários ou grupos); tampouco ordena a exposição de conteúdos com base em algoritmos, já que o emissor da mensagem decide diretamente quem serão seus destinatários, que receberão os materiais apenas nessa situação. O fato de as mensagens serem criptografadas de ponta a ponta significa de princípio que o Whatsapp sequer pode verificar cada teor, uma condição necessária para realizar uma moderação do conteúdo veiculado.

[16] Em 2015 os termos de uso do Facebook passaram a deixar claro que retratos de nudez em obras de arte eram aceitáveis. Apenas em 2019 o processo chegou ao fim, com um acordo amigável entre as partes, que destinaram valores a uma entidade artística francesa – "Facebook to French court: nude painting did not prompt account's deletion", *The Guardian*, reportagem de 1/2/2018; "Facebook and a French teacher settled their years-long lawsuit over Gustave Coubert's 'L'Origine du Mond'", reportagem publicada por *Artsty.net*, em 02/08/2019.

porque *permanecia oculta a extensão das regras de permissão ou proibição de discursos pelas plataformas, bem como da complexa estrutura criada para aplica-las.*

No caso do Facebook, foi apenas em 2017 que significativas reportagens jogaram luzes sobre como sua política de moderação de conteúdo, para além de ter um impacto significativo para o debate público na internet, era em si mesma construída sobre premissas e conceitos que incorporavam delicadas controvérsias morais, políticas e também jurídicas. Uma reportagem do jornal britânico *The Guardian* forneceu um panorama inicial sobre esse sistema e sua abrangência, revelando como essas regras compunham uma política corporativa ambiciosa[17].

A empresa chamava para si a responsabilidade de criação de um conjunto global de regras, que seria aplicado a seus usuários em todos os países – com toda a complexidade e diversidade de contextos culturais que isso implica. Essas regras eram consideravelmente específicas para um sem número de situações, traçando conceitos e premissas que, de um jeito ou de outro, teriam significativo impacto para a liberdade de expressão de seus usuários. Também tratavam de assuntos altamente propensos a dissensos morais ou política e legalmente sensíveis, tais como: opiniões revisionistas de negação do Holocausto, controle sobre casos de pornografia de vingança ou de *cyberbullying* contra crianças, regras de distinção entre discursos de ódio e debates aceitáveis, além de normas sobre postagens que incluíssem sexo, terrorismo ou violência. Entre essas regras, constavam, por exemplo, ainda segundo as revelações da reportagem do *The Guardian*:

> - Opiniões como '*Alguém deve atirar em Trump'* devem ser deletadas, porque como chefe de estado ele está em uma categoria protegida de pessoas. Mas pode ser permitido dizer '*Para quebrar o pescoço de uma vadia, assegure-se de aplicar toda a pressão no meio de sua garganta'*, ou '*vá se foder e morra'*, porque essas últimas frases não eram avaliadas como "ameaças críveis";
> - Vídeos de mortes violentas, ainda que categorizadas como perturbadoras, nem sempre precisam ser deletadas, porque poderiam ajudar a chamar atenção para debates para questões como saúde mental;
> - Fotos de abuso de animais podem ser compartilhadas, sendo deletadas apenas as imagens extremamente perturbadoras;

[17] "Revealed: Facebook´s internal rulebook on sex, terrorism and violence", reportagem publicada pelo jornal *The Guardian* em 21/5/2017. Ver ainda: "Social Media´s Silent Filter", reportagem publicada pelo site *The Atlantic* em 8/3/2017.

> - Transmissões ao vivo de tentativas de autolesões são permitidas, porque o Facebook quer evitar censurar ou punir pessoas em situações de alto *stress*;
> - Qualquer pessoa com mais de cem mil seguidores em uma plataforma de rede social seria considerada uma pessoa pública, o que significa que não têm o mesmo nível de proteção dado às pessoas em geral[18].

Para a publicação britânica, "por meio de milhares de slides e imagens, o Facebook define regras que podem preocupar críticos que dizem que a plataforma é agora um '*publisher*' e que deve fazer mais para remover conteúdos odiosos, lesivos e violentos. No entanto, essas regras podem também alarmar defensores da liberdade de expressão preocupados com o papel *de facto* do Facebook como o maior censor do mundo. Ambos os lados devem provavelmente demandar uma maior transparência"[19].

A revelação para o grande público desses episódios tornava evidente que a política de moderação de conteúdo do Facebook buscava nada menos do que traçar regras – de aplicação global, é importante repisar – para identificação de limites ao "legítimo" exercício da liberdade de expressão em sua plataforma, traçando linhas práticas, a serem seguidas por seus funcionários, entre quais conteúdos seriam permissíveis e quais seriam proibidos. Mais do que isso, ao chamar para si essa – que se tornou uma inevitável – tarefa, uma empresa construiria uma política de uso voltada a bilhões de usuários sobre alguns dos temas políticos e morais mais controversos, quando não insolúveis, a respeito da liberdade de expressão. A empresa via-se sob a necessidade prática e comercial de determinar o que diferenciava "discurso de ódio" de "legítima opinião política", além de ter de tomar posições em questões desprovidas de consenso, tal como se uma pessoa possui o direito de negar a ocorrência do Holocausto, por exemplo. Temas espinhosos que há décadas instigavam reflexões jurídicas e filosóficas sobre a liberdade de expressão passavam a ter respostas normativas adjudicadas globalmente, por meio de um sistema sempre em evolução e sujeito a revisões, bem como a inescapáveis erros.

Como pano de fundo desse cenário, era apenas natural que fossem levantadas indagações sobre o compreensível receio de censura e de remoções injustificadas de postagens. No caso do Facebook – e também do Twitter e do Youtube – a política de uso

[18] "Revealed: Facebook´s internal rulebook on sex, terrorism and violence", reportagem publicada pelo jornal *The Guardian* em 21/5/2017.

[19] "Revealed: Facebook´s internal rulebook on sex, terrorism and violence", reportagem publicada pelo jornal *The Guardian* em 21/5/2017.

de cada empresa, com suas regras internas e não públicas, possuía peso relevante para debates públicos, um peso potencialmente mais importante do que de decisões judiciais tradicionais, por exemplo.

Claro que a moderação de conteúdo em plataformas já era uma realidade na internet em geral. Qualquer *grupo de discussão* ou mesmo *sala de chat* necessitava de moderadores para manter tais ambientes funcionais. Se há um grupo de discussão com algumas dezenas de pessoas dedicadas a debater determinado assunto (cervejas, por exemplo), torna-se bastante razoável que alguma moderação garanta que os debates se mantenham dentro de seu respectivo tópico (vedando que as discussões enveredassem para o tema de cafés, por exemplo). Aqui, vale a regra de que quanto mais restrito um determinado grupo, mais natural que haja limitações e restrições por suas regras (tal como acontece em um pequeno clube privado no mundo real)[20]. Redes sociais, contudo, não se propõem a ser pequenos grupos restritos. Vistas em retrospectiva, foi tão somente natural que suas atividades de moderação de conteúdo se tornassem algo muito diferente do que faziam anteriormente os ambientes com dezenas, centenas ou algumas milhares de pessoas.

Soma-se àquela perplexidade inicial o fato de que, no âmbito da legislação brasileira, *o Marco Civil da Internet (lei federal n. 12.965/2014) dedica praticamente toda sua Seção III a regras que buscam evitar a derrubada de conteúdo postado por usuários nas plataformas da internet*, por meio de um regime de responsabilidade civil de quase imunidade aos provedores de aplicações. Com forte correspondência (embora não total) com o modelo norte-americano[21] e os objetivos declarados de "*assegurar a liberdade de expressão e impedir a censura*", a legislação brasileira determina que a responsabilidade civil de provedores de aplicações em razão de conteúdos gerados por terceiros ocorrerá "somente após ordem judicial específica" de remoção de conteúdo.

Ou seja: *o Marco Civil da Internet dispõe um regime jurídico que parece enfatizar a hipótese de que a retirada de conteúdos de plataformas será feita mediante*

[20] "Entidades *online*, de velhos '*messages boards'* dos anos 1980 e 90 até *blogs* e veículos tradicionais de mídia gerenciando suas áreas de comentários, sempre atuaram como guardiões ('gatekeepers') de conteúdo. As plataformas gigantes de hoje levam isso vários níveis à frente: elas se tornaram instituições de governança, completas com regras gerais e estruturas burocráticas de aplicação. E elas lutam para descobrir como policiar o conteúdo na escala que tomaram" – David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 16.

[21] Seção 230 do "Communications Decency Act*"*, que será abordada no Capítulo 4-C.

ordem judicial, já que silencia completamente com relação à possibilidade de retirada de conteúdo por decisão das próprias plataformas (provedores de aplicações)[22]. À época da promulgação da legislação, sequer havia informações públicas disponíveis sobre a extensão da moderação de conteúdo realizada pelas grandes redes sociais.

Ainda no cenário brasileiro, *a possibilidade de moderação de conteúdo pelas redes sociais foi objeto direto de controvérsias judiciais*. O Ministério Público Federal recentemente tomou medidas práticas para contestar juridicamente a possibilidade de o Facebook remover conteúdos por conta própria, sem que tenha havido prévio pedido de terceiro nesse sentido[23]. O tema também perpassa dois recursos extraordinários – ambos com repercussão geral reconhecida e ainda pendentes de julgamento pelo Supremo Tribunal Federal, que irá se manifestar a respeito da constitucionalidade do art. 19 do Marco Civil da Internet[24].

No nível constitucional, esse problema evoca diretamente o tema da eficácia horizontal de direitos fundamentais. Como compreender e conceituar essa relação entre a autonomia privada das redes sociais para criarem as regras de seus ambientes e o direito de liberdade de expressão de seus usuários? Na medida em que a infraestrutura da liberdade de expressão na sociedade concentra-se cada vez mais nas mãos de atores privados transnacionais – o que implica uma relativa diminuição da capacidade de estados para a regulação de discursos (Capítulo 1) – surge naturalmente a questão sobre quais parâmetros normativos devem nortear essas condutas à luz da liberdade de expressão.

Diante disso, o *primeiro objetivo desta tese* é fornecer, a partir do trabalho de pesquisa realizado, uma *descrição acurada sobre a governança privada de discursos realizada pelas grandes plataformas globais de redes sociais* – Facebook, Twitter e Youtube –, com *enfoque em suas políticas de moderação de conteúdo e decisões a respeito do que é permitido ou não ficar no ar*[25].

---

[22] A Seção III do Marco Civil da Internet também será abordada de forma mais detida no Capítulo 4-C. Para um breve relato sobre debates travados em torno do que se tornaria o art. 19 da lei, quando de sua tramitação legislativa: Francisco Carvalho de Brito Cruz, *Direito, democracia e cultura digital: a experiência de elaboração legislativa do Marco Civil da Internet*, dissertação de mestrado apresentada à Faculdade de Direito da Universidade de São Paulo, 2015, especialmente pp. 99-105.

[23] "Secretário de Direitos Humanos da PGR, Aílton Benedito quer impedir Facebook de banir mensagens de ódio", reportagem publicada pelo jornal *O Globo*, em 24/11/2019. Sobre a ação, ver Capítulo 4-C.

[24] Recursos extraordinários nº 1.057.258/MG e nº 1.037/396/SP; ver Capítulo 4-C.

[25] Logicamente, qualquer pesquisa que aborde plataformas de tecnologia está fadada a ficar em breve desatualizada. Não apenas o cenário de hoje pode mudar em pouco tempo – e essas plataformas perderem

A partir desse objetivo descritivo, pretendo apresentar argumentos conceituais e normativos que enfrentem o *problema* colocado – qual seja, *a remoção de conteúdo por decisão das próprias plataformas de redes sociais*. Torna-se relevante responder às seguintes perguntas de pesquisa, à luz da liberdade de expressão:

a) As redes sociais podem retirar do ar conteúdos de usuários por decisão própria? Se sim, essa retirada deve corresponder a critérios de ilicitude do conteúdo? Como articular parâmetros normativos possíveis que garantam o resguardo de direitos fundamentais, especialmente a liberdade de expressão?

b) Quais papéis cabem ao direito – e, especialmente, ao direito brasileiro, por meio da tutela judicial de direitos fundamentais e de eventual atualização legislativa – para lidar com esse tema?

A pesquisa mira as três plataformas gigantes e globais: Facebook, Twitter e Youtube, tal como descrito no item anterior. O fato de elas dominarem o mercado global de redes sociais significa que possuem as seguintes características que qualificam o objeto de pesquisa desta tese: a) essas empresas atuam de modo *transnacional*, o que levanta questões próprias a respeito de dinâmicas com relação a várias ordens jurídicas nacionais, incluindo diversos padrões culturais; b) o *volume de publicações gerenciado por elas é especialmente massivo e ocorre nas mais diversas línguas e contextos culturais*, o que por si só gera particularidades na tarefa de moderação de conteúdo, conforme será explicitado; c) para fazer frente a essa realidade, as plataformas tiveram que desenvolver *regras e procedimentos altamente institucionalizados para realizar essa governança de conteúdo*.

seus papéis de proeminência, por exemplo –, mas também elas podem ser radicalmente alteradas, quem sabe até mesmo ao prazo de defesa da tese. O Facebook, por exemplo, pode mudar sua estrutura de funcionamento, privilegiando interações entre grupos fechados, ao invés de basear sua experiência em uma "corrente de notícias" ("*News Feed*"). Por isso, esse objetivo inicial e descritivo cumpre um papel de iluminar esse *novo fenômeno qualitativo de governança da liberdade de expressão por atores transnacionais privados*, por vezes em franca concorrência com os direitos de estados nacionais.

Em resumo, como atores privados transnacionais, essas três empresas se destacam e justificam o enfoque do trabalho sobre elas. Ao longo da tese, haverá uma priorização de análise de casos do Facebook, pois além de ser a plataforma mais utilizada (e, por isso, *mais relevante),* é também a que mais disponibilizou e sistematizou informações públicas sobre sua governança de conteúdo[26]. Ainda assim, ao longo do trabalho, em diversos momentos serão mencionados casos e regras também do Youtube e do Twitter, conforme a proposta e seções de cada capítulo.

4. Estrutura da tese

O primeiro capítulo se dedica a uma análise sobre como a internet, a partir de sua arquitetura global, reconfigurou as capacidades de estados nacionais para a regulação de discursos. Valendo-se principalmente das ideias de Timothy Garton Ash a respeito da *Cosmópolis* e de Jack Balkin sobre a estrutura triangular de liberdade de expressão promovida pela rede mundial, será apresentado um argumento pela perda da capacidade relativa de estados nacionais para a regulação de discursos, diante da emergência de novos polos reguladores que são privados e transnacionais.

O segundo capítulo se afasta do foco sobre os estados nacionais e se dedica a analisar como as grandes redes sociais operam suas políticas de moderação de discursos, destacando as principais capacidades tecnológicas pelas quais essas políticas são implementadas, tais como: identificação automatizada de imagens, atuação de moderadores humanos, sistemas de "flagging", filtragem algorítmica, entre outros. Essa avaliação permitirá concluir que as redes sociais não são ambientes neutros voltados à publicação de usuários, além de explicitar os desafios de escala que qualificam essas políticas de moderação.

O terceiro capítulo complementa o enfoque operacional apresentado anteriormente com a apresentação de regras e aspectos substantivos da moderação de conteúdo pelas redes sociais. Inicialmente, será feito um relato sobre como Facebook, Twitter e Youtube desenvolveram seus sistemas de regras de moderação nos anos iniciais. Em seguida, serão apresentadas as regras do Facebook para os temas de "discurso de

[26] Além de estar avançado com formas institucionais inovadoras nessa sua governança, como demonstra a instituição de seu "Conselho Supervisor", inicialmente chamado de "Suprema Corte"; ver Capítulo 4-B.

ódio" e de "interesse noticioso". Ao final, será apresentado um argumento a favor do reconhecimento de uma nova espécie de liberdade editorial para as grandes redes sociais.

O quarto capítulo busca responder às perguntas de pesquisa apresentadas nesta introdução, que decorrem da indagação geral sobre como o direito pode responder à emergência desses novos "governantes de discursos". A partir do marco teórico do constitucionalismo digital, serão apresentados argumentos normativos que lidem com os dilemas que surgem das políticas de moderação de conteúdos pelas grandes redes sociais. Esses argumentos serão apresentados, em momentos distintos, para a seara do "direito das plataformas", no plano transnacional, e para o direito brasileiro, apontando critérios de atuação ao poder judiciário e de atualização das regras legislativas, partir dos dispositivos vigentes do Marco Civil da Internet. Essa visão constitucionalista busca conjugar ambos esses planos a partir de uma lógica de proteção a direitos e de limitação de poderes.

Ao final, complementando os argumentos dos capítulos anteriores, serão apresentadas conclusões gerais e adicionais do trabalho.

Capítulo 1 – Liberdade de expressão e internet: a reconfiguração da capacidade de estados nacionais para a regulação de discursos

Este primeiro capítulo apresenta um argumento inicial que servirá como ponto de partida para o desenvolvimento da tese. De certo modo, parte de um ponto de vista mais tradicional e familiar ao direito constitucional – uma perspectiva que remete à centralidade histórica de estados nacionais – para caminhar, então, em direção ao tema mais amplo no qual se insere o problema de pesquisa: a governança privada de discursos pelas grandes plataformas de internet.

Nesse sentido, serão apresentadas em linhas gerais as principais maneiras pelas quais a internet, cada vez mais organizada em torno de grandes intermediários, abalou aquela centralidade de estados na seara da regulação de discursos. O argumento principal a ser consolidado é o de que *o desenvolvimento da arquitetura da internet implicou uma diminuição da capacidade relativa por parte de estados para a regulação de discursos*.

Isso não quer dizer que estados não possuem uma capacidade *muito relevante* de atuação nessa seara, exercida por atos de governo, leis e decisões judiciais, por exemplo. Mas será apresentado um retrato sobre *como a internet criou um ambiente mais complexo e multifacetado no qual a regulação de discursos por parte dos estados nacionais coexiste – e muitas vezes compete – com a regulação de grandes atores privados e transnacionais.*

1.A – Cães, gatos e ratos no palco da Cosmópolis

A liberdade de expressão costuma ser compreendida como um direito fundamental liberal clássico, historicamente articulado como uma salvaguarda em face do estado. Suas origens remontam às do próprio constitucionalismo, porque essa liberdade buscava impor limites à então *forte capacidade* que estados possuíam de permitir ou proibir discursos. Por isso, as principais teorias em torno da liberdade de expressão costumam formular justificativas ou concepções desse direito em torno de teorias ou argumentações democráticas, normalmente a partir do paradigma de estados nacionais.

A defesa de uma robusta liberdade de expressão como direito individual feita por Ronald Dworkin decorre de sua teoria democrática – pela qual o estado deve tratar cada cidadão como dotado de dignidade e respeito próprios, argumento igualitário que legitima a imposição da lei nacional a cada um, mesmo quando vencido no debate

democrático, exatamente a partir do direito que se tem de dela discordar publicamente e advogar por sua mudança[27]. Outra forte corrente de autores sublinham a importância da liberdade de expressão para que a democracia conte com um vigoroso debate público que viabilize o autogoverno democrático[28]. Isso não significa que teóricos da liberdade de expressão guardem entre si um alto grau de concordância, muito pelo contrário: visões de democracia e de teorias a ela subjacentes comportam inúmeras vertentes, das mais libertárias e avessas a regulações por parte de governos[29] a aquelas que demandam de estados uma postura de maior intervenção visando a construção de uma esfera pública mais equilibrada[30].

Ainda assim, o que essas referências gerais sublinham neste momento é a centralidade que o paradigma de estado nacional tradicionalmente possui nas reflexões em torno da liberdade de expressão. Entretanto, talvez nenhum outro direito fundamental tenha tido seu exercício tão rapidamente modificado pela internet, por conta de sua arquitetura global e de seu desenvolvimento comercial, quanto a liberdade de expressão.

Fronteiras nacionais passaram a importar *menos do que antes* na observância das regras de divulgação e propagação de discursos. Novos obstáculos e limitações se colocaram diante da *pretensão de efetividade* de regras jurídicas nacionais. Foram criadas, sobretudo, condições que facilitam a transposição de discursos para amplos e diversificados contextos culturais e sociais, uma *globalização específica das condições do discurso livre.*

Timothy Garton Ash diz que a era do discurso digital nos levou a viver em uma "Cosmópolis" – uma espécie de "cidade global", onde as pessoas são "vizinhos digitais":

---

[27] Ronald Dworkin, *O direito da liberdade: a leitura moral da Constituição norte-americana*. Martins Fontes, 2006, pp. 311-343.

[28] Alguns exemplos: Meiklejohn e Sunstein identificam a importância do ideal de autogoverno democrático com um processo de deliberação pública que possibilite decisões políticas bem informadas e bem discutidas. Robert Post, por outro lado, vê como traço identificador do autogoverno democrático a possibilidade de participar da esfera pública e influenciá-la livremente, sob o valor da autonomia individual, sem ênfase em uma boa condução deliberativa – Alexander Meiklejohn, *Political Freedom: the constitutional powers of the people*. Greenwood Press, 1979; Cass Sunstein, *Democracy and the problem of Free Speech*. The Free Press, 1995; Robert Post, "Participatory Democracy and Free Speech", *Virginia Law Review Volume 97, n. 3* (2011), pp. 477-489.

[29] Martin Redish, *The Adversary First Amendment: free expression and the foundations of American Democracy*. Stanford Law Books, 2013.

[30]Owen Fiss, *A ironia da Liberdade de Expressão*. Editora Renovar, 2005.

> "*Cosmópolis é o contexto transformado para qualquer discussão sobre a liberdade de expressão ('free speech') em nossa época. A Cosmópolis existe na interconexão dos mundos físico e virtual e por isso, para emprestar uma frase de James Joyce em* Finnegans Wake*, é 'urbana e globall*[31]"".

Um primeiro aspecto dessa vivência na Cosmópolis – entrelaçada "entre os mundos físico e virtual" – é bastante aparente: a maior dificuldade que estados nacionais possuem de fazer valer suas antigas ferramentas para regular ou controlar discursos dentro de seus territórios. Em alguns casos, restrições há poucos anos extremamente poderosas se tornaram quase que inócuas. Como, por exemplo, as proibições a livros.

Veja a recente proibição judicial à venda do livro "Minha luta", de Adolf Hitler, determinada em 2016 por um juiz criminal de primeira instância no Rio de Janeiro. O caso gerou razoável repercussão na imprensa[32] e o teor da decisão foi objeto de atenção de dedicada análise acadêmica[33]. Afinal, a proibição pelo estado – por órgãos administrativos ou por decisões judiciais – à venda de livros, exibição de filmes ou peças de teatros constitui um conjunto de problemas clássicos frente à liberdade de expressão. Mas para além dos argumentos jurídicos que tradicionalmente emergem, *pouca atenção se dá à menor efetividade, nos dias de hoje, de uma ordem judicial que determina uma proibição do tipo*.

Na pratica, uma proibição do tipo pode ser inócua. O libelo autobiográfico do abjeto líder nazista pode ser encomendado pela internet a uma livraria em outro país, no qual sua venda seja permitida, com pagamento por cartão de crédito: uma opção para pessoas afeitas a livros físicos, que correriam um baixo risco de ter o produto retido pela alfândega. Mais fácil ainda seria baixar um exemplar digital por meio de rápida busca pelo Google, ou eventualmente recebê-lo diretamente de alguém, por e-mail ou Whatsapp.

---

[31] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 19. O termo "urbana e global" é uma tradução livre de "Urbi Et Orbi" – "da cidade [de Roma] ao mundo" – expressão tradicionalmente usada pelo papa católico em solenidades.

[32] "Justiça do Rio proíbe livrarias de venderem livro Minha luta, de Adolf Hitler", reportagem publicada pelo *Consultor Jurídico*, em 03/02/2016; "TJ-RJ proíbe venda e divulgação de 'Mein Kampf', autobiografia de Hitler, reportagem publicada pelo *portal G1*, em 03/02/2016.

[33] Ronaldo Porto Macedo Júnior, "Freedom of Expression: what lessons should we learn from US experience?", *Revista Direito GV Volume 13, n.1, São Paulo (2017)*, pp. 275-276.

Cerca de quinze anos depois do julgamento do caso Ellwanger[34] pelo Supremo Tribunal Federal, textos do autor circulam entre grupos de Whatsapp simpáticos ao nazismo[35], por meio de comunicação protegida por criptografia de ponta a ponta. Para Garton Ash, "[John] Milton, cujo *Aeropagitica* é um grande libelo contra autoridades inglesas serem capazes de restringir qual conteúdo impresso pode ser lido em seus domínios, está comemorando em seu túmulo"[36].

Poderia se argumentar que a natureza física de peças de teatro ou de exposições artísticas as tornam mais suscetíveis a ordens de proibição pelo estado, como no caso ocorrido em meados de 2017, no qual um juiz em Jundiaí proibiu a encenação da uma montagem que retratava Jesus Cristo como uma transexual[37].

Ainda assim, essa maior fragilidade também carrega consigo uma perda de potencial proibitivo pelo estado em torno da definição de padrões aceitáveis de discursos. Materiais semelhantes, seja qual for o exemplo, provavelmente estarão acessíveis com poucos cliques pela rede mundial. Proíbe-se uma encenação, mas uma gravação dessa ou de outra manifestação artística que retrate Jesus como uma transexual vai estar disponível em formato digital. Por isso, para além de um juízo de inconstitucionalidade sobre uma decisão judicial do tipo, importa sublinhar que essas decisões parecem ignorar que o judiciário *não tem mais o poder de outrora para determinar padrões de decência em um mundo tornado Cosmópolis*.

Esses argumentos não implicam a suposição de que antes da internet estados contavam com uma capacidade absoluta para a regulação ou proibição de discursos.

---

[34] Supremo Tribunal Federal, habeas corpus nº 82.424. A decisão manteve a condenação de Siegfried Ellwanger – autor de textos antissemitas e que negavam a existência do Holocausto judaico ("Shoah") – pelo crime de incitação à discriminação previsto pelo art. 20 da lei nº 7.716/89, ao considerar que o antissemitismo é uma forma de racismo e, por isso, seria delito imprescritível em razão do texto constitucional. A decisão é considerada um marco do STF no âmbito da liberdade de expressão, pois a maioria de seu plenário à época entendeu que discursos discriminatórios e racistas não merecem proteção constitucional sob o manto daquele direito fundamental.

[35] "Grupos de mensagens negam Holocausto, louvam de Hitler a Enéas e propagam nazismo", reportagem publicada por *Folha de S. Paulo*, em 24/08/2019.

[36] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 26.

[37] O magistrado Luiz Antônio de Campos Júnior motivou que não olvidava "a liberdade de expressão, em referência no caso específico, à arte, mas o que não pode ser tolerado é o desrespeito a uma crença, a uma religião, enfim, a uma figura venerada no mundo inteiro". A decisão seria revogada em segunda instância apenas cerca de cinco meses depois, depois de câmara cível do TJ-SP entender que houve censura prévia proibida pela constituição. A esse respeito: "Justiça suspende estreia de peça que traz mulher trans como Jesus, em Jundiaí", reportagem de *O Globo*, publicada em 03/10/2017; "Justiça de SP derruba liminar que proibiu peça com Jesus Cristo transexual", reportagem publicada pelo *portal UOL*, em 20/02/2018.

Mesmo em governos autoritários que praticavam censura prévia, era possível circular jornais ou publicações clandestinas – contra a chance de ser descoberto e punido. Regras jurídicas, mesmo se extremamente efetivas, dificilmente são totalmente efetivas.

Mas esses breves exemplos buscam sustentar que a internet implica, de um modo geral, uma significativa *perda de capacidade relativa dos estados para a regulação de discursos*, em oposição a um cenário anterior no qual havia uma expectativa de maior efetividade de suas regras jurídicas em seus próprios territórios.

Garton Ash destaca também que a Cosmópolis gera uma maior proximidade entre tradições e normas muito diferentes sobre a liberdade de expressão. Nela, é fácil que algo que seja publicado em um país seja acessível em outro. Fronteiras nacionais perdem relevância também sob esse aspecto: "se as normas de liberdade de expressão diferem muito entre dois lugares – se, por exemplo, é normal questionar o Islã em um país e isso é inaceitável em outro – então respostas violentas podem ocorrer, em um país ou em ambos[38]". Seu exemplo ilustrativo é o caso do vídeo "Inocência dos muçulmanos" – que, para o autor, marcou a "perda de inocência do Youtube"[39].

Em 2012, foram postadas duas versões de um vídeo ficcional, cada uma com aproximadamente 13 minutos, com os títulos de "A verdadeira vida de Maomé" e "Trailer da vida de Maomé". Era uma produção amadora (no Brasil, seria chamada de "filme B"), na qual atores de Los Angeles encenaram espécies de batalhas no deserto. Posteriormente, foram adicionados diálogos por meio de dublagem, que sedimentavam o sentido do filme: uma biografia do profeta Maomé – cuja representação visual é, por si só, um tabu religioso para grande parte dos muçulmanos. As cenas tornavam clara a orientação do filme sob responsabilidade da produtora "Mídia para Cristo": um grupo de muçulmanos atacando uma mulher cristã; um jovem Maomé sendo seduzido por uma mulher e fazendo referências escatológicas a animais, muçulmanos bradando morte a todos os infiéis e reivindicando suas mulheres como espólios de guerra, o Corão sendo mencionado como legitimação para massacres, entre outras cenas afins.

---

[38] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 19.

[39] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, pp. 62-72. "Innocence of Muslims" foi o título dado ao vídeo em sua repostagem mais popular no Youtube. Todas as informações que constam no relato a seguir constam nesta referência bibliográfica.

Depois de meses sem qualquer repercussão, o vídeo de denúncia de um líder muçulmano egípcio a um programa televisivo *foi postado no próprio Youtube* – e alcançaria mais de dois milhões de visualizações até o final do mês. Daí, foram questões de dias para que eclodissem protestos em frente à embaixada americana no Egito. Ainda que por evidente instrumentalização política, a "Inocência dos Muçulmanos" motivou levantes e protestos que se espalharam pelo mundo – mais de 50 pessoas morreriam, a maioria em países de maioria muçulmana, como Paquistão, Afeganistão e Líbia. Neste último, a onda de protestos foi marcada pela invasão à embaixada americana no país e a morte de seu embaixador. Foram registrados protestos também na Europa, Índia, África e Austrália. "Como vimos, nossa Cosmópolis é definida por sua combinação de mundos virtual e físico, global e local – e aqui estava um outro drama de liberdade de expressão urbana e global"[40].

O governo americano iria acionar formalmente o Google – empresa proprietária do Youtube – para chamar a atenção da empresa à importância do episódio e pedir que fosse verificado se o conteúdo violava seus termos de uso – depois de a então secretária de Estado, Hillary Clinton, classificá-lo de "revoltante e condenável". A empresa negou o pedido do governo, dizendo que já tinha feito essa análise e que os vídeos estavam claramente dentro de suas regras. Em meio à controvérsia, o Facebook também revisou o conteúdo do vídeo e decidiu pela sua manutenção[41].

Comentando esse episódio, Robert Post também ressaltou a realidade complexa que surge da convivência aproximada na Cosmópolis de tradições e culturas diversas: "a tradição americana de privilegiar direitos individuais sobre sensibilidades de grupos sempre teve custos. Enquanto o globo terrestre encolhe, e enquanto encontramos culturas que não compartilham dessa tradição, esses custos vão sem dúvida aumentar"[42].

---

[40] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 66.

[41] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1624-1625.

[42] Em continuação: "(...)Essas são consequências previsíveis dos valores de nossa própria Primeira Emenda. Esses tumultos nos forçam a nos perguntarmos o nível de nosso comprometimento a nossos valores constitucionais. Estamos preparados para apoia-los a despeito do dano que infligem a nossos interesses externos?" – Robert Post, "Free speech in the age of Youtube", artigo publicado na revista *Foreign Policy*, em 17/09/2012.

Esse episódio guarda semelhanças com a crise dos "quadrinhos dinamarqueses", ocorrida em 2005[43]. Mas desta vez havia um novo fator relevante: a publicação causa da controvérsia ocorria em uma plataforma de internet gigante e transnacional. Não apenas um grupo cristão poderia produzir um vídeo tosco e de baixo custo com viés anti-islâmico em Los Angeles, mas ele iria *viralizar através daquela própria plataforma* por meio das republicações feitas por moradores de países árabes. Neste caso, novamente, o discurso que era liberdade de expressão em um país com forte tradição constitucional liberal era a blasfêmia de outro – mas esse discurso agora transitava entre todos esses países por uma mesma plataforma digital neles presente. No caso dos países árabes, o funcionamento do Youtube permitia que suas populações tivessem acesso imediato a um vídeo cuja autorização seria inimaginável dentro de suas fronteiras, caso a decisão sobre a publicação estivesse sob responsabilidade de autoridades locais. As fronteiras nacionais – no caso, dos países árabes – tornavam-se mais porosas a um discurso feito em outro país, que neles seria de outro modo proibido.

Isso não significa que os governos nacionais não poderiam ou não iriam tomar medidas efetivas para fazer valer suas regras nesse episódio. Mas essas ações foram todas tomadas *posteriormente*, depois que o vídeo já havia sido divulgado e motivado adesões a protestos em vários países. Mais relevante ainda: *essas ações unilaterais de estados não esgotavam as esferas decisórias envolvidas, pois as reações autônomas do Youtube também iriam importar – compondo, assim, uma nova e relevante dinâmica normativa entre governos e a plataforma.* Esse aspecto em particular do caso "Inocência dos Muçulmanos" será retomado no Capítulo 2-C.

Entre a expressão feita a partir de um local concreto e particular ("urbi") e sua potencial repercussão global ("orbi"), os estados agora competem com atores privados

[43] Em 2005, o jornal dinamarquês Jyllands-Posten publicou uma série de 12 quadrinhos, a maior parte dos quais retratava Maomé – no mais famoso deles, o profeta carregava uma bomba no lugar de seu turbante. O jornal afirmou que buscava fomentar um debate sobre a liberdade de expressão, visões críticas ao Islã e a prática de autocensura. Apesar de reações de grupos muçulmanos naquele ano, apenas em 2006 os quadrinhos virariam um assunto nos países árabes, após virarem pauta de líderes religiosos. Isso levaria a protestos em todo o mundo, alguns deles violentos e que resultariam em mortes, incluindo ataques a igrejas cristãs e a cidadãos europeus. Durante a controvérsia, jornais de diversos países republicavam os quadrinhos, tanto pela notícia de interesse internacional, quanto por um senso de solidariedade à liberdade de expressão do periódico dinamarquês. Para um relato e uma análise desse caso em especial, ver: Iam Cram, "The Danish cartoons, offensive expression and democratic legitimacy", *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010; pp. 311-330. Cram analisa como a Corte Europeia de Direitos Humanos costuma invocar a doutrina da margem de apreciação em casos que confrontam a liberdade de expressão com regras de sensibilidades religiosas nacionais, dando precedência a restrições previstas por países membros.

transnacionais no exercício de uma mediação reguladora. Regras determinadas por essas grandes empresas importam – e podem importar, a depender do caso e do contexto, mais do que as regras jurídicas nacionais envolvidas. Essa *dinâmica em curso – que pode ser mais colaborativa ou mais competitiva, mas que é permanente* – faz com que estados tenham que desenvolver novas regras, políticas e reações que façam frente às rápidas mudanças que decorrem das condições de discursos na era digital.

Claro que sempre há a possibilidade de um estado banir o acesso a uma determinada plataforma ou serviço, ou até mesmo de romper com o acesso à internet no território – mas essas são medidas drásticas, com alto custo social, econômico ou político e, por isso mesmo, excepcionais[44].

Daí que Garton Ash empresta uma analogia de Jonathan Zittrain[45] para descrever a Cosmópolis como um palco de cães, gatos e ratos:

> *"Governos são os cães, empresas são os gatos e nós somos os ratos. Os maiores gatos são mais poderosos que quase todos os cães, exceto os maiores. O fascínio com o confronto entre o Google e a China, quando em 2010 a empresa retirou seu sistema de busca baseado no país, www.google.cn, mencionando a censura online chinesa e o hackeamento de contas de Gmail, foi pelo fato de ser um dos maiores gatos do mundo encarando um dos maiores cachorros. Ao menos tão comum quanto, porém, é uma colaboração próxima, por vezes secreta, entre governos e provedores de serviços de internet,* publishers *e empresas de dados ativos em seus territórios. Isso é o que eu chamo de "poder ao quadrado", ou P2. Enquanto isso, tanto governos quanto empresas trabalham para influenciar organizações internacionais que definem regras ou padrões técnicos para comunicações globais"*[46].

Dessa descrição também decorre que países possuem relevâncias e pesos variados nessa dinâmica global. Esse é um ponto importante, pois para compreender o desenvolvimento da governança privada de discursos pelas grandes redes sociais, não se

---

[44] Nesse mesmo sentido: David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 33.

[45] Jonathan Zittrain, *The future of the internet: and how to stop it*, Penguin, 2008: "Chapter 8: Strategies for a Generative Future".

[46] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, p. 26.

pode ignorar que todas elas são provenientes dos Estados Unidos, país que concentra grande parte das empresas que dominam a internet.

Essa dominância americana não é fruto do acaso. Desde meados dos anos 1990, o governo do país promoveu o que se provaria uma bem-sucedida política para avançar um ambiente de inovação e crescimento entre empresas de tecnologia voltadas à exploração comercial da rede[47]. Google, Amazon, Apple e Microsoft são gigantes que dominam diversas áreas da vida digital – e "nos últimos quinze anos, três empresas americanas – Youtube, Facebook e Twitter – estabeleceram-se como as plataformas dominantes para compartilhamento global de conteúdos"[48].

Nesse contexto, o direito americano assumiu uma importância fundadora para o desenvolvimento da internet comercial e suas consequências para as formas de exercício de – ou de restrições a – direitos fundamentais no mundo digital, nos mais diversos países[49]. No caso da liberdade expressão, isso significa que as regras legais e a cultura jurídica daquele país moldaram, em larga medida, a governança privada de discursos pelas grandes redes sociais[50]. Se as regras americanas de baixa proteção à privacidade

---

[47] Jack Goldsmith remonta o início dessa agenda aos anos do governo Bill Clinton, que anunciou uma política de "liberdade na internet" *('internet freedom'*) global. A agenda seria articulada em torno de dois princípios: o primeiro, de *não-intervenção estatal na área de comércio eletrônico* ("sua administração se oporia a tributos, impostos alfandegários e outras barreiras comerciais, restrições de telecomunicações, limitações de publicidade e outras formas de regulação para empresas, comunicações ou transações na internet"); o segundo, de *vedação à censura*, que advogava uma abordagem global à liberdade de expressão condizente com a tradição americana de forte proteção a esse valor ("esse princípio começou como um componente para promover o comércio eletrônico. Com o tempo, contudo, desenvolveu-se em uma consideração independente que procurava influenciar estruturas políticas de outros países") – Jack Goldsmith, "The Failure of Internet Freedom", *Knight First Amendment Institute – Columbia University*, artigo publicado em 13/06/2018.

[48] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1603.

[49] Debruçado sobre a tutela do direito à privacidade, por exemplo, Dennys Antonialli observa como "o fato de a internet ter sido originalmente povoada por atores localizados nos Estados Unidos foi determinante para os contornos que o direito à privacidade assumiu na rede". Isso porque, naquele país, prevalece nessa seara uma forte autonomia privada e, em análise comparativa, baixos níveis de proteção jurídica à privacidade. Assim, "do ponto de vista da tutela de direitos fundamentais, isso significa que a arquitetura da Internet abre caminho para que esses atores – eminentemente privados –, refugiados em países que adotem modelos regulatórios que lhes sejam mais convenientes, tais como os Estados Unidos no caso da privacidade, interfiram na eficácia das normas constitucionais de outros países sem que existam mecanismos jurídicos efetivos que permitam equalizar ou obstar essa interferência. (...) É como se a internet estivesse operacionalizando um movimento de ampliação extraterritorial da legislação estadunidense sem precedentes" – Dennys Antonialli, *A arquitetura da Internet e o desafio da tutela do direito à privacidade pelos Estados Nacionais*, tese de doutorado apresentada à Faculdade de Direito da Universidade de São Paulo, 2017, pp. 137-140.

[50] "As oportunidades de livre discurso global oferecidas pelas primeiras décadas da internet também têm muito a ver com a localização e origem americana. Seus fundadores viviam na – e suas operações se beneficiavam da proteção da – jurisdição sistematicamente mais a favor da liberdade de expressão em todo

potencializaram os modelos de negócios da internet, a lei daquele país que garante a modalidade mais forte existente de isenção de responsabilização civil para intermediários digitais por discursos de terceiros – a seção 230 do "Communications Decency Act*"* – também foi crucial para a consolidação da "Web 2.0*"*[51].

Com essa posição de franca dominância americana, seria questão de tempo para que os demais países reagissem para também tentarem moldar a internet à influência de suas regras e valores. A China trilhou um caminho à parte para criar uma rede mais insular e passível de controle direto por seu governo. Mas mesmo no âmbito de países mais abertos à internet global há importantes reações a essa hegemonia americana – e que costumam vir da Europa[52]. Essas reações também influenciam o funcionamento das grandes plataformas globais. Nenhuma delas possui capacidade de determinação absoluta. Acontece, essencialmente, um processo dinâmico de mútuas influências, ora colaborativas, ora competitivas, entre as regras jurídicas de países e também entre as decisões de grandes plataformas, que molda todo o espaço digital:

> "*O cyberspaço não é um estado unitário, separado, com suas próprias leis, cortes ou polícia – mas tampouco é simplesmente uma colcha de retalhos de jurisdições nacionais. É algo no meio disso tudo, com muitas formas de vida vira-latas – uma realidade confusa inadequadamente descrita com rótulos como 'dotada de interessados variados' ("multistakeholder") ou 'a comunidade da internet'*"[53].

---

o mundo" – Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, pp. 22-23.

[51] Para uma descrição sobre essa legislação, ver Capítulo 4-C.

[52] "Não são apenas as nações autoritárias que desafiaram o estilo americano de liberdade na internet. Nos anos recentes, países da União Europeia tem visto a hegemonia de empresas da internet daquele país como nada menos do que uma ameaça ao estilo de vida europeu. Em parte, isso decorre da revelação de que a *NSA* ['National Security Agency'] tem abocanhado quantidades massivas de dados sobre cidadãos europeus que foram inicialmente coletados por empresas americanas. E, em parte, é porque essas empresas americanas possuem um enorme poder para moldar a moral, a política, as notícias, as escolhas de consumidores, entre outras coisas, de maneiras que muitas autoridades europeias simplesmente abominam" – Jack Goldsmith, "The Failure of Internet Freedom", *Knight First Amendment Institute – Columbia University*, artigo publicado em 13/06/2018. Goldsmith menciona os exemplos de multas bilionárias por abuso de poder econômico, ameaças de sanções severas em caso de omissão no combate a discursos de ódio ou propaganda terrorista, o reconhecimento do "direito ao esquecimento" (que permite a remoção de informações pessoais de mecanismos de busca), critérios mais rigorosos de proteção a dados pessoais, além da promulgação do Regulamento Geral sobre Proteção a Dados do direito europeu. Para uma análise global sobre a dominância americana, a reação europeia, o isolacionismo chinês e os pesos de grandes mercados regionais, como Índia, Brasil e Rússia, ver também: Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, pp. 31-47.

[53] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 27. Para uma análise mais abrangente a respeito dos diversos tipos de conflitos que surgem em um mundo

1.B – Entre a "velha escola" e "nova escola" de regulação de discursos

Ainda mantendo uma perspectiva mais tradicional que enfoca as ameaças à liberdade de expressão provenientes de estados, essa nova dinâmica cosmopolita e fragmentada apresentada no tópico anterior não surge sem sua própria sorte de novos riscos e preocupações. O horizonte de uma dinâmica potencialmente competitiva entre regras de estados e de plataformas traz consigo também, ao mesmo tempo, a possibilidade de uma dinâmica colaborativa entre esses atores – o que Garton Ash chama de "poder ao quadrado"[54]. Neste tópico, serão apresentados os riscos que algumas formas de controle governamental sobre plataformas de internet apresentam à liberdade de expressão, a partir dos argumentos de Jack Balkin.

Balkin destaca que a liberdade de expressão sempre requer uma infraestrutura de expressão livre. E ao *se tirar o foco do momento em que ocorre a expressão* para mirar a "infraestrutura tecnológica, econômica e social que suporta e viabiliza essa expressão, nós podemos entender como essa infraestrutura é crucial para as liberdades de expressão e de imprensa[55]". Com isso em mente, Balkin aponta que a democratização da liberdade de expressão oportunizada pela nova estrutura da internet e de suas plataformas (incluindo tablets, smartphones, serviços de hospedagem em nuvens, novos serviços para compartilhamento de conteúdo, como redes sociais, entre outros) não acabou com a existência de intermediários ("gatekeepers") – apenas os transformou[56].

---

digitalmente conectado de modo transnacional, mas ainda dividido em jurisdições nacionais, ver: Bertrand de La Chapelle e Paul Fehlinger, "Jurisdiction on the Internet: from legal arms race to transnational cooperation", *Global Comission on Internet Governance Paper Series nº 28* (2016). Esses autores defendem que deve haver uma maior governança transnacional compartilhada entre diversos atores (estados, sociedade civil, plataformas, entre outros) no *nível do uso da internet (suas aplicações e serviços)*, a exemplo do que ocorre em seu nível técnico de protocolos de acesso, para desacelerar a "corrida armamentista jurídica" que caracteriza os conflitos jurisdicionais em curso.

[54] Ver referência da nota de rodapé nº 46 deste Capítulo.

[55] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), pp. 2301-2302. Ainda: "(...) o *New York Times* do meio do século XX que constava nos casos Sullivan e Pentagon Papers não era simplesmente um conjunto de papéis com tinta. Era o produto cumulativo de editores, repórteres, redações, escritórios, serviços de cabo, prensas, sindicatos, caminhões de entrega e serviços de assinatura; e também dependia de um conjunto maior de negócios, arranjos contratuais, costumes e convenções para produzir 'toda as notícias que cabem na página' ('all the news that's fit fo print')".

[56] "Audiências de massa ainda existem, mas muitos agora também são usuários finais que compartilham e transformam conteúdos; em muitos casos, são criadores de conteúdo. O movimento de 'publishers' para 'plataformas' é ao mesmo tempo um efeito e uma causa da revolução na infraestrutura da expressão livre" – Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2304.

Por isso, *do ponto de vista de governos*, "liberdades que dependem de uma infraestrutura podem ser atacadas ou controladas ao se controlar ou atacar a infraestrutura que lhes dá suporte". Falando de uma perspectiva particularmente americana, Balkin considera que uma marca significativa do começo do século XXI é que a infraestrutura da liberdade de expressão está cada vez mais se fundindo com a infraestrutura da regulação de discursos e a infraestrutura de vigilâncias pública e privada. As tecnologias, instituições associadas e práticas com as quais as pessoas contam para se comunicarem umas com as outras são as mesmas tecnologias, instituições e práticas que governos utilizam para a regulação de discursos e de vigilância[57] - como demonstra a revelação em 2013 de que órgãos de inteligência americana acessavam milhões de dados de usuários de todo o mundo a partir das informações coletadas pelas empresas de tecnologia daquele país[58].

Balkin considera que as dificuldades de governos para controlarem discursos com as antigas ferramentas faz com que eles se voltem à infraestrutura digital para implementar esses objetivos. Uma nova ameaça para a liberdade de expressão resulta do fato de que:

> *"muitas das mesmas características da infraestrutura digital que democratizaram os discursos ao mesmo tempo tornam essa infraestrutura o alvo mais tentador e mais poderoso para se buscar a regulação e o controle de discursos. Embora a infraestrutura digital libere as pessoas ('speakers') da dependência de velhos intermediários ('gatekeepers'), ela o faz através da criação de novos intermediários que oferecem a estados e entes privados novas oportunidades de controle e vigilância (...) [por isso], não surpreende que entes privados e estados procurem formatar essa infraestrutura de um modo que melhor facilite esse controle e vigilância"*[59].

---

[57] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), pp. 2297-2303.

[58] Shoshana Zuboff afirma que o paradigma de extração de dados pessoais que sustenta o modelo de negócios de grandes empresas de internet fomenta também – especialmente nos Estados Unidos – uma "afinidade eletiva" entre essas grandes empresas e órgãos governamentais de segurança interessados em maximizar o monitoramento de pessoas pela rede mundial – Soshana Zuboff, *The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power*, Public Affairs, 2019, em especial pp. 112-121.

[59] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), pp. 2304-2305.

Daí que o autor destaca sua preocupação com o que seria uma "nova escola" de regulação de discursos: "baixa saliência e o uso de entes privados pode ajudar governos a preservar sua legitimidade mesmo quando suas políticas bloqueiam, limitam ou monitoram a expressão livre. Essa é a grande história sobre as liberdades de expressão, imprensa e de associação na era digital"[60]. Assim, do ponto de vista tradicional (estatal), haveria uma "velha escola" e uma "nova escola" de regulação de discursos.

A "velha escola" é normalmente voltada a (i) pessoas, (ii) espaços e (iii) tecnologias pré-digitais de distribuição em massa. Suas táticas incluem a detenção ou deportação de pessoas, o controle de acesso a espaços públicos para protestos ou manifestações, bem como o monopólio, regulação, apreensão e apropriação das estruturas de transmissão de massa, como prensas, torres de transmissão televisivas ou radiofônicas, projetores de vídeos ou mesmo livros[61]. "Nessa concepção tradicional, liberdade de expressão e de imprensa significa simplesmente estar livre de uma regulação da velha escola"[62].

A "nova escola" de regulação, por sua vez, é fruto das técnicas de controle das redes digitais. Ela prioriza a prevenção de publicação (proibição "ex ante") sobre a punição posterior, além de sofisticadas formas de cooperação entre governos e empresas. Utiliza-se ao mesmo tempo de "chicotes e cenouras" e está "profundamente conectada com novas tecnologias de vigilância digital por entes privados e pelo estado"[63].

Balkin menciona como alguns exemplos de "nova regulação" a promulgação em 2017 da legislação alemã conhecida como NetzDG, que prevê a imposição de multas a plataformas que falhem em retirar do ar conteúdo considerado ilegal naquele país[64], assim como o "direito ao esquecimento" reconhecido inicialmente em 2014 pelo Tribunal de Justiça da União Europeia e que mira a exclusão de mecanismos de buscas das

[60] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2298.

[61] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2306.

[62] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), p. 2015.

[63] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2306.

[64] A legislação entrou em vigor no começo de 2018 e é mote de fortes controvérsias ainda em debate. Para uma análise detida a respeito, ver Capítulo 4-F.

informações consideradas não mais relevantes ao interesse público[65]. "Por essa razão, a nova escola (...) afeta a habilidade prática de discursar tanto quanto a velha escola"[66]. Essas regras, apesar de proferidas por estados nacionais, blocos regionais ou organismos internacionais, podem acabar implicando uma reestruturação mais ampla das práticas daqueles atores transnacionais, que a depender do peso ou da influência de tais regras, incorporam esses novos critérios na reformulação de seus parâmetros de atuação em escala mais ampla, regional ou global.

Por isso, Balkin se preocupa com as maneiras pelas quais os estados podem exercer controle ou pressão sobre as plataformas privadas digitais. Uma das principais é a "censura colateral" ("collateral censorship"):

> *"Censura colateral ocorre quando o estado responsabiliza um ente privado A pelo discurso de um entre privado B, e A possui o poder de bloquear, censurar, ou de qualquer outra maneira controlar o acesso ao discurso de B. Isso leva a bloquear o discurso de B ou retirar dele seu suporte de infraestrutura. De fato, como isso não envolve o discurso de A, cria-se incentivos para que A cometa erros em nome da cautela e restrinja também mesmo aqueles discursos completamente protegidos [juridicamente] de modo a evitar qualquer chance de responsabilização"*[67].

O conceito de "censura colateral" evoca diretamente a questão jurídica da responsabilização civil de intermediários digitais pelo conteúdo de autoria de terceiros, tema que será abordado no Capitulo 4-A.

Uma segunda preocupação é que essa "nova escola" implique uma "censura prévia digital", como em caso de filtragem oculta de conteúdos em ambientes digitais,

---

[65] Os debates sobre o direito ao esquecimento não tratam apenas de seus aspectos substanciais, mas também sobre as disputas em torno da abrangência territorial (eventualmente transnacional) que devem ter as decisões que acolhem os pedidos de exclusão de informações, que normalmente envolvem o Google. A esse respeito: Nicolas P. Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 51-53. Em setembro de 2019, o Tribunal de Justiça da União Europeia decidiu que tais exclusões não podem ter abrangência para além da própria União Europeia; "Right to be forgotten privacy rule is limited by Europe´s top court", reportagem publicada pelo *The New York Times*, em 24/09/2019.

[66] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), p. 2016.

[67] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2309. Além de incentivos para que A seja mais restritivo com relação a discursos de B, Balkin também destaca que essas decisões ocorrem sem nenhuma pré-determinação judicial ou mesmo sem nenhuma transparência para com B.

com a retirada de algum material do ar, sem que isso seja feito com transparência para com as pessoas afetadas, além de permanecer oculta para o público em geral, já que esses procedimentos seriam implementados por burocratas das empresas privadas ou por algoritmos que, em alguma medida, servem ao interesse estatal. Além disso, essa dinâmica se desenvolve mediante relações de interesses mútuos ou de "soft power" entre governos e empresas, a exemplo que pode ocorrer também com empresas de mídia tradicionais. Por fim, estados também preservam a possibilidade de criação de regras jurídicas que aumentem a prerrogativa de governos de bloquear conteúdos hospedados no exterior, por exemplo[68].

A existência dessa "nova escola" de regulação não significa o abandono das velhas táticas. Balkin entende que esses dois conjuntos passaram a conviver e a complementar um ao outro[69]. Mas torna-se "muito melhor controlar as redes digitais nos bastidores ou cenários de fundo, para que o controle e a vigilância pareçam indistinguíveis das condições normais, ao invés de uma demonstração de força extraordinária, singular ou intermitente", já que "demonstrações abertas e excessivas de força deslegitimam o próprio estado"[70].

Balkin sublinha corretamente o fato de que intermediários da internet se tornam alvos naturais para projetos de controle ou regulação de discursos por parte de estados. *Mas essas plataformas não são apenas alvos: são também atores.*

---

[68] Foge ao escopo desta seção enumerar todas as formas de controle que Balkin identifica sob a "nova escola" de regulação, até porque elas envolvem searas muito distintas – desde a coleta de dados pessoais e monitoramento por órgãos de segurança nacional, até relações mais sutis de influência política. Para essa análise mais exaustiva do autor, ver: Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), pp. 2318-2329; Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), pp. 2016-2021.

[69] Balkin aponta que, no contexto da perseguição dos órgãos de inteligência americanos e ingleses a Edward Snowden, partiu-se da premissa de que ele e seus colaboradores não usavam as plataformas digitais convencionais porque as consideravam inseguras; por isso, teriam utilizado métodos clássicos para repassar os arquivos secretos (levando-os fisicamente em dispositivos offline); por essa razão, David Miranda, parceiro do jornalista Glenn Greenwald, ficou detido por horas no aeroporto internacional de Heathrow, suspeito de estar cruzando fronteiras na posse física de tais informações – Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2307.

[70] Jack Balkin, "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014), p. 2309. Balkin escreve de uma perspectiva propriamente americana, ou seja, de uma realidade na qual o governo daquele país possui de fato meios mais efetivos de moldar iniciativas – inclusive de controle – que terão alto impacto em todo o mundo. Como já mencionado, as regras daquele país possuem um peso de influência diferenciado no desenvolvimento da internet mundial, até porque são as empresas daquele país que lideram a exploração comercial da rede. No caso dos Estados Unidos, isso gera um ciclo que favorece uma liderança também do sofisticado aparato de segurança e vigilância daquele país. Por isso, embora esse ponto de vista sobre uma maior capacidade de vigilância e de controle por parte de estados nacionais esteja correta, ele possui uma incidência diferenciada a depender de qual país se trata.

E, no caso das grandes redes sociais, são atores globais e multibilionários, que guardam consigo uma alta capacidade de definir regras de discursos e de moldar *esferas corporativas de debates públicos* povoada por centenas de milhões ou por bilhões de pessoas. Daí que para Balkin, hoje:

> *"a liberdade de expressão é um triângulo. O conceito de liberdade de expressão – e os perigos para a expressão livre – que caracterizaram muito dos séculos XIX e XX diziam respeito sobre se estados-nações e suas subdivisões políticas poderia censurar ou regular os discursos de pessoas que viviam dentro de suas fronteiras. Esse retrato ainda descreve muitos importantes problemas de liberdade de expressão, mas é cada vez mais defasado e inadequado para proteger a liberdade de expressão hoje em dia. No começo do século XXI, a liberdade de expressão depende cada vez mais de um terceiro grupo de atores: uma infraestrutura de comunicação digital de propriedade privada composta por empresas que suportam e governam a esfera pública digital que as pessoas usam para se comunicarem"*[71].

Essa nova disposição "triangular" da liberdade de expressão evidencia um novo polo regulador de discursos:

> *"Estados nacionais regulam discursos e mídias de massa tradicionais por meio da velha escola de regulação. Estados nacionais regulam e tentam cooptar e coagir a infraestrutura de internet por meio da nova escola de regulação. E, finalmente, a infraestrutura da internet regula atores privados e mídias tradicionais por meio de técnicas de governança privada. Essa é a nova estrutura da regulação de discursos no começo do século XXI, e debates sobre direitos de liberdade de expressão online devem levar em conta essa estrutura"*[72].

[71] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), p. 2012.

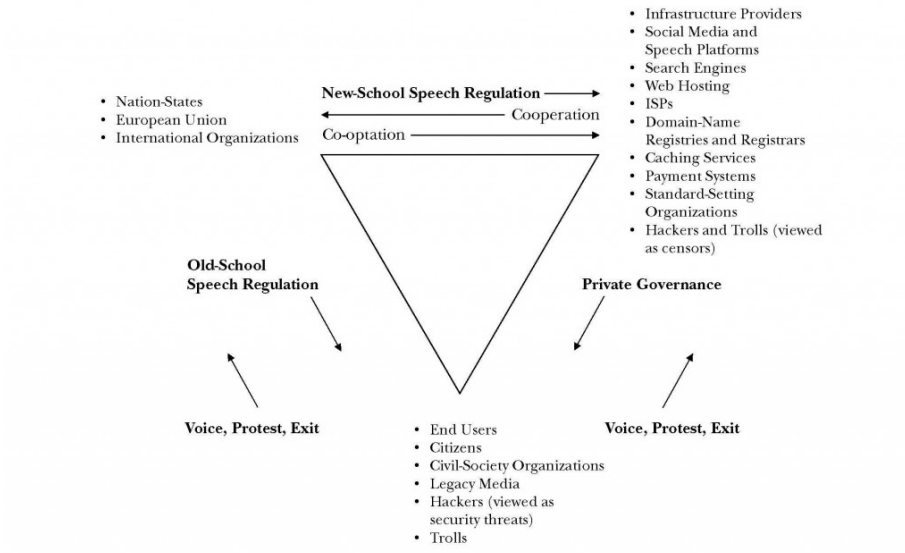
[72] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), pp. 2014-2015. O autor resume esse cenário triangular com o seguinte gráfico (com os termos originais em inglês):

Nesse novo cenário, a concentração da internet comercial na mão de alguns poucos e gigantes intermediários é um fator extremamente relevante:

> "*Nenhum único ator poderia policiar uma rede horizontal. Essa era a beleza da internet quando Derakhshan perdeu sua conexão com ela. A internet de hoje requer atores poderosos e identificáveis que policiam suas próprias plataformas e provêm as chaves para que governos as policiem também*"[73].

1.C – Considerações finais do capítulo

A realidade de hoje traz alguns desafios para as maneiras mais tradicionais de se pensar problemas da liberdade de expressão. Essa erosão da centralidade de estados nacionais significa não apenas a existência de problemas práticos (como uma maior dificuldade para a efetividade de regras estatais), mas também novos desafios sobre a



[73] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 13-14. O autor continua: "Danny O'Brien, o chefe de `advocacy` internacional da Eletronic Frontier Foundation, um dos maiores e mais antigos grupos de advocacy de direitos digitais do Vale do Silício, colocou a questão para mim desta maneira: 'Você precisa que essas empresas existam se pretende controlar discursos'. A velha internet era difícil de policiar. Como Bill Clinton notoriamente disse sobre a esperança da China de controlar a internet, 'Boa sorte! Isso é como tentar pregar geléia na parede'. Mas a internet contemporânea não é nada como uma geléia. Ela facilita controle por empresas e por governos".

legitimidade das práticas da governança privada de discursos na internet e de seus impactos a direitos fundamentais.

No ambiente de uma internet cada vez mais organizada em torno de grandes intermediários, as condições de exercício da liberdade de expressão são em grande medida dadas por atores corporativos, transnacionais e que atravessam e são influenciados por diversas culturas, diferentes padrões políticos e mercados variados[74].

Os argumentos deste capítulo foram desenvolvidos em torno da reconfiguração da capacidade de estados para a regulação de discursos. Pode-se dizer que há uma perda da capacidade relativa de regulação de discursos com parte desses entes, ainda que com as ressalvas e observações feitas, entre elas as possibilidades da "nova escola" de regulação de discursos.

[74] "O Google pode não ser um país, mas é um superpoder. Assim como o Facebook, Twitter e alguns poucos gigantes da indústria de informação. Eles não possuem a autoridade de legislações formais de estados soberanos. E seus líderes não são responsabilizáveis ("accountable") por seus usuários assim como governantes são por seus eleitores (...). Ainda assim, suas capacidades para permitir ou limitar as liberdades de expressão ou de informação é maior do que aquelas da maioria dos estados" – Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, 2016, p. 47-48.

Capítulo 2 – Como as redes sociais operam a moderação de discursos: entre o permitido, o proibido, o visível e o invisível

A reconfiguração da capacidade de estados para a regulação de discursos, descrita no capítulo anterior, traz consigo uma nova perspectiva necessária: o *emergente sistema de governança privada de discursos* merece especial atenção por quaisquer interessados em debates em torno da liberdade de expressão na internet. Se hoje prevalece uma nova disposição "triangular" da liberdade de expressão[75], então torna-se mais do que justificado *compreender e iluminar as maneiras pelas quais agem os atores localizados nesse novo polo regulador de discursos*[76]. Especialmente porque essas novas maneiras correspondem a *novas formas e possibilidades de regulação de discursos, em comparação às ferramentas estatais tradicionais.*

Facebook, Twitter e Youtube são, a rigor, empresas de tecnologia – que apenas em um segundo momento, e de modo praticamente inadvertido ou inesperado, tiveram que fazer frente aos dilemas e complicações decorrentes do gerenciamento de mercados de ideias alimentados por centenas de milhões ou bilhões de usuários. Por isso, este segundo capítulo irá *apresentar as principais capacidades tecnológicas e operacionais que caracterizam a regulação e controle de discursos pelas grandes redes sociais*. Cada tópico irá abordar um tipo de operação técnica relevante para viabilizar a implementação de suas políticas de conteúdo[77], o que, no conjunto, termina por revelar uma sorte de complexidades próprias a esse ramo.

Os argumentos feitos ao final da tese podem ser sustentados apenas com uma compreensão consistente dessas complexidades, referente às *possibilidades* e aos *desafios* que essas empresas de tecnologia possuem ao lidar com uma escala massiva e sem precedente de discursos em seus ambientes. Dito de outra maneira: não é possível enfrentar o problema da remoção de conteúdos pelas grandes redes sociais apenas com análises de suas regras substantivas sobre discursos, pois *tão importante quanto isso é*

[75] Como defende Jack Balkin; Capítulo 1-B.

[76] É certo que as redes sociais não são a internet propriamente dita. Aquele polo regulador envolve diversos outros atores igualmente relevantes: provedores de conexão, redes de segurança para hospedagem de conteúdo, mecanismos de busca, registros e registradores de domínios, entre outros. O problema de pesquisa proposto, entretanto, mira as decisões de derrubada de discursos pelas grandes redes sociais.

[77] O *problema específico* a ser enfrentado– a remoção de conteúdos pelas grandes redes sociais – *insere-se nesse tema mais amplo*, que é a emergente governança privada de discursos por intermediários digitais.

compreender o contexto tecnológico (fático) no qual esses discursos são produzidos e as possibilidades pelas quais aquela moderação pode ocorrer[78].

O desenvolvimento em si dessas tecnologias carrega consigo implicações políticas relevantes para a liberdade de expressão. Lawrence Lessig há anos já destacou como a governança da internet se sustenta sobre *quatro diferentes modais reguladores: o direito, o mercado, as normas sociais e o código*. Na internet, *o código também é uma espécie de lei*. Do mesmo modo que um protesto de rua é configurado pelo espaço físico de praças ou avenidas, *a arquitetura tecnológica* (código) determina as condições de funcionamento da internet; determina o que é possível fazer online, quais são as "leis da física" daquele ambiente. Como códigos nunca são neutros, é necessário estar atento às escolhas e às possibilidades de controle que neles estão embutidas[79]. Daí o enfoque, neste capítulo, sobre as novas *capacidades técnicas* de regulação (e controle) de discursos pelas redes sociais.

Algoritmos são um exemplo claro sobre como códigos importam para a governança de discursos nesses ambientes. Eles não apenas determinam aquilo que terá mais ou menos visibilidade para os usuários; com isso, também criam incentivos para que as pessoas produzam conteúdos e discursos "valorizados" pelo "código".

Por muitos anos, o Twitter foi considerado uma plataforma mais propícia a discursos agressivos e de ódio – de fato, esse problema se mostrava muito maior ali que no Facebook, por exemplo. Um fator colaborativo era o próprio *design* de suas regras de conversações. Enquanto no Facebook uma pessoa tinha controle sobre os comentários de terceiros feitos em sua própria postagem (podendo apaga-los, por exemplo), o Twitter durante quase toda sua existência não permitia que um usuário controlasse ou moderasse as reações de terceiros, que passavam a ser públicas para todos que visualizavam a postagem inicial[80]. Foi apenas em meados de 2019 que a empresa lançou – gradativamente

---

[78] Nesse ponto, este capítulo irá revisar parte significativa de influente pesquisa de Kate Klonick – "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018).

[79] Lawrence Lessig, *Code: version 2.0*, Basic Books, 2006, em especial: pp. 5-7; 122-137; 311 e seguintes.

[80] "Parte do que torna o Twitter tão poderoso é sua habilidade de nivelar o campo da comunicação; tweets de uma pessoa não famosa, ou de usuários com poucos seguidores, são na teoria tão prováveis de vir à tona do que aqueles provenientes de celebridades. O Twitter simplesmente é construído de modo diferente da maior parte das redes sociais. Duas contas não precisam seguir uma à outra para interagir, e uma vez que um tweet está publicado, o tuitador original não possui a habilidade de moderar respostas, como se possui no caso dos comentários do Facebook ou Instagram. Esse design único é responsável por alguns dos momentos mais revolucionários e de felizes acasos ('serenpidity') do Twitter; também é perfeito para abusos" – "A Honeypot for assholes: inside Twitter's 10 years failure to stop harassment", reportagem

– a opção de bloqueio ou de ocultação a comentários pelo autor da postagem original, exatamente para incentivar um ambiente de discurso mais civilizado e evitar que uma única pessoa pudesse "estragar" um tópico de discussão entre várias outras[81].

No caso desta pesquisa, arquiteturas tecnológicas e procedimentais (código) construídas pelas grandes redes sociais importam sobremaneira para a compreensão sobre como ocorre o controle de conteúdo em seus ambientes. *Uma apresentação sob esse aspecto revela, sobretudo, as novas maneiras e possibilidades da regulação de discursos pelas redes sociais, em comparação às formas tradicionais da "velha escola"*. Por isso, este capítulo constrói um retrato geral e atualizado sobre os principais procedimentos técnicos que caracterizam a moderação de conteúdo, enfocando a implementação de regras pelo Facebook, Twitter e Youtube. Ainda assim, em cada tópico serão apresentados problemas e dilemas substantivos que surgem a partir dessas atividades.

2.A – Controle prévio à publicação por revisão automatizada de imagens

O *controle prévio* é certamente o momento mais incisivo e restritivo de se realizar a moderação de conteúdo, pelo fato de que ele remonta à ideia de censura prévia. É um tipo especialmente importante de filtragem de conteúdo exatamente porque constitui uma exceção à regra geral de que as pessoas normalmente podem postar conteúdos em plataformas sem avaliação prévia, já que a maior parte da moderação e de eventual retirada do ar ocorre "a posteriori".

Esse tipo de controle é comum em casos de "uploads" de vídeos, especialmente no Facebook e no Youtube. A avaliação é prévia porque é feita entre o envio do conteúdo e sua efetiva publicação – quando no Facebook, por exemplo, aparece a mensagem: "Processando o Vídeo: o vídeo da sua publicação está sendo processado. Nós iremos enviar uma notificação quando estiver pronto e seu vídeo apto a visualização". Em regra,

publicada pelo *Buzzfeed News*, em 11/08/2016. Ainda assim, o fator predominante para esse problema era a insistência convicta da empresa de não moderar – ou não censurar – as postagens de seus usuários. Com resultado, o Twitter passou a ficar identificado como um ambiente raivoso e cheio de "trolls", no qual pessoas que se engajavam em conversas públicas se viam com frequência diante de mensagens de ódio ou mesmo ameaças concretas. Esse último ponto será retomado no Capítulo 3-A.

[81] "Twitter launches the 'Hide replies' feature, in hopes of civilizing conversations", reportagem publicada por *Techcrunch*, em 17/07/2019.

decorre de um processo automatizado, levado a cabo por filtragem algorítmica, sem revisão humana[82].

Embora as tecnologias aplicadas para a implementação desse controle prévio sejam importantes por si só, é também revelador como o processo de seu desenvolvimento se deu a partir de restrições da legislação americana sobre pornografia infantil e proteção a direitos autorais, em um primeiro momento, para depois se expandir, por decisão das próprias plataformas, à área de combate à propaganda terrorista.

Nos anos 1990, a expansão da internet atraiu consigo, especialmente no debate público e político americano, uma alta preocupação com a disseminação de pornografia em geral – e a pornografia infantil, em particular[83]. A legislação federal obriga as empresas a reportarem a existência de pornografia infantil ao "International Center for Missing and Exploited Children" apenas quando elas tiverem ciência da existência de conteúdo do tipo em suas plataformas. Ou seja, não há uma obrigação jurídica específica de monitorar de maneira proativa a existência de pornografia infantil; mesmo assim, compreensivelmente, esse monitoramento se tornou uma prática padrão da indústria, como maneira de responder positivamente a essa demanda social. Ainda assim, quando detectam um material do tipo, por monitoramento voluntário, tornam-se legalmente obrigadas a notificarem as autoridades competentes[84].

Esse monitoramento promove uma checagem sobre se uma determinada imagem corresponde ou não a um material previamente cadastrado como pornografia infantil. Nos Estados Unidos, diversas autoridades contam com um banco de cerca de 720 mil imagens conhecidas, mantido em cooperação com o governo federal. É em face desse banco de referência, organizado pelo "International Centre for Missing and Exploited Children",

---

[82] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1636.

[83] Ver: Jeff Kossef, *The Twenty-Six Words That Created the Internet*, Cornell University Press, 2019, capítulo 3. Não é por outra razão que, desde a promulgação da seção 230 do "Communications Decency Act", em 1996, essa legislação – que garante uma forte e decisiva imunidade de responsabilidade civil para plataformas em razão de conteúdos gerados por usuários – contém uma exceção específica para crimes federais, o que abrange, na prática, especialmente a pornografia infantil. A seção 230 do "Communications Decency Act" será abordada no Capítulo 4-C.

[84] Jeff Kossef, *The Twenty-Six Words That Created the Internet*, Cornell University Press, 2019, capítulo 12. A Microsoft e America Online, por exemplo, também monitoram conteúdos em nuvem ou em servidores de e-mail em busca de imagens de pornografia infantil.

que cada postagem é checada pelas redes sociais, através de um algoritmo de reconhecimento de imagem denominado PhotoDNA[85].

Em casos de pornografia infantil, por óbvio, não existe qualquer tipo de controvérsia sobre a ilicitude de divulgação desse tipo de materiais, ação considerada crime em todos os países. Mas essa tecnologia – e o controle prévio que ela possibilita – avançou também para um outro campo como resultado da forte disputa por parte da grande indústria cultural para garantir que as aplicações da internet não se tornassem uma ameaça letal a suas atividades.

Aqui, novamente, o desenvolvimento da tecnologia de monitoramento e filtro prévio serviu como resposta a uma demanda de um segmento social (a indústria cultural), que também contava com respaldo em uma importante peça da legislação americana: o "Digital Millenium Copyright Act*"* de 1998 (DMCA) previa uma exceção à regra geral de imunidade civil das plataformas por conteúdo de terceiros, quando violassem direitos autorais. No caso, as plataformas fazem jus àquela imunidade legal apenas se implementarem a possibilidade de derrubada célere de conteúdo violador de direitos autorais, mediante notificação de sua existência, em procedimento próprio, por parte dos detentores daqueles direitos (procedimento de "notice and takedown*"*).

Nesse cenário, o Youtube tomou uma iniciativa que seria fundamental para seu futuro sucesso. Quando a plataforma foi adquirida pela Google, em 2006, por $1.6 bilhões de dólares, sua popularidade crescente devia bastante a conteúdos que infringiam direitos autorais. "Donos de conteúdo reclamavam fortemente que o sistema de 'notice and takedown' não dava ao Youtube qualquer incentivo para lidar de modo proativo com esses conteúdos irregulares"[86]. Ficou claro para o Google que para sua nova aquisição

[85] "Ao converter cada uma dessas imagens a uma escala de cor cinza, com uma grade de divisão por cima, e designando um valor numérico a cada quadrado, pesquisadores foram capazes de criar uma assinatura, que é mantida mesmo se as imagens são alteradas. Como resultado, plataformas podem determinar se uma imagem contém pornografia infantil nos microssegundos entre o upload e a publicação" – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1637. Chamado de "Project VIC", esse banco de dados é mantido pela entidade não governamental "International Centre for Missing and Exploited Children" e tem sido utilizada por forças policiais dos Estados Unidos (nesse país, sob coordenação do "U.S. Department of Homeland Security"), Reino Unido, Canadá, Nova Zelândia e Australia, pois possibilita a unificação das imagens acessadas por diversos órgão policiais locais – "Cloud-based archive tool to help catch child abusers", reportagem publicada pela *BBC News*, em 24/03/2014. Ver, ainda, www.projectvic.org.

[86] Nicolas P. Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p. 71.

prosperar, ela devia operar em bons termos com a indústria cultural. A resposta foi o desenvolvimento, por cerca de $60 milhões de dólares, do software "Content ID".

O programa permite a inclusão de materiais em um banco de dados, que passam a contar com uma "marca identificadora digital" ("digital fingerprint"), que serve para verificação dos vídeos publicados em suas páginas. Essa checagem automatizada não exclui a possibilidade adicional de os vídeos serem marcados ("flagged") pelos detentores de direitos autorais, utilizando-se do sistema "notice and takedown" previsto na legislação. Mas a automatização desse controle, evitando a postagem de vídeos com direitos autorais protegidos, é uma maneira de o Youtube viabilizar sua plataforma aberta de vídeos em parceria com a indústria cultural: quando um vídeo é identificado como sendo coberto por direitos autorais, a plataforma *de antemão* oferece aos "copyright owners" as escolhas de bloqueá-lo, aguardar para avaliar se ele se torna muito popular ou incluir propagandas que lhes tragam retorno direto. "O Content ID foi um investimento massivo que rendeu um retorno espetacular para o Google – e aos detentores de direitos autorais", diz Nicolas Suzor[87]. Nos termos de Lessig, a lei americana pode ter dado um limitado incentivo inicial; nesse caso, contudo, o mercado foi determinante para o desenvolvimento e formatação do "código".

Os dois campos mencionados acima – controle prévio e automatizado para combater a pornografia infantil ou a reprodução não autorizada de conteúdos protegidos por direitos autorais – encontram cada um deles correspondências em incentivos de relevantes leis americanas, que criam nessas searas importantes exceções à regra geral de imunidade jurídica das plataformas em razão e conteúdo produzido ou postado por usuários. Essas leis fixaram obrigações, restrições ou incentivos que terminaram por

[87] Estima-se que o Content ID tenha gerado mais de $2 bilhões de dólares em receita de anúncios para detentores de direitos autorais que tenham escolhido monetizar sobre vídeos postados por usuários. Mais de 8 mil detentores de direitos autorais estão registrados no sistema, que já identificou mais de 400 milhões de vídeos. Nicolas Suzor contextualiza a criação do Content ID pelo Youtube nas chamadas "Copyright Wars" – a grande disputa protagonizada pela industrial cultural para garantir a sobrevivência de seus negócios diante de ameaças surgidas com a popularização da internet e possibilidade de acesso gratuito massificado a conteúdos. Se no início as grandes empresas chegaram a processar individualmente usuários em função de "downloads*"* ilegais (como no célebre caso Napster) – uma estratégica custosa, de baixa efetividade e repercussões negativas de imagem pública –, atualmente o movimento tem sido de garantir que as estruturas da internet (seja em seu acesso, em suas grandes plataformas ou em suas aplicações) operem de partida em acordo com as regras de direitos autorais. O autor problematiza, ainda, como esse processo automatizado tem gerado inúmero equívocos de avaliação, muitas vezes derrubando do ar conteúdos que seriam lícitos mesmo à luz da legislação protetiva de direitos autorais – como vídeos informativos, pequenos trechos reproduzidos, ou de críticas sobre conteúdo de terceiros. – *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 70-73.

formatar a criação e a disseminação de tecnologias específicas para a realização dessa filtragem prévia – ressaltando que, como apontamos, as iniciativas das plataformas foram além daquilo que o direito requeria de maneira estrita. Em ambos os casos, imperativos de normas sociais e de mercado tiveram maior peso para o desenvolvimento do "código". Mas uma vez que essa nova possibilidade técnica estava cada vez mais avançada e consolidada, ela poderia ser ampliada para potencialmente qualquer nova seara. Inclusive por decisões autônomas das próprias plataformas, por uma espécie de autorregulação. Foi isso o que ocorreu – desta vez, na área de combate ao terrorismo.

No final de 2016, Facebook, Microsoft, Twitter e Youtube anunciaram que passariam a coordenar ações no combate ao terrorismo online. Isso ocorreria pelo compartilhamento de uma base de dados de "hashes*"* ("digital fingerprints") sobre imagens terroristas violentas ou vídeos de recrutamento que tivessem sido previamente removidos de uma de suas plataformas. "Nossas empresas começarão a compartilhar as marcas das imagens e vídeos terroristas mais extremos e odiosos que removemos de nossos serviços – conteúdos que provavelmente violam todas as nossas respectivas políticas"[88].

Cada empresa seria capaz de adicionar vídeos a esse banco de dados; em seguida, as demais participantes se utilizariam da atualização dessas referências para promover uma checagem própria em suas plataformas, preservada a autonomia de cada uma pela decisão de fazer ou não a remoção. "Nenhuma informação pessoal identificável será compartilhada, e os conteúdos correspondidos ("matched") não serão automaticamente removidos. Cada empresa continuará a aplicar usas próprias políticas e definições de conteúdo terrorista ao decidir sobre a remoção de conteúdo", apontou o comunicado conjunto. "E cada empresa continuará a aplicar suas práticas de transparência e revisão para pedidos governamentais, além de manter seu próprio processo de recursos ("appeal process") para decisões de remoção e demais insatisfações", destacaram as quatro

[88] "Partnering to help curb the spread of terrorist content online", comunicado conjunto do Facebook, Microsoft, Twitter e Youtube, publicado em 5/12/2016, disponível em https://blog.google/around-the-globe/google-europe/partnering-help-curb-spread-terrorist-content-online/

gigantes[89]. Atualmente, há pelo menos 15 empresas participando da iniciativa – e um banco de dados de cerca de 100 mil imagens consideradas como propaganda terrorista[90].

A aplicação de um filtro prévio a imagens, por meio de um banco de dados comum, passava a ser uma decisão das próprias empresas envolvidas – poucas, pois inicialmente apenas quatro, mas que dominam boa parte da internet comercial do mundo. Implementada em um terreno politicamente bastante delicado e dado a controvérsias, como é o caso da *conceituação de propaganda terrorista*, a iniciativa chamou a atenção também pelo que revelava sobre essa capacidade técnica existente para se articular, *sem iniciativa governamental aparent*e, uma base concentrada e compartilhada voltada a servir de parâmetro de controle de conteúdo.

Esse aspecto não passou desapercebido ao "Center for Democracy and Technology" (CDT), que manifestou "profunda preocupação de que esse projeto conjunto possa criar um precedente para censura interplataformas ("cross-site censorship") e se torne um alvo para governos e atores privados que busquem a supressão de discursos em toda a rede". A organização avaliou que a iniciativa decorria de pressões intensas de governos para medidas de combate ao terrorismo – e que havia o risco de essa ação sinalizar um caminho de censura concentrada do discurso online. A CDT também destacou que, para além de não haver qualquer definição internacional sobre o que constitui propaganda terrorista, as empresas possuem a prerrogativa jurídica de serem muito mais restritivas com relação a discursos em comparação com Estados[91]. E concluiu:

> *"Sem diretrizes claras sobre qual discurso pode se tornar um alvo e sem uma garantia de avaliação independente de que as empresas participantes estão restritas aos termos de seu próprio acordo, [a iniciativa] pode minar a confiança de usuários nas informações que eles acessam nesses serviços. Se*

[89] O comunicado conjunto finaliza: "Através dessa colaboração, estamos comprometidos em proteger a privacidade de nossos usuários e sua habilidade de se expressarem livremente e de modo seguro em nossas plataformas. Também esperamos nos engajarmos com a comunidade mais ampla de interessados (´stakeholders´) de um modo transparente, ponderado e responsável enquanto aprofundamos nosso objetivo comum de prevenir o avanço de conteúdo terrorista online enquanto respeitamos direitos humanos".

[90] Evento "The State of Online Speech and Governance", debate promovido pelo *Berkman Klein Center for Internet and Society* (*Harvard University*), entre Jonathan Zittrain e Monika Bickert, em 03/12/2018 – vídeo disponível em https://www.youtube.com/watch?v=lWkhFB0f2tw

[91] "É altamente provável de que boa parte do conteúdo nesse banco de dados será discurso legal nos Estados Unidos", até porque as empresas teriam "desenvolvido definições bastante idiossincráticas de 'conteúdo destinado a recrutramento para organizações terroristas', 'organizações perigosas' e 'ameaças violentas (diretas ou indiretas)'" – "Takedown collaboration by private companies creates troubling precedent", editorial publicado pelo *Center for Democracy and Technology*, em 6/12/2016, acesso em http://cdt.org

> *uma imagem ou vídeo pode ser suprimida de todas as maiores plataformas de uma única vez, o que irá acontecer com a habilidade de pessoas de acharem esse material para reportar as notícias, o debate político, a pesquisa acadêmica, as discussões sobre políticas públicas, arte, literatura, ou qualquer uma da miríade de razões legítimas existentes para postar e acessar até mesmo os exemplos mais horríveis de propaganda terrorista? Mesmo considerando nossas recomendações, essa proposta é um exemplo claro da ameaça fundamental que a centralização traz para a liberdade de expressão e da subestimação da importância de um ecossistema de informações que inclua não apenas diversidade de vozes, mas diversidade de provedores de serviços e de conteúdo ('content hosts and service providers')"*[92].

A tendência tecnológica, entretanto, aponta para uma direção clara. O Facebook já trabalha para avançar a possibilidade de *controle e filtragens prévias de vídeos ao vivo* – uma das razões são os diversos episódios de suicídios transmitidos em tempo real pela plataforma, muitos deles por jovens[93]. Por ora, porém, transmissões ao vivo ainda são moderadas de modo mais eficaz a partir de denúncias ("flagging") de outros usuários.

Mas embora esse controle seja mais difícil quando a transmissão é ao vivo, a tecnologia permite marcar um determinado vídeo e evitar sua republicação ou proliferação. Foi o que o Facebook fez quando o atentado de Christchurch, ocorrido em março de 2019 na Nova Zelândia e que resultou em 51 pessoas mortas, foi transmitido ao vivo pelo atirador na plataforma. O massacre a tiros foi considerado singular porque foi

---

[92] A manifestação da CDT, embora voltada a esse banco de dados em especial, levanta diversas outras questões mais amplas e relacionadas à moderação de conteúdo pelas grandes plataformas, incluindo a transparência de critérios, possibilidade de procedimentos recursais, a publicação de relatórios, entre outros – aspectos que serão abordados ao longo do Capítulo 4.

[93] Kate Klonick, "Inside the team at Facebook that dealt with the Christchurch shooting", *The New Yorker*, artigo publicado em 25/04/2019. Mencionando a possibilidade de monitoramento de vídeos transmitidos *ao vivo*, Monica Bickert, "head of global policy management" do Facebook, comentou ao final de 2018: "com certeza nós gostaríamos de poder impedir vídeos [ao vivo] de ir para o ar, em algumas áreas. Estamos trabalhando em questões como suicídio. Buscamos tecnologia que identifique imediatamente se alguém ameaça se matar ou se machucar, inclusive para acionar serviços de proteção. Se alguém transmitir ao vivo algo como abuso infantil ou um estupro, queremos impedir isso na hora. Mas hoje a tecnologia funciona melhor quando lidamos com imagens já conhecidas [por conta dos "hashes"] - Evento "The State of Online Speech and Governance", debate promovido pelo *Berkman Klein Center for Internet and Society* (*Harvard University*), entre Jonathan Zittrain e Monika Bickert, em 03/12/2018 – vídeo disponível em https://www.youtube.com/watch?v=lWkhFB0f2tw.

planejado, dirigido e executado para se valer exatamente da transmissão ao vivo pelo Facebook[94].

Com a repercussão do vídeo, a plataforma iniciou seu protocolo de crise, que inclui equipes de plantão especializadas sempre a qualquer momento, ao redor do globo. Depois de algumas poucas horas, o procedimento padrão a ser adotado seria, depois de apagar a postagem original, marcar a identidade digital do vídeo para evitar sua republicação. Mas ainda doze horas após a transmissão, o Facebook encontrou uma dificuldade inesperada: alguns usuários manipulavam digitalmente o vídeo exatamente para driblar essa restrição. A solução encontrada foi fazer uma identidade digital a partir da trilha de áudio; essa técnica, alinhada a algumas outras, deu certo. Nas 24 horas subsequentes, 1.5 milhões de cópias do vídeo foram removidas do Facebook, sendo 1.2 milhões delas no momento do upload[95].

2.B – Análise automatizada de linguagem

Há uma tendência crescente também de desenvolvimento de ferramentas automatizadas para análise de linguagem – que busca contrabalancear a ainda premente necessidade de atuação permanente de numerosas equipes de moderadores humanos, revisores de conteúdo[96].

Ao longo de 2018, ao menos metade do conteúdo removido por vedações a "discursos de ódio" foi feita de *modo proativo* pelo Facebook com uso de tecnologia automatizada – o que inclui o monitoramento de imagens e também de textos. Como Monica Bickert, "head of global policy management*"* da empresa, ressalta, o uso desse tipo de tecnologia é atualmente mais eficaz para o monitoramento de imagens *já previamente conhecidas* ("hashes*"*), tal como explicado na seção anterior deste capítulo, mas tem sido utilizada cada vez mais para avaliação generalizada de conteúdos –

94 "(...) uma característica surpreendente [do episódio] foi sobre como a violência foi sobremaneira *online*, e como o atirador aparentava durante a transmissão estar ciente sobre como seus atos seriam vistos e interpretados por distintas subculturas da internet" – Kevin Roose, "A mass murderer of, and for, the Internet", artigo publicado pelo *The New York Times*, em 15/03/2019. Mencionado também por Kate Klonick, "Inside the team at Facebook that dealt with the Christchurch shooting", *The New Yorker*, artigo publicado em 25/04/2019.

95 Kate Klonick, "Inside the team at Facebook that dealt with the Christchurch shooting", *The New Yorker*, artigo publicado em 25/04/2019.

96 Sobre a atuação de revisores humanos de conteúdos nas redes sociais, ver Capítulo 2-E.

especialmente na área de avaliação de discursos de ódio – a partir da evolução das formas de "aprendizado" de inteligência artificial[97].

Enquanto o uso dessas tecnologias aumenta, ficam evidentes também as limitações existentes e os riscos decorrentes desse tipo específico de monitoramento. As principais críticas são relativas ao fato de que a inteligência artificial é incapaz de entender o contexto ou interpretar o real significado e intenção de quem produz o discurso, além do risco de algoritmos que atuam por "machine learning" incorporarem vieses discriminatórios existentes na sociedade ou embutidos em sua formulação.

Alessandra Gomes, Dennys Antonialli e Thiago Dias Oliva ressaltam essas observações a partir de um estudo conduzido pelo centro de pesquisas Internetlab[98], que fez uso da ferramenta "Perspective" – tecnologia de inteligência artificial desenvolvida pela Jigsaw, por sua vez de propriedade da Alphabet, empresa-mãe do Google. A "Perspective" mede o nível de "toxicidade" em um texto, considerando "tóxico" como um "comentário rude, desrespeitoso ou não-razoável que provavelmente fará com que você abandone uma discussão". A modelagem e treinamento do programa são feitos em cima de avaliações de pessoas sobre se um conteúdo é "muito saudável" ou "muito tóxico".

O estudo aplicou a ferramenta em postagens do Twitter para comparar a "toxicidade" das postagens de 80 "drag queens" famosas (ex-participantes de um programa televisivo de sucesso internacional) e de políticos reconhecidos como líderes de extrema-direita nacionalista. Postagens de outras personalidades – como Michelle Obama e Donald Trump, por exemplo – também foram avaliadas, tomando-as como referencias intermediárias para aqueles dois grupos. No total, foram analisados 114 mil tuítes (em língua inglesa) com a versão então mais recente do "Perspective".

---

[97] Evento "The State of Online Speech and Governance", debate promovido pelo *Berkman Klein Center for Internet and Society* (*Harvard University*), entre Jonathan Zittrain e Monika Bickert, em 03/12/2018 – vídeo disponível em https://www.youtube.com/watch?v=lWkhFB0f2tw . A respeito do papel crescente, no último ano, do uso de inteligência artificial para a identificação de conteúdos que firam a política de uso do Facebook, inclusive na área de discurso do ódio: "Facebook says it is more aggressively enforcing content rules", reportagem publicada pelo jornal *The New York Times*, em 23/05/2019; "Facebook: New AI tech spots hate speech faster", reportagem publicada pelo portal C-Net em 01/05/2019.

[98] Alessandra Gomes, Dennys Antonialli e Thiago Dias Oliva, "Drag queens e inteligência artificial: computadores devem decidir o que é 'tóxico' na internet?", artigo publicado pelo centro de pesquisas *InternetLab*, em 28/06/2019.

Os resultados demonstraram que alguns perfis de drag queens tiveram um nível de toxicidade maior do que de líderes de extrema-direita; para os pesquisadores, isso resultou do fato de que vieses foram incorporados à ferramenta, que avaliava palavras como "gay", "lesbian" (lésbica), "fag" (bicha) ou "bitch" (vadia) como possuidoras de um alto nível tóxico. No entanto, da mesma maneira como acontece com a palavra "nigger", no contexto norte-americano, muitas vezes minorias e grupos vulneráveis se apropriam dessas palavras para utilizá-las de forma antidiscriminatória e autoafirmativa, exatamente de modo a problematizar ou permitir a superação de preconceitos sofridos[99]

Ao mesmo tempo, a pesquisa destacou também postagens de líderes de extrema-direita que veiculavam opiniões discriminatórias de modo sutil, sem fazer uso aberto de linguagem ou termos agressivos – e que acabavam sendo avaliadas com baixo grau de toxicidade. "A ferramenta Perspective parece dar mais prevalência às palavras, do que às mensagens veiculadas ('underlying messages')", conclui Dennys Antonialli[100]. Em resumo, se a ferramenta fosse aplicada ao Twitter, muitas das publicações das "drag queens*"* seriam removidas, mesmo que tivessem a tônica central de proferir uma linguagem de afirmação de identidade do público LGBT.

Outro estudo também reforça, no mesmo sentido, o risco de viés discriminatório racial na detecção automatizada de "discursos de ódio", parte dele também se valendo do uso da ferramenta Perspective, com resultados falso-positivos que identificavam expressões comuns à população negra americana[101].

---

[99] "Como o uso de linguagem 'pseudo-ofensiva' foi identificado como forma de interação que serve para preparar membros da comunidade LGBTQ para lidar com hostilidade externa ao grupo, análises automatizadas que desconsiderem essa função social do discurso podem ter repercussões significativas na capacidade de membros da comunidade de reclamar esses termos e reforçar vieses danosos. Uma possível razão para tais distorções e vieses é o desafio que a análise do contexto representa para tecnologias baseadas em inteligência artificial. O Perspective parece apoiar-se, em grande medida, no nível 'geral' de toxicidade de palavras específicas ao invés de analisar a toxicidade de ideias ou ideologias, o que pode ser muito contextual. Essas discrepâncias provavelmente também têm origem em vieses presentes nos dados usados para treinar a ferramenta de inteligência artificial, assunto que é tópico de intenso debate no âmbito da comunidade das ciências da computação" – Alessandra Gomes, Dennys Antonialli e Thiago Dias Oliva, "Drag queens e inteligência artificial: computadores devem decidir o que é 'tóxico' na internet?", artigo publicado pelo centro de pesquisas *InternetLab*, em 28/06/2019.

[100] Dennys Antonialli, "Drag Queen vs. David Duke: whose tweets are more 'toxic'?, artigo publicado em *Wired*, em 25/07/2019, também comentando os resultados da pesquisa.

[101] Marteen Sap, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A. Smith, "The risk of racial bias in hate speech detection", *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp.1668–1678; acesso direto em "Google's algorithm for detecting hate speech is racially biased", artigo publicado por *MIT Technology Review*, em 13/08/2019.

Nesse sentido, o Center for Democracy and Technology (CDT) também enfatiza essas limitações ainda prevalecentes em análises automatizadas de conteúdo na internet. Também focando nas ferramentas de processamento de "linguagem natural", muitas vezes voltadas à identificação de "discursos de ódio", as conclusões do CDT ressaltam cinco limitações-chave: a) o melhor funcionamento da ferramenta para contextos determinados, sem ampliação com a mesma confiabilidade para demais áreas; b) risco de reprodução de vieses discriminatórios contra grupos vulneráveis ou marginalizados; c) necessidade de definição clara e precisa do conteúdo a ser mirado, que não pode se limitar a "extremismo" ou "radicalização"; d) baixa acuidade de resultados – média entre 70% e 75% de acerto, o que demanda cautela na aplicação da ferramenta, preferencialmente a ser usada para gerar uma análise posterior por revisão humana, e; e) facilidade de evasão do monitoramento, mediante alterações de elementos contextuais ou privilegiando mensagens mais sutis[102].

2.C – Bloqueio geográfico

Outra capacidade técnica relevante que plataformas possuem de realizar controle de conteúdo é o *bloqueio geográfico da publicação* – ou seja, a eventual proibição em função da localização dos usuários, normalmente associada a seu país de origem. Afinal, se as grandes redes sociais operam globalmente, é natural que haja conflitos de suas regras e operações principalmente com culturas e leis nacionais, criando assim um *cenário de potencial dinâmica competitiva entre os sistemas normativos das plataformas e das ordens jurídicas tradicionais* – como ilustram os exemplos a seguir.

Em 2006, o governo da Tailândia ameaçou bloquear o acesso ao Youtube no país por conta da veiculação do que considerava "vídeos ofensivos". O cerne da celeuma estava em uma série tida como crítica (ou desrespeitosa) para com o rei tailandês. Alguns dos vídeos eram apenas imagens satíricas da realeza, com pés no lugar da cabeça, por exemplo. Na Tailândia, contudo, insultar o rei constitui crime punível por até quinze anos de prisão. Como resultado, o Google/Youtube aceitou bloquear os vídeos "ofensivos" ao

[102] "Mixed Messages? The limits of automated social media content analysis", relatório publicado pelo *Center for Democracy and Technology*, em 28/11/2017, acesso em https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf. Como recomendações a "policymakers", usuários e desenvolvedores, o CDT também defende que o uso de análise automatizada de linguagem não deve nunca ser determinada por lei.

rei dentro da fronteira da Tailândia – mantendo, porém, alguns conteúdos críticos aos militares. A rede concordou que havia no caso ofensa à legislação local[103].

Pouco tempo depois, o Youtube vivenciou outra experiência semelhante, desta vez na Turquia. Em março de 2007, a plataforma foi bloqueada em todo o país em função de uma ordem judicial que julgava uma ação motivada um programa humorístico que sugeria que o "pai fundador" do estado turco moderno, Mustafa Kemal Ataturk, era gay. No país, insultos a Ataturk também são considerados crimes. Na esteira da ordem judicial e com a plataforma já bloqueada, o governo turco apresentou uma lista com diversos outros vídeos que deveriam também ser retirados do ar. A equipe da empresa se viu na situação de estudar a legislação local que proibia a difamação de Ataturk e então analisar caso a caso quais desses vídeos realmente, a seu ver, violavam as leis nacionais. Isso levou a um breve acordo que restabeleceu o serviço no país, mas que iria durar pouco: em junho, a Turquia exigiu ao Google que derrubasse os vídeos em todo o mundo. Como a empresa não aceitou, o Youtube foi bloqueado no país.

O Facebook também já se deparou com um caso semelhante. Em 2010, um grupo da rede social originado nos Estados Unidos ganhou certa notoriedade ao propor o "Dia de todas as pessoas desenharem Maomé", passando da marca de cem mil integrantes. A ideia era ridicularizar a proibição eclesiástica muçulmana de representação gráfica do profeta. Se um número muito alto de pessoas publicasse imagens de Maomé, a proibição religiosa cairia no ridículo, sendo exposta como absurda – além de tornar impraticável qualquer ameaça de retaliação contra determinada pessoa, já que seu alvo seria uma multidão. Com a repercussão do grupo e a aproximação da data agendada, uma corte superior do Paquistão chegou a proibir o acesso dos cidadãos do país ao Facebook, bem como a outras plataformas, como Youtube e páginas do Wikipedia e Flickr. O Facebook resolveu, diante da ordem, "apagar" a existência do grupo nas buscas realizadas dentro

[103] Segundo Nicole Wong, advogada que estruturou a política de conteúdo da empresa naqueles anos: "minha primeira reação foi: isso é apenas um cartoon. É um photoshop estúpido. Mas então tornou-se um momento de aprendizagem para mim sobre padrões internacionais de liberdade de expressão versus padrões da Primeira Emenda de liberdade de expressão, e então havia muito mais excepcionalismo americano da Primeira Emenda do que à primeira vista (...)".Ela viajou até a Tailândia para tentar resolver o impasse e, segundo relata, ficou surpresa com as demonstrações de amor que a população oferecia ao seu rei: "Toda segunda-feira literalmente 85% das pessoas aparece no trabalho em camisas e vestidos dourados ou amarelos e há uma razão histórica para isso: a única fonte de estabilidade naquele país é seu rei. Eles absolutamente o veneram. (...). Algumas pessoas choravam ao falar sobre como os insultos ao rei os ofendiam. Essa é a parte que me impactou. Quem sou eu, uma advogada americana sentada na Califórnia, para dizer: "não, nós não vamos derrubar esses vídeos. Vocês vão ter que conviver com isso." – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1623.

do Paquistão. "Para usuários do Paquistão, a página ofensiva simplesmente não estava lá; até sua remoção havia sido invisível (...). Mas o grupo continuou disponível no site em outros países; seus membros sem saber que ele havia sido filtrado"[104]. Depois dessa medida, o bloqueio ao Facebook no país foi suspenso pelos órgãos de governo.

Esses casos demonstram como as plataformas de alcance global podem reagir quando são cobradas por governos nacionais para promover a derrubada de algum conteúdo: podem aceitar o pedido com base nas leis nacionais e retirá-lo do ar, ou podem se negar a fazê-lo e se arriscarem a ter o serviço inteiro bloqueado em um determinado país[105]. Sob essa hipótese, estados nacionais possuem uma alavancagem relevante para fazer valer seus ordenamentos jurídicos – *uma alavancagem tão mais alta quanto mais importante for o mercado constituído por sua população*.

A arquitetura global da internet e das próprias plataformas, por sua vez, também lhes garante um espaço de liberdade para agir a despeito de regras jurídicas nacionais. Embora aparente ser uma situação muito excepcional, o Facebook já implementou uma política previamente definida que deixava de bloquear conteúdos considerados ilegais em determinados países ao considerar que isso não implicaria um risco jurídico real à empresa, pois seus respectivos governos não demonstravam interesse em promover a aplicação de suas próprias legislações. No caso, o Facebook tradicionalmente permite a seus usuários publicações com conteúdo que negue a existência do Holocausto. Em 2017, orientações corporativas internas da empresa foram publicadas pelo jornal The Guardian, revelando que, embora esse tipo de conteúdo fosse ilegal em 14 países onde o Facebook atua, a empresa iria proibi-los apenas na França, Alemanha, Israel e Austria – cujos governos efetivamente cobravam o Facebook a esse respeito[106].

---

[104] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 191.

[105] Monica Bickert, do Facebook, explica que a empresa possui um time jurídico dedicado a, quando são recebidos pedidos de remoção de conteúdo por parte de autoridades, avaliar se aquele determinado pleito está de acordo com as legislações nacionais (tanto com relação a aspectos materiais, quanto formais). Nesses casos, o Facebook produz relatórios públicos semestrais sobre os pedidos acatados – Evento "The State of Online Speech and Governance", debate promovido pelo *Berkman Klein Center for Internet and Society* (*Harvard University*), entre Jonathan Zittrain e Monika Bickert, em 03/12/2018 – vídeo disponível em https://www.youtube.com/watch?v=lWkhFB0f2tw

[106] "How Facebook flouts Holocaust denial laws except where it fears being sued", reportagem publicada pelo jornal *The Guardian*, em 24/05/2017; "Zuckerberg defends Facebook users right to be wrong – even Holocaust deniers", reportagem publicada pelo jornal *The Guardian*, em 18/07/2018.

Um último e mais singular caso demonstra, por fim, que bloqueios geográficos podem ser feitos por decisões autônomas das próprias plataformas. O caso do vídeo "Inocência dos Muçulmanos" já foi apresentado no Capítulo 1-A. Ali, foi destacado que o governo americano havia pedido publicamente ao Youtube a revisão sobre se o conteúdo do vídeo violava suas regras de uso – a resposta da empresa, em seguida, foi negativa, asseverando que o vídeo estava "claramente dentro de suas regras". O Facebook igualmente manteve o vídeo no ar[107].

Mas esse pedido não partiu apenas do governo americano. Em apenas duas semanas, o Youtube recebeu pedidos de bloqueio ou de revisão do vídeo de pelo menos 21 governos nacionais. Em alguns países – como Índia, Indonésia, Malásia, Cingapura, Jordânia e Arábia Saudita – nos quais a plataforma possuía tanto uma versão de uso local, quanto representação legal em cada país, o vídeo foi bloqueado sob o argumento de que seu conteúdo violava as leis locais. No Paquistão, Afeganistão e Bangladesh – onde o Youtube não possuía representação local – os pedidos foram negados. Esses governos, então, bloquearam todo o serviço da plataforma em seus respectivos países. O Bahrein e os Emirados Árabes Unidos não entraram em contato com a empresa, mas bloquearam os vídeos utilizando-se de meios próprios. "Em resumo, instalou-se mundialmente um frenético jogo jurisdicional, político e tecnológico de gato e rato"[108].

Mas no meio dessa frenética disputa, o episódio também foi marcado por uma decisão autônoma da própria empresa. Depois de ataques a representações diplomáticas norte-americanas no Egito e na Líbia, o Youtube bloqueou o acesso ao vídeo em ambos os países, mesmo sem qualquer pedido por parte de seus governos. Em comunicado, a empresa explicou seu posicionamento: "esse pode ser um desafio porque o que é *okay* em um país pode ser ofensivo em outro lugar". O Youtube reiterou que seu conteúdo estava de acordo com as regras da plataforma, mas "diante da situação bastante difícil na Líbia e no Egito nós temporariamente restringimos o acesso naqueles dois países"[109].

Comentando esse episódio, Garton Ash apontou que a obediência a legislações nacionais "pode levantar questões sobre as diferentes qualidades de leis em diversos

[107] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1624-1625.

[108] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, p. 67.

[109] "Youtube temporarily censors offensive vídeo in Egypt and Libya", reportagem publicada pelo portal *The Verge*, em 13/09/2012.

estados, mas possui uma lógica transparente e consistente". Por outro lado, bloquear um conteúdo sem um pedido jurídico de um governo transformaria o Youtube/Google, um "poder privado, em um árbitro – altamente não transparente e arbitrário – sobre o que pessoas em outros países podem ver"[110].

2.D– "Flagging"

Outro aspecto operacional relevante para compreender a prática de moderação de conteúdo pelas redes sociais é o sistema de "*flagging*" – pelo qual os próprios usuários de uma plataforma marcam ("flag") uma determinada publicação, reportando-a de maneira negativa (como irregular ou indesejável) e, por consequência, levando-a a passar por um processo de revisão, normalmente feito por moderadores humanos[111]. Essa ambivalência entre os termos "irregular" ou "indesejável" é proposital. Marcar a publicação é um ato de reclamação, que pode ser motivada por inúmeras razões: "um 'flag', em sua forma mais pura, é uma objeção"[112].

Contar com essa marcação feita por usuários é uma maneira direta de lidar com o problema de escala das grandes redes sociais. O volume massivo de publicações diárias é tão gigantesco que se torna impossível qualquer revisão completa de conteúdo. Por isso, diante de um desafio (moderação) no qual a escala de publicações se torna um problema em si, o "flagging", mais do que uma possibilidade operacional, é uma necessidade[113].

A adoção desse sistema decorre, assim, de dois incentivos ou ganhos principais: em primeiro lugar, é um modo de viabilizar a revisão de uma quantidade gigantesca de conteúdo, contando com um papel fiscalizatório pelos próprios usuários do sistema. Além disso, esse procedimento aumenta o grau de legitimidade da política de moderação de

---

[110] Timothy Garton Ash, *Free Speech: ten principles for a connected world*, Yale University Press, p. 68. Nesta passagem, Garton Ash classifica a postura do Youtube como "compreensível", ainda que diante das ponderações feitas. Ele também problematiza o chamado "veto do assassino": argumento segundo o qual o fato de a plataforma ceder às pressões resultantes das ameaças ou condutas violentas apenas fomentaria mais reações desse tipo no futuro, criando um ciclo perverso de ameaça à liberdade de expressão.

[111] Conforme próximo tópico deste capítulo.

[112] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 91.

[113] Para um relato panorâmico sobre o sistema de "flagging" no Youtube, Twitter e outras plataformas, ver Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, pp. 87-97 e pp. 128-133.

conteúdo perante esses mesmos usuários, já que revisões são feitas em regra quando há pedidos expressos nesse sentido[114].

Dados de 2016 já apontavam que o Facebook ultrapassava a marca de um milhão de publicações marcadas por usuários, diariamente, no mundo[115]. Como em outras plataformas, a maior parte desse conjunto, no entanto, não se refere a efetivas violações de políticas de conteúdo – pelo contrário. Parte significativa das denúncias decorrem de disputas ou conflitos interpessoais entre os usuários – é comum que os conteúdos sejam reportados por divergências opinativas ou posturas idiossincráticas. É por isso que, ao reportar um conteúdo, o usuário normalmente é levado a preencher em sua tela um questionário, que auxilia a plataforma a organizar as demandas e dar prioridades a casos urgentes[116].

O fato de uma comunidade "policiar a si mesma" pode ser conveniente para a gestão da plataforma, mas carrega consigo riscos e efeitos colaterais relevantes:

> *"ainda mais quando essa 'comunidade' não é de modo algum uma comunidade, mas uma base de usuários de milhões ou bilhões de pessoas, ao longo do globo, falando diferentes línguas e com crenças tão antagônicas entre si, além de às vezes contrariar os objetivos da própria plataforma (...). É como um grupo de vigias de bairro ('neighbourhood watch'); suas intenções podem ser boas, mas muita coisa pode dar errado na prática"*[117].

---

[114] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard* Law Review Volume 131 (2018), p. 1638.

[115] "Facebook gets 1 million user violation reports a day", reportagem publicada por *CNN Business*, em 12/03/2016.

[116] "Usuários [do Facebook] que reportam algum conteúdo primeiro clicam em uma opção 'Reportar/Marcar como Spam', que em seguida guia-os rapidamente a descrever a situação em termos como 'Discurso do Ódio', 'Comportamento Violento ou Prejudicial', ou 'Eu não gosto deste post'. Alguns tipos de relatos, como de abusos ou de lesões auto infligidas, guiam os usuários para a opção de 'relato social' – uma ferramenta que 'permite às pessoas reportar conteúdos problemáticos não apenas ao Facebook mas também diretamente a seus amigos, de modo a ajuda-los a resolverem conflitos'. Para melhorar o tempo de resposta da equipe de moderação, o fluxo dos relatos também possui o propósito instrumental de criar uma triagem do conteúdo marcado para revisão. Isso torna possível ao Facebook priorizar imediatamente alguns conteúdos e, quando necessário, notificar as autoridades a respeito de situações emergenciais, como suicídios, ameaças violentas iminentes, terrorismo ou autolesões. Outros conteúdos, como possível discurso do ódio, nudez, pornografia, ou assédio, podem entrar na fila em bancos de dados menos urgentes para posterior revisão" – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1639.

[117] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 88.

Um caso ocorrido em 2014 é representativo de como o sistema de "flagging" pode ser transformado em "arma" em contendas políticas ou ideológicas. Nesse ano, o Facebook começou a suspender diversos perfis de "drag queens", alegando violação de sua política de obrigatoriedade de uso de "nomes reais". Centenas de "drags" que usavam seus nomes artísticos tiveram suas contas suspensas, boa parte delas proveniente da cidade americana de São Francisco. Os episódios geraram reclamações públicas e pressão por reversão da medida por grupos que atuam em favor de direitos da população LGBT. Depois de algumas semanas, o Facebook se desculpou e passou a aceitar que *nomes sociais* de "drag queens" fossem aceitos pela plataforma como "nomes reais", alterando a interpretação sobre essa sua própria regra. Mas o mais surpreendente nesse episódio foi a revelação pelo próprio Facebook de que a suspensão das contas havia ocorrido como decorrência de um único usuário ter marcado as centenas de contas de "drag queens", acusando-as de nomes falsos. A empresa demorou para perceber que se tratava de uma ação coordenada, proveniente de uma única pessoa. No Twitter, esse usuário aproveitava para também divulgar seus "alvos" e motivar colegas a reforçarem as reclamações. Segundo uma reportagem da época, a pessoa tinha motivação religiosa e marcava as contas de drag queens acusando-as de "pervertidas e sodomitas"[118].

Danos decorrentes de ações desse tipo se tornaram ordinários, como ilustra um outro caso brasileiro de 2018. O perfil do Instagram "@vulva.livre" chegou a 12 mil seguidores, a partir de postagens com conteúdos sobre sexualidade e saúde feminina. Tornou-se a atividade profissional de sua autora, que passou a participar de eventos, palestras e escrever colunas na imprensa sobre o tema. Depois de ela participar de uma entrevista em um grande portal de notícias, em meio ao período eleitoral e a críticas de caráter político-partidário, a página foi alvo de diversas denúncias e terminou derrubada pela plataforma, ainda que nenhuma razão específica tenha sido fornecida à autora. A página somente foi restabelecida depois do ajuizamento de uma ação que pleiteava também danos morais. A posição do Facebook (controladora do Instagram) nos autos deixa claro que a derrubada acabou ocorrendo apenas pelo alto número de denúncias

[118] "RealName Police and the real story behing Facebook's name policy fumble", reportagem publicada pelo site *The Daily Dot*, em 03/10/2014; Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, pp. 94-95.

contra o perfil, mesmo que nenhuma violação aos termos de uso tenha sido apontada ou reconhecida pela empresa[119].

2.E – Moderadores: a aplicação das regras por revisores humanos

O sistema de moderação das redes sociais – em especial, do Facebook – reserva um papel central a um exército de moderadores contratados, cujas funções são atualmente imprescindíveis: apenas uma análise humana permite que conteúdos de publicações possam ser devidamente contextualizadas (inclusive culturalmente) e, portanto, analisadas diante das regras pré-estabelecidas. Essa análise por moderadores costuma ser feita após a publicação – em geral, ela é *reativa*, promovendo a revisão de conteúdo que tenha sido marcada ("flagged") por outros usuários[120].

Atualmente, o Facebook possui cerca de 15 mil moderadores contratados em todo o mundo – o que equivale à metade de sua equipe dedicada a aspectos de segurança. O número aumentou vertiginosamente a partir de 2017 – quando esses revisores somavam 4,5 mil pessoas – na esteira do aumento de críticas públicas a conteúdos "tóxicos" veiculados pela plataforma. Uma pequena parte dos empregados é da própria empresa, mas a necessidade de ampliar essa atividade levou à contratação majoritária de moderadores terceirizados, nos Estados Unidos e em diversos outros países. Esse modelo possibilitou ao Facebook "escalar globalmente" seu time de revisores, de modo a avaliar postagens em mais de 50 línguas diferentes[121]. Como consequência de sua atuação global para bilhões de usuários, impõe-se a necessidade de contratação de revisores fluentes em diversas línguas nativas – habilidade necessária para poder fazer inclusive a interpretação dos conteúdos postados diante de contextos culturais e linguísticos específicos.

A adoção de um sistema de regras por plataformas implica a expectativa de que elas serão aplicadas de maneira uniforme por todos esses moderadores espalhados pelo planeta. Monica Bickert, "head of global policy*"* do Facebook, explica:

---

[119] Tribunal de Justiça de São Paulo, processo nº 1010421-22.2019.8.26.0566, em trâmite na 5ª vara cível do foro de São Carlos.

[120] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1638.

[121] "The impossible job: inside Facebook's struggle to moderate two billion people", reportagem publicada pelo portal *Motherboard*, em 23/08/2018; "The Trauma Floor: the secret lives of Facebook moderators in America", reportagem publicada pelo portal *The Verge*, em 25/02/2019.

> "*Uma das maneiras com que tiramos os preconceitos da equação é dizendo que não esperamos que os revisores determinem se eles consideram algo odioso ou não, pois isso poderia gerar decisões bem diferentes entre um revisor situado na Califórnia versus um outro em Dublin. Ao invés disso, estamos dizendo: aqui está o critério objetivo; se atinge esses critérios ('and it ticks theses boxes'), remova. É assim que atingimos uma consistência*"[122].

No Facebook, a atuação dos moderadores é organizada em um esquema de pirâmide. Em um primeiro nível, mais amplo e normalmente formado por revisores terceirizados de empresas espalhadas pelo mundo, é feita a moderação do dia a dia, aplicando as regras mais claramente estabelecidas para "flags" referentes a nudez, pornografia, insultos baseados em religião, etnia ou orientação sexual, incitação à violência contra pessoas ou animais e etc. Um segundo nível, formado por funcionários mais experientes e normalmente sediados nos Estados Unidos, é responsável pela supervisão da equipe mais básica (incluindo as decisões que são repassadas do primeiro nível ao segundo) além das decisões sobre temas previamente definidos como prioritários, tais como ameaças iminentes de violência, autolesões, terrorismo ou suicídio. No terceiro e último nível atuam empregados mais graduados do próprio Facebook, sediados na Califórnia[123]. As decisões tomadas no processo de moderação são sempre revistas por amostragem, de modo a garantir um nível satisfatório de *consistência* na aplicação das regras. Se há divergências sobre um determinado conteúdo, ele é enviado para um nível superior da cadeia decisória, visando ao fim e cabo uma atualização do livro de regras – um processo constante.

O nível de consistência esperado nunca é de 100% - se milhões de postagens são revisadas por dia, a própria escala da empreitada pressupõe que erros ocorram. Uma das

---

[122] "The impossible job: inside Facebook's struggle to moderate two billion people", reportagem publicada pelo portal *Motherboard*, em 23/08/2018.

[123] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1640-1643. Segundo Tarleton Gillespie, "no topo, a maioria das plataformas possui uma equipe interna de política de uso ('policy') encarregada de supervisionar a moderação. Essa equipe define as regras daquela plataforma, supervisiona sua aplicação, decide os casos particularmente difíceis e molda as novas políticas em resposta. São equipes normalmente pequenas, em regra apenas um punhado de empregados fixos da empresa (...). Essas equipes são obscuras para os usuários, por razões design e política ('by design and policy'). (...). Ainda assim, tem uma influência abrangente sobre onde as linhas são colocadas, quais tipos de punições são impostas, bem como sobre a abordagem filosófica que as plataformas aplicam em suas próprias governanças" - Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 117.

empresas terceirizadas – Cognizant, localizada no Arizona – estabelece um padrão de 95% de acertos. Normalmente, sua equipe fica entre picos na casa de 80% ou alguns pontos abaixo daquela meta[124].

Nos anos iniciais da empresa, o primeiro nível de moderação era feito por jovens recentemente formados em universidades e que trabalhavam na cidade de São Francisco. Hoje, esses moderadores trabalham em "call centers" nas Filipinas, Irlanda, México, Turquia, Índia, Leste Europeu ou nos próprios Estados Unidos. Essa diversidade cultural e linguística – necessária para a revisão de conteúdo de usuários de inúmeros países – é implementada a partir da premissa de que todos esses revisores passam por um processo de treinamento profissional para que as regras sejam aplicadas de maneira uniforme em todo o mundo. De qualquer modo, não se trata de um trabalho fácil. Cada postagem do Facebook levanta ao moderador duas questões: primeiro, se há alguma violação às regras da empresa. Em seguida, em caso positivo, o moderador deve apontar especificamente qual foi a regra violada. Se há uma decisão correta sobra a remoção do conteúdo, mas com a opção equivocada por uma determinada regra, o resultado não será contabilizado como um acerto. Acompanhando o trabalho diário de um moderador, uma reportagem contabiliza cerca de 400 avaliações diárias; cada postagem é revisada, na média, em menos de 30 segundos[125].

Um dos problemas levantados a respeito da atividade desse exército de moderadores são os efeitos psicológicos sobre a saúde mental dos trabalhadores em função do contato constante, por horas seguida em cada expediente diário, com conteúdos violentos e problemáticos. "Aqui está uma piada racista. Aqui está um homem fazendo sexo com um animal de uma fazenda. Aqui está um vídeo explícito de um assassinato gravado por um cartel de drogas" é uma descrição prosaica de um dia de trabalho. Sem surpresa, parte desses funcionários terceirizados – que ganham por volta de 15 dólares por hora no Arizona, por exemplo[126] – reclamam de depressão profunda ou "stress*"* pós-

[124] "The Trauma Floor: the secret lives of Facebook moderators in America", reportagem publicada pelo portal *The Verge*, em 25/02/2019. Para uma investigação social sobre a consolidação desse novo ramo profissional e das condições de trabalho de moderadores de contéudo que prestam esses serviços para plataformas de internet em uma economia globalizada, ver: Sarah T. Roberts, *Behind the screen: content moderation in the shadows of social media*, Yale University Press, 2019.

[125] "The Trauma Floor: the secret lives of Facebook moderators in America", reportagem publicada pelo portal *The Verge*, em 25/02/2019.

[126] Um empregado do Facebook ganha em média 240 mil dólares por ano, considerando salário, bônus e ações. Um moderador de conteúdo terceirizado que trabalha no Arizona ganha cerca de 28,8 mil dólares por ano - "The Trauma Floor: the secret lives of Facebook moderators in America", reportagem publicada

traumático como decorrência da exposição contínua e prolongada a tais tipos de publicações. Há ainda um rígido controle de horário de trabalho, além de uma vedação à possibilidade de uso de celulares ou contato com terceiros para conversas pessoais durante o expediente – com o intuito de proteger a privacidade dos usuários da plataforma[127].

O fato de as regras de discursos serem assimiladas e internalizadas pelos moderadores não quer dizer que eles concordem com a lógica que as determina, ou que vislumbrem consistência nessas razões. Miguel, um dos moderadores do Facebook que trabalham no Arizona, relata em reportagem que busca aprender como aplicar as regras e fazê-lo de modo diligente, embora as vezes não entenda suas razões. Uma postagem que diz que "pessoas autistas devem ser esterilizadas" lhe parece extremamente ofensiva, mas é permitida pela política da empresa porque o autismo não identifica uma "categoria protegida de pessoas", como ocorre com critérios de raça ou gênero, por exemplo. A postagem "homens deveriam ser esterilizados" deveria ser removida, por exemplo[128].

Kate Klonick traça um paralelo entre o papel de moderadores com a função tradicional de juízes, igualmente responsáveis pela adjudicação profissional de regras – na medida em que estão adstritos a normas pré-estabelecidas e de quem se espera que suas predisposições pessoais sejam desconsideradas[129]. A comparação, contudo, parece falha ou insuficiente. Moderadores são empregados adstritos às regras definidas por seus superiores, com pouco ou quase nenhum poder para superar o que, em suas percepções, seriam inconsistências ou inequidades. Além disso, espera-se deles um papel muito mais

pelo portal *The Verge*, em 25/02/2019. Depois de uma série de reportagens a esse espeito por parte de vários veículos ao longo de anos, o Facebook anunciou aumento de salários para seus moderadores terceirizados (ao menos aqueles sediados nos Estados Unidos) e padrões mais altos de suporte psicológico – "For once, we have good news about Facebook and Content Moderadors", reportagem publicada pelo portal *Slate*, em 13/05/2019.

[127] The Trauma Floor: the secret lives of Facebook moderators in America", reportagem publicada pelo portal *The Verge*, em 25/02/2019.

[128] O uso do conceito de "categorias protegidas" pelo Facebook será abordado especificamente no Capítulo 3-B.

[129] "Uma análise dessas regras internas revela uma estrutura que de muitas maneiras replica o processo decisório presente na jurisprudência moderna. Moderadores de conteúdo agem por uma capacidade muito similar àquela de um juiz: moderadores são treinados para exercer um juízo profissional a respeito da aplicação das regras internas de uma plataforma e, ao aplicar essas regras, espera-se que moderadores usem conceitos legais como relevância, razões por meio de exemplos e analogias, além de aplicar testes multifatoriais" – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1641-1642.

automatizado do que aquele esperado de juízes – que precisam articular as razões de suas decisões com razoável independência.

Aqui, novamente, prevalece o alto nível de automatização que caracteriza a moderação de conteúdo pelas redes sociais, que encaram essa questão como um problema de escala, solucionável por uma abordagem sistêmica – na qual uma margem de erros é prevista e contabilizada de partida.

2.F – Filtragem algorítmica: entre o visível e o invisível

Este capítulo não estaria completo se ficasse restrito apenas a operações vinculadas diretamente com políticas de moderação (e decisões de derrubada) de conteúdos – que, em resumo, operam por meio de um sistema de regras sob a lógica permitido/proibido. Um outro aspecto central compõe a curadoria de conteúdo feita pelas grandes redes sociais: *a customização do conteúdo exibido, feita por filtragem algorítmica, e que opera sob o binômio visível/invisível*.

Embora este possa ser o tópico de maior complexidade deste capítulo, é também a tecnologia mais opaca e blindada do debate público, em parte porque muitas vezes algoritmos são protegidos como segredos de negócios, em parte também porque, mesmo quando esse não é o caso, sua operação é tão complexa a ponto de dificultar sua compreensão por pessoas sem formação técnica especializada. Algoritmos, por exemplo, não se limitam a aplicar instruções previa e expressamente definidas em sua programação, pois podem operar por meio de "aprendizagem automática" ("machine learning*")*, que evoluem a partir de seu próprio uso ou da alimentação de dados, incluindo a interação entre algoritmos distintos[130].

[130] Jonathan Zittrain argumenta que o processo de "machine learning" – pelo qual algoritmos "aprendem" a identificar padrões em meio a uma torrente de dados – pode levar a um "déficit intelectual" social significativo, já que as pessoas, empresas e demais organizações podem saber que o algoritmo funciona para uma determinada finalidade, mas não eventualmente *como* ou *porquê*. Ele compara essas situações a remédios que são usados e funcionam há tempos, embora não exista explicação científica sobre a relação causal que leva a seus efeitos – como no caso da aspirina, descoberta em 1897 e "explicada" cientificamente apenas em 1995. Para além de vieses (eventualmente discriminatórios) decorrentes de sua programação, Zittrain se preocupa com as consequências sociais do amplo uso de algoritmos caso eles "deem certo" – e defende que haja maior acesso por pesquisadores aos algoritmos e seus conjuntos de dados nos casos de amplo uso social de modo a manter um nível de controle e conhecimento sobre como operam – "The hidden costs of automated thinking", artigo publicado por *The New Yorker*, em 23/07/2019. Para uma visão mais geral sobre as diversas aplicações de algoritmos nas esferas públicas e privadas, bem como uma defesa de parâmetros para uma regulação jurídica eficiente, ver: Diogo R. Coutinho e Beatriz Kira, "Por que (e como) regular algoritmos? ", artigo publicado no portal *Jota*, em 02/05/2019.

Desde já, é importante sublinhar que algoritmos nunca são neutros. No caso das grandes redes sociais isso é bastante evidente: algoritmos determinam, entre a torrente interminável de conteúdos publicados, *quais serão aqueles que serão exibidos para cada usuário* – inclusive em qual ordem ou com qual destaque. *Funcionam, assim, como uma nova maneira de editar a quais conteúdos será dada visibilidade.* No limite, dar maior visibilidade a um conteúdo significa promove-lo; *sua invisibilidade, por outro lado, pode alcançar a equivalência prática de proibi-lo*[131].

A chave para entender essa nova forma de organização da visibilidade de discursos nas grandes redes sociais deve partir da percepção de que *o barateamento da possibilidade de publicação pelas pessoas aumentou a oferta de expressões*, o que em contrapartida *torna a visibilidade ou a atenção do público um bem mais escasso – e, portanto, mais valioso*. "A mudança mais importante no ambiente expressivo se resume a uma ideia: não é mais o discurso em si que é escasso, mas a atenção dos ouvintes", explica Tim Wu[132].

Para Wu, o corpo jurídico da jurisprudência da Primeira Emenda americana, que começou a emergir no começo do século XX, lidava com um mundo no qual a maior ameaça à liberdade de expressão era a supressão de discursos de dissidentes políticos. A proteção ao "panfletário solitário" ("lonely pamphleteer") foi o embrião dessa mentalidade[133], que pressupunha um "mundo pobre em informações", além de focar em ameaças vindas, sobretudo, de governos e suas ações de aprisionamento. Havia, assim, três premissas para a cultura da Primeira Emenda no século XX: escassez informacional; uma audiência com tempo suficiente para lidar com as opiniões e informações colocadas

[131] Um exemplo aproximado, mas fora do ambiente das redes sociais, pode ser verificado no Google: em setembro de 2019, a plataforma anunciou que alterou o algoritmo que determina a classificação de notícias em seu sistema de busca para priorizar aquelas que publicaram uma determinada história primeiro – ou seja, que deram um furo jornalístico. "A mudança é uma forma de tentar garantir que a receita de anunciantes seja direcionada às companhias que investiram no trabalho de reportagem. A medida também é um aceno às empresas de mídia que há anos demandam do Google repasse de verba que ele obtém a partir do conteúdo de terceiros indexado em seu site": "Google muda algoritmo e prioriza notícias originais na busca", reportagem publicada pelo jornal *Folha de S. Paulo*, em 12/09/2019.

[132] Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University,* artigo publicado em 01/09/2017.

[133] O termo "panfletário solitário" se refere ao caso Abrams v. United States, julgado pela Suprema Corte americana em 1919, no qual o voto dissidente de Oliver Wendell Holmes, seguido também por Louis Brandeis, marca o início da moderna jurisprudência americana sobre liberdade de expressão no século XX. No caso, com ambos juízes vencidos, a Corte manteve a condenação criminal de réus que distribuíam panfletos contrários ao recrutamento militar à guerra contra a Rússia. Ver: Anthony Lewis, *Make No Law: The Sullivan case and the First Amendment,* Vintage Books, 1991, pp. 80-89.

em debate público; e ameaças à liberdade de expressão identificadas exclusivamente com governos, cuja não-interferência seria uma garantia suficiente de um mercado de ideias saudável. "Cada uma dessas premissas se tornou, de um jeito ou de outro, obsoleta no século XXI, graças à ascendência dos mercados de atenção e às mudanças nas tecnologias de comunicação (...) Em outras palavras, se era difícil falar, hoje é difícil se fazer ouvir"[134].

O autor levanta a possibilidade de a jurisprudência da Primeira Emenda se tornar obsoleta para lidar com questões e problemas que surgem no século XXI, ao menos que sejam revistas algumas daquelas suas premissas ou lógicas. Ele se preocupa principalmente com o fato de a "produção barata de discursos" ser utilizada como uma "arma" tática para silenciar vozes contrárias[135], mencionado os exemplos de uso de "exército de *trolls*" que assediam pessoas em redes sociais (como ativistas ou jornalistas, por exemplo) ou mesmo a reorientação diversionista da atenção do público por meio de inflação proposital de determinado assunto para abafar a repercussão de outro, o que seria *uma maneira de silenciar vozes sem a necessidade de censura direta*:

> *"Em um mundo de atenção escassa, esses tipos de métodos são mais efetivos do que poderiam ser há algumas décadas (...). Em um ambiente assim, inundar a rede com um conteúdo ("flooding") pode ser tão efetivo quanto formas mais tradicionais de censura"*[136].

As grandes redes sociais têm um papel central nesse contexto de barateamento das expressões e de valorização da atenção, com todas as consequências que advêm desse cenário. *E o uso da filtragem algorítmica é o fator técnico que viabiliza esse papel*. As redes sociais facilitam a publicação de discursos por qualquer um, gerando a possibilidade

---

[134] Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University, Emerging Threats Essay*, publicado em 01/09/2017.

[135] "Os baixos custos para se expressar, de modo paradoxal, têm facilitado o uso de discursos como armas ('weaponized speech') enquanto ferramenta de controle de outros discursos. A triste verdade é que o discurso barato pode ser usado para atacar, assediar e silenciar tanto quanto pode ser usado para iluminar ou debater. E o uso de discurso como uma ferramenta para suprimir outros discursos é, por sua própria natureza, algo muito desafiador para a Primeira Emenda lidar. Diante desses desafios, a doutrina da Primeira Emenda parece, no melhor cenário, despreparada" - Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University, Emerging Threats Essay*, publicado em 01/09/2017.

[136] Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University, Emerging Threats Essay*, publicado em 01/09/2017. O autor menciona que os governos chineses e russo têm liderado as iniciativas de uso dessa última tática ("*flooding*") para abafar vozes contrárias ou assuntos controversos na internet, diminuindo a necessidade de uso constante de censura direta.

de audiências cativas formadas a partir das conexões interpessoais. Fazem isso, sobretudo, por um modelo de negócios desenvolvidos dentro do que Wu denomina "indústria da atenção", formada por atores "cujo modelo de negócios é a revenda da atenção humana" [137].

Isso porque o fomento do exercício individual (e gratuito) da liberdade de expressão significa que as postagens geradas, bem como as conexões interpessoais por detrás delas, sejam traduzidas em dados aptos a produzir receita publicitária. Por meio desse ciclo, as opiniões e ideias (e emoções) dos usuários tornam-se elas próprias mercadorias, quando processadas como dados que permitem a micro-segmentação publicitária – do mais *cotidiano marketing comercial* ao *marketing político de uma eleição nacional*[138]. E, para garantir que usuários passem a maior parte do tempo possível em seus ambientes, consumindo e postando conteúdos, enquanto se engajam com eles mediante reações (curtidas, comentários, compartilhamentos), as redes sociais customizam aquilo que lhes é visível com o uso da filtragem algorítmica – ou "bolhas de filtro" ("filter bubble"), nos termos de Wu – que se refere à "tendência dos mercadores ou corretores de atenção de maximizar suas receitas ao oferecer à audiência um pacote de informações altamente customizado, filtrado e desenhado a alcançar seus interesses preexistentes"[139].

Nesse contexto, a filtragem algorítmica, como operação técnica, é central para a governança de discursos mais ampla que essas redes sociais exercem, possibilitando *uma nova forma de edição de conteúdos baseada no binômio visibilidade/invisbilidade e*

---

[137] Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University, Emerging Threats Essay*, publicado em 01/09/2017.

[138] Esse "ecossistema de publicidade digital" – que nasceu com popularização da internet comercial – beneficiou-se do desenvolvimento de tecnologias de coleta e tratamento de dados de usuários, que permitem uma segmentação cada vez mais refinada de alvos publicitários. "É como se o espaço publicitário genérico ('anuncie aqui') fosse substituído por um mais específico ('anuncie aqui para mulheres moradoras do bairro de Moema, entre 30 e 35 anos, interessadas em moda e artigos de luxo')", comenta Dennys Antonially, ao traçar um histórico sobre a evolução desse sistema – *A arquitetura da Internet e o desafio da tutela do direito à privacidade pelos Estados Nacionais*, tese de doutorado apresentada à Faculdade de Direito da Universidade de São Paulo, 2017, pp. 21-33. Para uma descrição do microdirecionamento da propaganda política eleitoral e do célebre caso Cambridge Analytica, ver: Francisco Carvalho de Brito Cruz, Heloisa Massaro, Thiago Oliva e Ester Borges, "Internet e eleições no Brasil: diagnósticos e recomendações", relatório publicado pelo centro de pesquisas *Internetlab* (2019). Para problemas mais gerais sobre armazenamento de dados, privacidade, modelos de negócios de empresas de tecnologia, ver: Soshana Zuboff, *The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power*, Public Affairs, 2019; Roger McNamee, *Zucked: Waking up to the Facebook catastrophe*, Harper Collins, 2019.

[139] Tim Wu, "Is the First Amendment Obsolete?", *Knight First Amendment Institute – Columbia University, Emerging Threats Essay*, publicado em 01/09/2017.

customizando essa edição, no limite, a um nível individual. Como toda forma de edição ou curadoria, isso envolve escolhas e suas respectivas consequências. Por isso, *as escolhas sobre a quais tipos de conteúdo será dada visibilidade importam*, pois constituem um vetor essencial de regulação do mercado de ideias de cada plataforma. Não apenas determinam o que será visto, mas com isso também criam incentivos para que a produção ou consumo de conteúdo pelos usuários priorize determinadas características.

Uma das consequências mais debatidas sobre essa nova forma de organização da "dieta informacional" é a extensão de uma resultante polarização opinativa – que seria fruto da falta de exposição a opiniões diversas ou contrárias no interior dessas "bolhas de filtro"[140]. No mesmo sentido, as novas possibilidades de promoção de campanhas de desinformação nas redes sociais, abrangidas pelo popular termo "fake news"[141], podem ser melhor compreendidas como produção de "informações de combate" (ou hiper-partidárias), sempre no contexto de disputas por atenção, apoio ou engajamento[142]. A operação dos algoritmos (definição daquilo que terá visibilidade) importa para *o que se vê e o que se produz* em cada plataforma: o Youtube, em especial, tem sido o alvo mais recente no debate público por conta das alegações de que os algoritmos que baseiam seu sistema de recomendações (que passam automaticamente, em *autoplay*, ao final e cada

[140] Cass Sunstein se baseia em estudos de psicologia comportamental que analisam a dinâmica de radicalização de grupos e de "efeitos manada" para alertar que esse cenário apresenta riscos reais para sistemas democráticos que, pelo dever de zelo por uma dimensão republicana de deliberação pública, pressupõem a exposição de seus cidadãos a opiniões diversas e contraditórias, especialmente por meio de encontros fortuitos ao longo do debate público. É nesse sentido que Sunstein propõe que plataformas como o Facebook, por iniciativa própria, criem botões de "contato inesperado" ("serendipity") que sugiram acesso a notícias e opiniões "contrárias" ou "diversas" ao lado daquelas recomendadas por afinidade prévia pelo "feed" de notícias – *#Republic: divided democracy in the age of social media*, Princeton University Press, 2017, especialmente: pp. 1-30; 231-233.

[141] Ronaldo Porto Macedo Júnior, "Fake News: a novidade de dizer mentiras", *Revista de Jornalismo ESPM*, edição julho-dezembro 2018, pp. 42-47.

[142] Pablo Ortellado e Márcio Moretto Ribeiro, "O que são e como lidar com as notícias falsas", *Sur – Revista Internacional de Direitos Humanos Volume. 15, nº 27* (julho 2018), pp. 71-83. No texto, os autores analisam alguns exemplos de "notícias hiper-partidarizadas" repercutidas por veículos alternativos de direito e de esquerda no contexto da crise política brasileira no último par de anos. Eles também apresentam algumas definições correntes para o termo "notícias falsas", ressaltando que a maior parte das divergências recaem sobre dois aspectos: a) se deve incluir apenas conteúdo noticioso comprovadamente falso, ou também a outras técnicas de desinformação e engano, como exageros, omissões e informações tiradas de contexto, por exemplo e; b) se deve incluir apenas conteúdo falso intencional ou também equívocos factuais verificáveis. Em sentido semelhante: Francisco Carvalho de Brito Cruz, Heloisa Massaro, Thiago Oliva e Ester Borges, *Internet e eleições no Brasil: diagnósticos e recomendações*, Internetlab, 2019, pp. 12-14. Esses autores, contudo, abrem mão de uma definição mais precisa de "fake news" – que chamam de "alcunha simplificadora" – em favor de uma problematização mais geral sobre essas mudanças estruturais nas formas de produção e consumo de informações.

visualização) têm uma especial predisposição para levarem usuários a conteúdos extremados ou inflamatórios, com ênfase em teorias conspiratórias[143].

Se este é um tópico complexo, que levanta ramificações para vários outros problemas atuais dentro do tema mais amplo da governança privada de discursos feita pelas redes sociais, *importa para os fins deste capítulo destacar como esse gradiente de visibilidade/invisibilidade pode ser utilizado com o propósito específico de ocultar discursos – o que, na prática, aproxima-se (podendo se confundir) com o ato de derrubá-los.*

Isso fica claro quando se analisa alguns dos critérios operados pelo Facebook para *reduzir o alcance de conteúdos indesejados*. Essa estratégia passou a ser adotada com mais vigor pela empresa em 2016, ano em que começou a sofrer com cumulativas crises de imagem por conta de conteúdos considerados tóxicos, incluindo os debates sobre "fake news" durante a eleição presidencial americana daquele ano. Desde então, para além da moderação feita por retirada de conteúdos proibidos, a empresa assumidamente passou a *adotar uma estratégia de reduzir o alcance de conteúdos problemáticos que não violam suas regras de comunidade*[144].

Em 2018, o Facebook anunciou que seu algoritmo iria priorizar a exibição de veículos de mídia tidos como dotados de credibilidade depois de pesquisas realizadas pela empresa com usuários[145]. Pouco tempo depois, anunciou também que, entre eles, seriam priorizadas reportagens de veículos locais (regionais) a seus usuários[146]. Além disso, ao

---

143 Zeynep Tufekci, "Youtube, the great radicalizer", artigo publicado no jornal *The New York Times*, em 21/03/2018; "How Youtube drives people to the Internet's darkest corners", reportagem publicada pelo jornal *The Wall Street Journal*, em 07/02/2018; "How Youtube radicalized Brazil", reportagem publicada pelo *The New York Times*, em 11/08/2019; "How Youtube misinformation resolved a Whatsapp mistery in Brazil", reportagem publicada pelo *The New York Times* em 15/08/2019. Para uma visão geral sobre o assunto, métodos de pesquisa e demais referências bibliográficas específicas ao Youtube, ver as seguintes publicações de Jonas Kaiser e Adrian Rauchfleisch: "Unite the right? How Youtube´s recommendation algorithm connects the U.S. far right", artigo publicado em *Medium*, em 11/04/2018; "The implications of venturing down the rabbit hole", artigo publicado em *Internet Policy Review: Journal of Internet Regulation*, em 27/06/2019.

144 "Desde 2016, nós usamos uma estratégia chamada 'remover, reduzir e informar' para gerenciar conteúdo problemático no Facebook. Isso envolve remover conteúdo que viola nossas regras de comunidade, reduzir o alcance de conteúdo problemático que não viola nossas regras e informar as pessoas com informações adicionais para que eles possam escolher quando clicar, ler ou compartilhar. Nosso trabalho de 'redução' [de alcance] é em grande parte centrado no 'News Feed' e em como priorizamos as postagens nele" – "People, Publishers, the Community", *Facebook Newsroom*, artigo publicado em 10/04/2019.

145 "Helping ensure News on Facebook is from trusted sources", *Facebook Newsroom*, publicado em 19/01/2018.

146 "More local news on Facebook", *Facebook Newsroom*, publicado em 19/01/2018.

final daquele ano, Mark Zuckerberg anunciou que conteúdos "fronteiriços" ("borderline") em relação às regras de comunidade – ou seja, que se *aproximassem* de regras de proibição contra discursos de ódio de ou agressão, por exemplo – também teriam sua visibilidade reduzida[147].

O Facebook também declarou que o combate à desinformação e às notícias falsas seria uma prioridade: desde então, a empresa mantém parceria com diversos "fact checkers" (organizações, normalmente com credenciais jornalísticas), em vários países, que reportam à empresa postagens que considerem problemáticas – sejam de conteúdo falsificado de modo intencional, ou deturpado por falta de contexto, etc. Nessa área, um sistema de inteligência artificial também tenta identificar essas postagens por meio de frases ou palavras normalmente usadas por "caça-cliques" ("clickbaits"). O sistema de inteligência artificial se desenvolve cada vez mais na medida em que é usado, inclusive com informações de conteúdos reportados por "fact checkers". Em regra, conteúdos assim ou páginas responsáveis também têm seu alcance reduzido. No final de 2018, o Facebook anunciou que essas medidas seriam adotadas também com relação a fotos e vídeos[148].

Em 2019, uma nova mudança do algoritmo do *News Feed* iria diminuir o alcance de postagens que – segundo uma nova métrica denominada "*Click-Gap*" – tivessem um número muito elevado de acessos no Facebook em comparação à internet em geral, pois esse seria um indicativo de páginas sem credibilidade com conteúdos feitos apenas para se tornar viral na plataforma[149]. Em julho desse ano, também para "melhorar a qualidade das informações no *News Feed*", o Facebook anunciou que iria diminuir o alcance de postagens que fizessem alegações sensacionalistas ou exageradas na área de saúde, bem como de postagens que buscassem vender produtos ou serviços com base em argumentos

[147] Mark Zuckerberg, "A Blueprint por content governance and enforcement", artigo publicado pelo *Facebook Newsroom*, em 15/11/2018. Sobre isso, Zuckerberg argumentou que "de modo interessante, nossas pesquisas demonstram que esse padrão natural do fato de conteúdo fronteiriço ('borderline') conseguir mais engajamento [dos usuários] se aplica não somente para notícias, mas a quase todas as categorias de conteúdo. Por exemplo: fotos próximas da linha de nudez, como aquelas que revelam roupas ou posições sexualmente sugestivas, ganhavam mais engajamento em geral antes de nós mudarmos a curva de distribuição para desencorajar isso. O mesmo vale para postagens que não alcançavam nossa definição de discurso de ódio, mas ainda assim eram ofensivas".

[148] "Expanding Fact-Checking to Photos and Videos", *Facebook Newsroom*, publicado 13/09/2018.

[149] "Remove, Reduce, Inform: New steps to manage problematic content", *Facebook Newsroom*, publicado em 10/04/2019.

de saúde (um exemplo mencionado é o de pílulas que prometem emagrecimento). Uma página identificada com postagens desse tipo pode ter seu alcance reduzido em geral[150].

Esses são alguns exemplos que demonstram como o Facebook tem avançado sua política de se utilizar da filtragem algorítmica para *aumentar o nível de invisibilidade de conteúdo considerado problemático* – e como essa *possibilidade técnica pode implementar novos tipos de regras editoriais*, de maneiras antes impossíveis para antigos intermediários.

2.G– Considerações finais do capítulo

As grandes redes sociais não são ambientes neutros, onde prevaleceria uma irrestrita possibilidade de publicação por indivíduos. Se é verdade que há uma liberdade sem precedentes de publicação praticamente imediata em seus ambientes, ela convive com *diversos e relevantes mecanismos de filtragem que operam entre os campos do permitido/proibido e do visível/invisível*. A derrubada de conteúdo pode ser uma intervenção mais extrema, mas o próprio funcionamento regular das redes sociais pressupõe que a plataforma escolha o nível de visibilidade/invisibilidade de veiculação das postagens, conforme sua *curadoria algorítmica*.

Esse é o ponto chave de tensão inerente ao problema desta tese: usuários possuem a expectativa (normalizada) de publicação irrestrita na internet (ou um direito de liberdade de expressão "a priori"), mas quando essas publicações ocorrem no ambiente desses intermediários (grandes redes sociais), estão sujeitas a regras substantivas de regulação de discursos, implementadas com uso dessas novas tecnologias.

Logo, este capítulo cumpre um primeiro objetivo de apresentar as operações técnicas das estruturas de controle de discursos das redes sociais, *iluminando uma esfera de poder privado sobre a liberdade de expressão que ainda é razoavelmente desconhecida* e cujas informações têm se tornado públicas há cerca de um par de anos.

Isso porque essa curadoria/moderação de conteúdos era feita até há pouco tempo de modo bastante opaco e longe do escrutínio público. Se no caso de antigos intermediários (como jornais impressos ou canais de televisão) poderia ser evidente que o conteúdo veiculado era fruto de escolhas (que implicavam a escolha de não-publicação do que ali não estava), essa mesma lógica, também presente nas grandes redes sociais,

[150] "Adressing Sensational Health Claims", *Facebook Newsroom*, publicado em 02/07/2019.

ainda que de modo qualitativamente diverso, acaba sendo mais dissimulada. Afinal, decisões de "não moderar" um conteúdo são, a rigor, decisões de moderação. Como aponta Tarleton Gillespie:

> "*Moderação é um prisma para compreender o que plataformas são, bem como as maneiras com que sutilmente direcionam a vida pública. Nossa compreensão de plataformas, tanto especificamente quanto como categoria conceitual, tem largamente aceitado os termos com que são vendidas e celebradas por suas próprias direções: abertas, imparciais, conectivas, progressistas e transformadoras (...). E se moderação é algo central ao que plataformas fazem, e não um aspecto periférico? (...). Moderação é, em muitos aspectos, a commodity que plataformas oferecem*"[151].

Essa descrição da nova governança privada de discursos pelas redes sociais traz consigo novas lógicas – processos, dinâmicas e, literalmente, novas engenharias – que impactam diretamente a liberdade de expressão na internet para centenas de milhões ou bilhões de usuários.

No caso da capacidade de remoção de conteúdo publicado por usuários, essas tecnologias implicam *novas possibilidades de controles de discursos* que, mais do que não terem paralelos com atores privados pretéritos, rivalizam (ou chegam a superar) as próprias capacidades de regulação de discursos de estados nacionais. A arquitetura global da internet e do próprio funcionamento dessas plataformas possibilita que *atuem com*

---

[151] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 13. O autor utiliza o termo "moderação" referindo-se primordialmente à política de permissão ou proibição de discursos, embora tangencie em sua obra também as escolhas algorítmicas que levam à visibilidade ou invisibilidade de postagens. Seu argumento, de qualquer modo, é válido no sentido da *curadoria* que a governança privada de discursos descrita neste capítulo promove. Para ele, "a moderação está lá desde o começo, e sempre; ainda assim, deve permanecer desautorizada, escondida, em parte para manter a ilusão de uma plataforma aberta e em parte para evitar uma responsabilização legal e cultural. Plataformas encaram o que pode ser uma contradição irreconciliável: são representadas como mero conduítes e também são baseadas no fato de que fazem escolhas sobre o que usuários dizem ou vêm. Ver a moderação sob esse aspecto poderia alterar a maneira como vemos o que as plataformas de redes sociais realmente fazem: de transmitir o que postamos a constituir o que vemos. Não há posição de imparcialidade" (p. 21); "quando a moderação é mantida tão invisível quanto possível, o conteúdo que vemos parece que está simplesmente ali, um fenômeno natural – incluindo qualquer tipo de material racista, homofóbico, obsceno ou abusivo que seja permitido a ficar no ar. A opacidade esconde não apenas o fato da seleção, mas também os valores que motivam a seleção" (p. 124). Por isso, "dado o escopo e variedade desse trabalho, nós deveríamos parar para refletir quando plataformas continuam apresentando a si próprias como abertas e desimpedidas ao fluxo de participação dos usuários" (p. 136).

significativo grau de autonomia com relação um determinado ordenamento jurídico, criando as *condições para dinâmicas competitivas entre seus sistemas normativos e de estados nacionais* (Capítulo 2-C).

Além disso, as grandes redes sociais lidam com um problema fundamental de escala. O que leva esses atores a ter de lidarem com problemas de moderação de discursos a partir de uma abordagem nada artesanal. Como atuam globalmente e lidam com um volume de discursos sem precedentes[152], as respostas de moderação de conteúdo só podem ser construídas com soluções técnicas que enfrentem problemas nesse novo nível de escala – *soluções que já levam em conta a contabilização de uma margem aceitável de erros*. Uma fração aceitável de erros (tirar do ar um conteúdo que deveria ter permanecido; ou manter no ar um conteúdo que deveria ter sido retirado) pode significar milhares de postagens. Nesse exato sentido:

> "*Plataformas de redes sociais lutam com uma tensão fundamental. Não importa como elas lidem com a moderação de conteúdo, quais são suas políticas e suas premissas, quais táticas elas escolhem; isso deve ser implementado em uma escala de dados ('data scale'). Essas plataformas massivas devem tratar seus usuários como dados, subpopulações e estatísticas, e suas intervenções devem ser semiautomatizadas de modo a acompanhar o interminável ritmo de violações e de reclamações. Não se trata de atendimento ao cliente ou de administração de comunidades, mas de logística – pela qual conteúdo, os usuários e as preocupações devem ser distribuídos e abordados de modo procedimental*"[153].

---

[152] "Dada a escala na qual o Twitter está, uma chance de um em um milhão acontece quinhentas vezes por dia. É o mesmo que ocorre com outras empresas lidando com esse tipo de escala. Para nós, os casos limite ("edge cases"), aquelas raras situações que são improváveis de acontecer, são basicamente normais. Imagine que 99.9% dos tweets não impliquem nenhum risco a alguém. Sem o envolvimento de qualquer ameaça. Quanto você retira esses 99.9%, aquela minúscula porcentagem de tweets remanescentes acabam somando cerca de 150 mil tweets ao mês. A magnitude da escala com que lidamos cria um desafio" – Del Harvey, vice presidente do Twitter para "Trust and Safety", *TED Talk "Protecting Twitter Users (Sometimes from Themselves)*", março de 2014; Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 74.

[153] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 141. Em entrevista ao autor, um funcionário da área de "content policy" do Facebook avalia que "a imensa escala das plataformas roubou, de qualquer uma das pessoas familiares com a torrente de relatos que chegam [aos revisores de conteúdo], a ilusão de que há algo como um caso único... Em qualquer rede social suficientemente grande tudo o que você pode possivelmente imaginar acontece toda semana, certo? Então não há situações hipotéticas, e não há casos que sejam diferentes ou realmente limítrofes ('edgy'). Há apenas casos mais ou menos frequentes, todos acontecendo ao mesmo tempo" (p. 77).

O problema de escala significa que deve haver um grau fundamental de automatização de interpretação e aplicação de regras. Mas o uso de ferramentas tecnológicas – potencialmente dotadas de inteligência artificial – não dispensa, por ora, o trabalho de equipes de milhares de revisores humanos. Se por um lado, a automatização é necessária para fazer frente ao volume de postagens publicadas, o fator humano também é, por ora, igualmente indispensável: somente uma pessoa pode interpretar o contexto de um texto ou de uma imagem para tentar identificar seu real significado e pertinência diante das regras pré-estabelecidas. Uma postagem com uma imagem de um ataque terrorista pode veicular uma crítica ou uma exaltação, a depender das palavras que a acompanham. Uma frase literalmente absurda pode ser fruto de uma intenção irônica. Uma ameaça à primeira vista pode ser, de fato, apenas uma figura de linguagem – ou realmente uma ameaça.

Nesse cenário, qualquer resposta normativa ou regulatória por parte do direito – sendo ele apenas um dos quatros modais regulatórios da internet – deve levar em consideração *a complexidade específica desse novo contexto tecnológico*. Como Lessig adverte, "a lição mais importante sobre o direito no cyberespaço é a necessidade de se levar em conta o efeito regulatório do código"[154]. Dos riscos embutidos nesse efeito regulatório do código, mas também das necessidades legítimas que se busca enfrentar.

[154] Lawrence Lessig, *Code: version 2.0*, Basic Books, 2006, p. 155.

Capítulo 3 – Aspectos substantivos da moderação de conteúdo pelas redes sociais: dos anos iniciais às atuais encruzilhadas valorativas e editoriais

No capítulo anterior, foram apresentadas as principais formas como as redes sociais operam a governança de discursos em suas plataformas – mantendo o enfoque em suas políticas de moderação de conteúdo, que implicam decisões sobre a derrubada ou ocultação de postagens. Ao final, essa descrição sobre o funcionamento operacional das redes sociais levou à afirmação de que elas não são meros espaços neutros de veiculação de discursos de terceiros.

Este capítulo irá avançar com aquela análise, desta vez abordando *aspectos substantivos de regras de moderação de conteúdo*. A partir das recentes publicações acadêmicas e das informações públicas disponíveis, seu primeiro tópico vai apresentar os anos iniciais de estabelecimento dessas regras pelo Facebook (principalmente), Youtube e Twitter.

Em seguida, serão apresentados tópicos voltados a aspectos particulares da política de moderação do Facebook – suas regras para discursos de ódio e também para proteção ao debate público – por considerar que são recortes representativos e reveladores de temas, problemas e dilemas que geralmente surgem na formulação e aplicação dessas regras pelas plataformas.

Por fim, com base nessa análise, será desenvolvido o argumento segundo o qual a governança privada de discursos pelas redes sociais deve ser compreendida como um novo tipo de liberdade editorial.

3.A – Os anos iniciais: da aplicação de "standards" genéricos à construção de um sistema de regras

O gigantismo que hoje caracteriza as redes sociais era certamente inesperado para elas próprias – e recuperar o histórico da fase inicial da política de moderação do Facebook, com comparações adicionais aos casos do Youtube e do Twitter, revela como esse processo foi acidentado e construído à medida em que se tornava cada vez mais necessário. Há poucos anos, sequer estava claro para essas empresas o tamanho do desafio que elas iriam encarar ao se tornarem responsáveis pela regulação de discursos de multidões de milhões de pessoas.

Em seus anos iniciais, as plataformas possuíam uma abordagem que se assemelhava à então tradicional moderação realizada em páginas de comentários ou fóruns de internet – que costumam agregar centenas ou poucos milhares de usuários. Os departamentos responsáveis pela revisão de conteúdos publicados eram ínfimos se comparados às práticas atuais. Além disso, esses parâmetros eram poucos, genéricos e limitados[155].

No caso do Facebook, até 2009 não havia a publicação de quaisquer "community standards" a seus usuários. Nesse período, a política interna da empresa "tinha cerca de uma página; uma lista de coisas que deveriam ser deletadas: coisas como Hitler e pessoas nuas". A moderação de conteúdo era feita com base em um senso comum intuitivo: "se te faz se sentir mal em seu estômago, apague", descreveu uma funcionária do setor naqueles anos iniciais. A regra "*Sente-se mal? Então apague*" seria o cerne do treinamento sobre moderação de conteúdo até a formação de um grupo especializado, ao final de 2009[156]; o grupo, com cerca de doze pessoas, teria a tarefa de sistematizar o que se tornaria a primeira versão das "regras de comunidade".

O modelo inicial tornava-se cada vez mais insustentável na medida em que a plataforma deixava de se limitar a um público mais homogêneo de universitários americanos, tornando-se cada vez mais global e diversificada. Como disse Dave Willner, um dos responsáveis pela área na empresa:

> "*nos anos iniciais nós tínhamos muitas políticas que determinava 'derrube todas as coisas ruins. Derrube as coisas que sejam grosseiras, racistas, ou bullying'. Esses são conceitos importantes, mas são julgamentos de valor. Você tem que ser mais granular e menos abstrato do que aquilo. Porque se você diz para quarenta universitários [que atuam como moderadores], 'apague todo o discurso racista', eles não vão concordar uns com os outros sobre o que é racista*"[157].

[155] Introdução – Tópico 3.

[156] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1631.

[157] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1633.

Willner baseou seu primeiro conjunto de regras de conteúdo para o Facebook tendo como modelo os códigos contra assédio ("anti-harassment codes") de sua universidade, já que era um profissional recém-formado. Mas percebeu que "standards" vagos – como a proibição de discursos que criem um "ambiente hostil" – levavam a dúvidas e decisões divergentes entre os moderadores, que àquela altura já se espalhavam pela Califórnia, Texas, Irlanda e Índia, envolvendo, portanto, diferentes contextos culturais. Ciente da oposição entre as concepções norte-americana e europeia sobre os temas da liberdade de expressão e do discurso de ódio, Willlner passou a buscar uma política contra discursos de ódio "que focassem em ações concretas, facilmente categorizáveis", de modo que a decisão a respeito de eventual remoção, "pudesse ser baseada em nada além das informações contidas no formulário que usuários do Facebook se utilizam para fazer uma reclamação sobre posts ofensivos, aplicadas como um algoritmo". Ou, em resumo: "ele procurou uma resposta de um engenheiro para um problema histórico e jurídico espinhoso – uma abordagem ao estilo do Vale do Silício"[158].

A primeira versão das regras do Facebook tinha quinze mil palavras; "seu objetivo era consistência e uniformidade: obter o mesmo julgamento em um determinado conteúdo, independentemente de quem fosse a pessoa responsável pela moderação"[159]. Desde então, essa não tem sido uma empreitada fácil. Ao chamar para si a tarefa de criar um sistema de regras de discursos com a pretensão de aplicá-lo globalmente, o Facebook abraçava dilemas e controvérsias que há muito pautavam debates acalorados a respeito da liberdade de expressão:

> *"Arte não existe como uma propriedade de imagem. Não há 'pixels' de arte que você pode encontrar nas imagens que consideramos belas ou inspiradoras... O que descobrimos sobre arte é que as questões de moderação de conteúdo que surgiam sobre arte não eram sobre a arte em si, mas quando ela poderia se tornar uma exceção a uma restrição previamente existente. Então não há problemas com a maior parte da arte. Mas é quando você discute casos que podem chegar em definições de nudez, racismo ou violência, por exemplo, que pessoas se importam"*[160].

---

[158] Jeffrey Rosen, "The Delete Squad", artigo publicado pelo *The New Republic*, em 29/04/2013.

[159] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1633-1634.

[160] Dave Willner, funcionário do Facebook à época – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1634.

Ou ainda:

> "*Nós não podíamos fazer uma política que simplesmente proibisse o uso da "N-Word'. Isso poderia ser completamente insensível para a comunidade afro-americana nos Estados Unidos. Mas você também não deseja que ela seja usada como discurso de ódio. Então é quase impossível tornar isso uma decisão objetiva, porque o contexto importa muito"*[161].

Essa abordagem não seria muito diferente daquela aplicada na mesma época pelo Youtube. A advogada Nicole Wong, especializada em direito da Primeira Emenda, começou a trabalhar para o Google em 2004; dois anos depois, a empresa iria adquirir o Youtube. Conforme seu relato, um problema que diagnosticou à época era que "essas novas corporações ainda se enxergavam como empresas de computação ('software companies') – e não refletiam sobre os efeitos permanentes gerados sobre a expressão ('speech') como parte de suas atividades". No Youtube, ela foi responsável por implementar a política de moderação de conteúdo sob uma base de proteção à liberdade de expressão: nenhum conteúdo lícito seria removido ao menos que violasse regras da plataforma[162].

Nesses anos iniciais, contudo, três episódios obrigaram o Youtube a tomar decisões de relevância pública – e que iriam se tornar paradigmáticas para demonstrar que, ao se lidar com a regulação de discursos, normas pré-estabelecidas de modo muito genérico não seriam sempre suficientes para tomadas consistentes de decisões, diante do peso que a análise de contexto frequentemente possui em casos que envolvam a liberdade de expressão.

No final de 2006, surgiram dois vídeos no Youtube logo depois da morte do ditador iraquiano Saddam Hussein. Um deles retratava seu enforcamento, após sua

---

[161] Jud Hoffman, funcionário do Facebook à época – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1634.

[162] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1618. Para Jeffrey Rosen, "no despertar da era da internet, alguns dos maiores 'players' da indústria nascente advogavam com fervor uma política de não intervenção em discursos de ódio; Wong foi parte de uma geração que descobriu em primeira mão que essa posição libertária extrema era insustentável" – "The Delete Squad", artigo publicado pelo site *The New Republic*, em 29/04/2013.

captura em um esconderijo, no rastro da invasão militar de seu país. O segundo vídeo retratava seu corpo no necrotério. Pelas regras da plataforma, os dois vídeos deveriam ser proibidos, em razão da exibição de atos violentos e relacionados à morte de uma pessoa. "Mas nós decidimos manter o vídeo do enforcamento, porque sentimos que sob uma perspectiva histórica, ele tinha real valor", disse Wong[163].

No ano seguinte, o Youtube removeu um vídeo que mostrava um homem sendo espancado no interior de uma cela, em função da violação da regra de exibição de violência gráfica. Posteriormente, no entanto, o vídeo foi restabelecido após a empresa receber a informação de que o vídeo tinha sido postado por um ativista de direitos humanos egípcio, que buscava denunciar abusos policiais pelas forças de governo daquele país. Em 2009, o Youtube também manteve no ar um vídeo da morte, após momentos agonizantes, de um manifestante iraniano alvejado no peito por forças de segurança do governo[164].

No caso do Youtube, a empresa empregava cerca de sessenta pessoas em meados de 2006, responsáveis por revisar todos os vídeos denunciados ("flagged") por usuários, pelas mais diversas razões. Para casos de violações dos termos de uso, dez funcionários trabalhavam utilizando *uma diretriz de uma única página*, contendo uma lista de "standards" com materiais proibidos – tais como abuso animal, vídeos mostrando sangue, nudez e pornografia. Alguns meses depois, as diretrizes alcançariam seis páginas. Cinco anos depois, em 2011, o Youtube gerenciava o dobro de volume de vídeos; a essa altura, a equipe de moderação de conteúdo tinha sido expandida e terceirizada – além disso, ela era responsável por aplicar regras mais precisas e pré-definidas. Esse conjunto de regras seria regularmente comentado e atualizado internamente, por meio de sucessivas novas edições[165].

Por isso, o movimento mais marcante desse estágio inicial do crescimento das grandes redes sociais, como aponta Kate Klonick, é mudança de um paradigma de

[163] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1618.

[164] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1619-1620.

[165] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1632-1633.

governança orientado por "standards" normativos para um novo – e cada vez mais complexo – sistema baseado em regras.

Um exemplo de "standard" seria "é proibido dirigir sob velocidade excessiva". Uma regra, por outro lado, definiria um limite de velocidade máxima – como 80 quilômetros por hora, por exemplo. "Há ganhos e custos ('trade-offs') em escolher uma forma de resolução sobre a outra. 'Standards' normalmente reforçam propósitos ou valores, mas porque são geralmente vagos e abertos, podem estar sujeitos a uma aplicação arbitrária ou preconceituosa por parte de tomadores de decisões", diz Klonick. Essa categoria, contudo, pode ser mais eficiente para acomodar diferentes circunstâncias, especialmente não previstas:

> *"Regras, por outro lado, implicam questões reversas: são comparativamente mais baratas e fáceis de aplicar, mas podem ser adaptáveis demais ou de menos a algumas situações e, por isso, podem levar a resultados injustos (...). Regras permitem pouca discricionariedade e nesse sentido limitam os ímpetos de tomadores de decisões, mas também contêm lapsos e conflitos, criando complexidades e litigância"*[166].

A autora credita essa *mudança qualitativa* a três fatores principais: a) o vertiginoso aumento tanto de usuários, quanto de conteúdo publicado; b) a globalização e diversificação da "comunidade" da plataforma; c) a crescente dependência de moderadores humanos com perfis culturais e linguísticos diversos[167].

Nesse cenário no qual as demais plataformas já desenvolviam suas políticas de moderação de conteúdo[168], o Twitter anunciava em seus anos iniciais que iria incorporar para si o "ethos" da tradição libertária norte-americana da liberdade de expressão e, com isso, tolerar praticamente todos os tipos de expressão discursiva. Na prática, isso

[166] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1631-1635. Para Klonick, "esse desenvolvimento observado no Youtube e Facebook de [um sistema de] standards para [outro de] regras na política de moderação de conteúdo reflete esses 'trade-offs'".

[167] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1631-1635.

[168] Danielle Keats Citron & Helen Norton, "Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age", *Boston University Law Review Volume 91* (2011), pp. 1453-1456.

significava que, em seu âmbito interno, o Twitter não iria ainda iniciar o desenvolvimento de um processo interno estruturado para revisão e eventual retirada de conteúdos do ar[169].

Em 2012, um alto executivo do Twitter descreveu a própria empresa como sendo "a ala mais libertária do partido da liberdade de expressão" ("the free speech wing of the free speech party")[170]. Afinal, a rede era notoriamente conhecida por assumir uma postura praticamente absolutista de não intervenção sobre os conteúdos postados por seus usuários. O Twitter anunciava publicamente sua postura de ser um mercado de ideias de que operava declaradamente sob o marco do "laissez faire". E que, em seus anos iniciais, advogava o otimismo convicto de que dar voz plena a todas as pessoas era parte da revolução tecnológica que aprofundava a causa democrática. Foi nesses anos, por exemplo, que a plataforma alcançou a reputação de ter sido um instrumento relevante de levantes populares e de aspirações democráticas durante a Primavera Árabe, especialmente ao longo de 2011.

Essa postura remontava à própria gênese da empresa – que tinha entre seus fundadores e primeiros diretores ex-funcionários da popular plataforma Blogger, de propriedade do Google. Ao fundarem o Twitter, eles buscaram transmitir para a nova rede social o prévio e forte compromisso "com o direito universal de publicar, a despeito de críticas externas"[171]. Embora a mentalidade tenha sido transplantada para a nascente rede social, o tempo deixaria evidente que *a dinâmica social do Twitter não poderia ser comparada com a velha realidade da blogsfera*, na qual páginas de blogs individuais, esparsas e conectadas apenas por links de interesse podiam lidar com a muito mais simples moderação de comentários feitos por visitantes, normalmente ao rodapé do texto principal.

Em poucos anos, o Twitter cresceu vertiginosamente – mas ao alcançar uma escala global e gigante, passou a sofrer cada vez mais com crises de imagem, até se

---

169 "A devoção a um padrão fundamental de liberdade de expressão refletia-se não apenas no fato de o Twitter não monitorar o conteúdo de usuários, mas também no que fazia para protegê-los". Klonick explica que a gestão jurídica da empresa, caracterizava-se também pela forte resistência a pedidos de governos para a retirada de conteúdo ou para obtenção de informações de usuários – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1621.

170 "Twitter's Tony Wang: "We are the Free Speech Wing of the Free Speech Party", reportagem publicada pelo *The Guardian*, em 22/05/2012.

171 "A Honeypot for assholes: inside Twitter's 10 years failure to stop harassment", reportagem publicada pelo *Buzzfeed News*, reportagem publicada em 11/08/2016.

deparar com a estagnação de seu número de usuários, basicamente por ser identificado como um ambiente tóxico, cheio de mensagens cada vez mais agressivas, raivosas e, bem, de ódio[172]

Uma mudança de rumo se tornou inevitável[173]. Em maio de 2016, o Twitter já havia se juntado ao Facebook, Microsoft e Youtube, quando – em meio à crescente pressão europeia, na tradição daquele continente de forte presença de leis contra discursos de ódio – as quatro gigantes aderiram ao "Código de Conduta da União Europeia para combate ao discurso de ódio online ilegal"[174]. No final de 2017, um alto executivo da empresa deu a seguinte declaração a políticos britânicos, em meio à pressão e cobranças de autoridades europeias a representantes dessa empresa – e também do Google e Facebook – sobre a "proliferação do ódio" na internet:

> "*Eu olho para trás durante o período de 5 anos e meio – e as respostas que eu daria a essas questões, cinco anos atrás, seriam bem diferentes. O Twitter estava em um lugar no qual acreditava que o antídoto mais efetivo ao mau discurso ('bad speech') era o bom discurso ('good speech'). Era basicamente uma filosofia estilo John Stuart Mill. [Mas] nós percebemos que o mundo em que vivemos mudou. Nós tivemos que embarcar em uma jornada com ele, e*

[172] "Por anos, o Twitter vinha sendo criticado por permitir que uma cultura de assédio se alastrasse livremente em seu serviço, particularmente atacando mulheres, mas também comunidades LGBTQ, minorias étnicas e raciais, participantes de várias subculturas, e figuras públicas. Isso era mais do que conversas duras ('harsh talk') ou insultos pessoais. Isso era discurso misógino e discurso de ódio, ameaças explícitas de estupro e de violência, ataques concertados e incessantes contra indivíduos, além de doxxing (postar as informações privadas de uma vítima como uma ameaça velada, ou um convite a outros para que também ameacem) - Tarleton Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 24.

[173] "A transformação do Twitter de herói da internet, por sua forte recusa a policiar o conteúdo de usuários, a vilão da rede aconteceu de modo relativamente célere. Embora a atenção pública sobre assédio e discursos de ódio no ambiente online já fosse crescente, a controvérsia Gamergate em 2014 atingiu novos níveis de atenção mundial sobre o tema. Como a plataforma menos policiada ou regrada, muito da culpa caiu sobre o Twitter. Em 2015, a mudança em valores culturais e expectativas começou a se refletir em novas regras e políticas públicas no Twitter. O site adicionou nova linguagem proibindo a 'promoção de violência contra outros... com base na raça, etnia, origem nacional, religião, orientação sexual, gênero, identidade de gênero, idade ou deficiência'" - Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1628. Em meio às discussões corporativas internas, um diretor partidário de uma maior moderação de conteúdo teria dito: "se as coisas forem do jeito que vocês querem, as únicas pessoas que vão usar o Twitter serão do ISIS [Estado Islâmico] ou da ACLU [American Civil Liberties Union]" – "A Honeypot for assholes: inside Twitter's 10 years failure to stop harassment", reportagem publicada pelo *Buzzfeed News*, reportagem publicada em 11/08/2016.

[174] Para uma contextualização sobre a crescente pressão europeia sobre grandes plataformas americanas quanto aos temas de discurso de ódio e combate ao terrorismo (especialmente a partir de 2015), ver: David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 65-74. Esse movimento contou com um protagonismo do governo alemão, que em 2017 culminou com a promulgação da legislação conhecida como NetzDG, que será analisada no Capítulo 4-E.

> *percebemos que não é mais possível defender todos os discursos ('stand up for all speech') com a esperança de que a sociedade irá se tornar um lugar melhor porque o racismo será contestado, a homofobia será contestada ou o extremismo será contestado. E nós temos sim que tomar medidas para limitar a visibilidade de símbolos de ódio, para banir pessoas da plataforma afiliadas com grupos violentos – esta é a jornada em que estamos agora"*[175].

Isso não seria o fim da questão: pelo contrário, seria apenas o começo do tortuoso processo de chamar para si a responsabilidade de traçar linhas aceitáveis para discursos – um processo interminável. Ainda hoje, o Twitter desenvolve suas regras básicas no campo de "discursos de ódio": uma proposta recente que buscava proibir todas as formas de discursos "desumanizadores" (comparação de pessoas com insetos, animais, objetos fétidos e afins), depois de debate aberto a considerações de usuários e de discussões internas na empresa, terminou restrito apenas para insultos dirigidos a grupos religiosos. A empresa sustenta que esse é um passo inicial e que o Twitter pretende ampliar suas regras contra discursos desumanizadores para demais áreas[176]. De qualquer modo, *a experiência libertária do Twitter fracassou* – e a empresa também aderiu, com alguns anos de atraso, a uma política regulatória de discursos, calcada em um sistema de regras.

Como essa breve recapitulação dos anos iniciais de criação dos sistemas de regras de moderação de conteúdos pelas redes sociais já transparece, o tema da proibição a discursos de ódio é um dos mais centrais e polêmicos na seara da liberdade de expressão – e analisar como o Facebook formulou suas regras a respeito é, por si só, revelador da complexidade dessa atividade exercida pelas grandes redes sociais, bem como dos impasses que elas necessariamente enfrentam. Por isso, esse será o tema do próximo tópico.

3.B – A proibição do Facebook a discursos de ódio ("hate speech")

---

[175] "Twitter was once a bastion of free speech but now says it's 'no longer possible to stand up for all speech'", reportagem do *The Business Insider*, publicada em 19/12/2017. Na véspera, a empresa havia divulgado suas novas regras contra discursos de ódio – e, como companhia para o anúncio, realizou também uma onda de suspensão de perfis "xenófobos, racistas" e "afiliados com grupos de ódio ou discursos de ódio online ou off-line" – "Twitter suspended Britain First leaders Paul Golding and Jayde Fransen in a massive hate purge", reportagem do *The Business Insider*, publicada em 18/12/2017.

[176] "Twitter backs off broad limits on 'dehumanizing speech'", reportagem publicada pelo jornal *The New York Times*, em 09/07/2019.

Poucos temas no âmbito da liberdade de expressão levantam tantas controvérsias políticas, morais e jurídicas quanto as questões ligadas ao chamado "discurso de ódio". Conceito por si só de difícil e complicada definição, é também o centro de intermináveis disputas teóricas e mesmo culturais sobre os "limites" ou "fundamentos" daquele direito fundamental. Ao lidarmos com discursos extremistas, vislumbramos também como lidamos com os extremos daquela liberdade. Por essas razões, o "hate speech" é um tema conveniente e profícuo para apresentarmos os níveis de complexidade, detalhamento e controvérsias que emergem da política de moderação de conteúdo do Facebook, não raro por suas frequentes conexões com pontos de vista políticos. Além disso, os próprios ambientes das redes sociais têm apresentado ao problema dos "discursos do ódio" novas roupagens e preocupações, que serão abordadas a seguir.

Conceituar o "discurso de ódio" não é tarefa fácil – até porque inexiste uma definição universal a seu respeito[177]. A expressão normalmente carrega consigo o significado de vedação ou punição a discursos que promovam a discriminação de pessoas – seja em função de identificação por religião, raça, gênero, orientação sexual, nacionalidade ou critérios afins. Mas essa definição é apenas um ponto de partida para diversas controvérsias: mesmo entre países que proíbem de algum modo modalidades de discurso de ódio, os tratamentos dispensados por sistemas jurídicos nacionais variam consideravelmente[178]; pode-se discutir sobre se o que define essa categoria é exclusivamente um eventual teor discriminatório de seu conteúdo ou se a forma e o tom de sua apresentação são traços determinantes para isso[179]; pode-se debater se apenas

[177] Como aponta Robert Post, "todas as tentativas jurídicas de se suprimir o ódio, seja contra grupos raciais ou o Rei, enfrentam uma profunda dificuldade conceitual. Elas devem distinguir o ódio de desgostos ou discordâncias ordinárias. Até mesmo aqueles que acreditam que o ódio deve ser punido porque é 'extremo' aceitam que discordâncias, mesmo discordâncias que decorrem de antipatias ('dislike'), devem ser protegidas porque são a matéria prima ('lifeblood') da política*"* – "Hate Speech", *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010; p. 125. A menção ao "Rei" foi feita em razão do breve paralelo às leis de sedição que vigoravam na Inglaterra, punindo o "ódio" ao governante, em trecho anterior do artigo. E é oportuna, em geral, porque o debate em torno da punição a discursos de ódio levanta com frequência a questão sobre se essas regras podem acabar sendo usadas, na prática, para perseguir a oposição ou dissidentes políticos.

[178] Frederick Schauer, "The Exceptional First Amendment,", *in:* Michael Ignatieff (ed.), *American Exceptionalism and Human Rights*, Princeton University Press, 2005, pp. 33-36; para análise dos casos do Canadá e da França, por exemplo, ver: L. W. Summer, "Incitement and the Regulation of Hate Speech in Canada: a Philosophical Analysis" e Pascal Mbongo, "Hate Speech, Extreme Speech, and Collective Defamation in French Law", ambos *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford Univeristy Press, 2010. Entre outros países que possuem algum tipo de legislação do tipo, destacam-se Inglaterra, Holanda, Noruega, Suécia e África do Sul.

[179] Robert Post, por exemplo, entende que a categorização de "discurso de ódio" normalmente ocorre quando um conteúdo discriminatório também viola, adicionalmente, um conjunto de normas sociais de respeito mútuo que sejam observadas por uma comunidade. Nesse sentido, o autor problematiza se um

grupos historicamente marginalizados podem ser alvos de ódio que mereçam intervenção estatal ou se critérios como raça se aplicam sem considerações desse tipo, indistintamente a todos. E, no limite, como pano de fundo e raiz do problema, impõe-se a questão sobre se a democracia demanda uma postura de maior tolerância ou de intolerância a discursos de ódio – especialmente quando restritos ao âmbito de discursos *strictu sensu*, sem conexão direta com ações violentas ou atos discriminatórios concretos[180].

Para os objetivos desta tese, porém, bastam por ora essas breves referências à complexidade e a controvérsias que cercam o tema, a título introdutório deste tópico, com o intuito de abrir caminho para uma análise sobre a operacionalização do tema "discurso de ódio" pela política de moderação de conteúdo do Facebook.

A empresa desde cedo adotou a prática francamente dominante entre as plataformas norte-americanas de internet ao impor restrições ao que considera como discursos de ódio – e, como visto, constrói essas regras de modo que sejam aplicadas a todo o mercado global. Depois de reportagens jornalísticas tornarem públicas em 2017 as regras de moderação de conteúdo do Facebook em searas controversas[181], a empresa começou a tornar públicas, no ano seguinte, suas diretrizes e parte de seus critérios de moderação, até então restritos a seus ambientes internos[182].

---

ponto de vista apresentado com roupagem intelectual ou acadêmica (como um artigo defendendo uma correlação entre raça e criminalidade, por exemplo) não termina por sofrer menor grau de oposição social e baixa possibilidade de punição formal do que a opinião pouco articulada de uma pessoa comum do povo, proferindo ideias racistas em uma esquina – "Hate Speech", *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010; pp. 134-136.

[180] Esse é o debate central que emerge do tema "discurso de ódio" e do qual se ocuparam inúmeros autores. Como referência recente em meio à vasta bibliografia, registro a seguinte tese elaborada a partir do embate das visões de Ronald Dworkin e de Edwin Baker (que defendem fundamentos democráticos em favor de uma tolerância com relação a discursos de ódio) em contraposição a Jeremy Waldron (que defende a legitimidade de restrições com base em um conceito de dignidade dos cidadãos em uma democracia): Clarissa Piterman Gross, *Pode dizer ou não? Discurso do ódio, liberdade de expressão e a democracia liberal igualitária*, tese de doutorado apresentada à Faculdade de Direito da USP, 2017. Ainda a esse respeito, vide o debate entre Ronald Dworkin e Jeremy Waldron, realizado na conferência *Challenges to Multiculturalism: a conference on migration, citizenship and free speech*, na The House of Literature de Oslo, em junho de 2012 – vídeo disponível no em https://www.youtube.com/watch?v=DoSbp8pdbM8

[181] "Revealed: Facebook´s internal rulebook on sex, terrorism and violence", reportagem publicada pelo jornal *The Guardian* em 21/5/2017; "Social Media´s Silent Filter", reportagem publicada pelo site *The Atlantic* em 8/3/2017.

[182] Ao mesmo tempo, a empresa ampliou o processo de "recursos" ("appeal") contra suas decisões: "nós decidimos publicar essas diretrizes internas por duas razões. Primeiro, as diretrizes irão ajudar as pessoas a entenderem onde nós traçamos as linhas em questões nuançadas. Em segundo lugar, fornecer esses detalhes torna mais fácil para que todas as pessoas, incluindo especialistas em diferentes áreas, possam dar para nós um retorno sobre o que podemos melhorar nas regras – e também nas decisões que tomamos – ao longo do tempo"."Publishing our internal enforcement guidelines and expanding our appeals process", *Facebook Newsroom*, publicado em 24/04/2018.

Atualmente, as regras para o tema de "discursos de ódio" integram a categoria de "conteúdo questionável", com cinco divisões ao total: violência e conteúdo explícito, nudez e atividades sexuais, abordagem sexual, conteúdo cruel e insensível – e discursos de ódio. O Facebook apresenta da seguinte maneira o que chama de sua "racionalidade" para esta política em especial:

> *"Nós não permitimos discurso de ódio no Facebook porque isso cria um ambiente de intimidação e exclusão, além de em alguns casos poder promover violência no mundo real.*
>
> *Definimos discurso de ódio como um ataque direto a pessoas baseado no que chamamos de características protegidas – raça, etnia, nacionalidade, afiliação religiosa, orientação sexual, casta, sexo, gênero, identidade de gênero, doença grave ou deficiência. Nós também damos algumas proteções ao status de imigrantes. Nós definimos ataque como discurso violento ou desumanizador, afirmações de inferioridade ou chamadas para exclusão ou segregação. Nós dividimos os ataques em três níveis de severidade, como descritos a seguir.*
>
> *Às vezes as pessoas compartilham conteúdo contendo o discurso de ódio de uma terceira pessoa com o objetivo de chamar atenção para isso ou educar outras pessoas. Em alguns casos, palavras ou termos que violariam nossas regras são usadas de modo autorreferenciado ou de uma maneira empoderadora. Pessoas às vezes expressam desprezo no contexto de um fim de um relacionamento romântico. Outras vezes, elas usam linguagem exclusiva de gênero pra controlar a filiação em um grupo de apoio positivo ou de saúde, como no caso de um grupo exclusivo para mulheres que amamentam. Em todos esses casos, nós permitimos tais conteúdos mas esperamos que as pessoas indiquem claramente seu intuito, o que nos ajuda a melhor compreender porque isso foi compartilhado. Quando a intenção não for clara, nós podemos remover o conteúdo.*
>
> *Nós permitimos humor e comentário social relacionados a esses tópicos. Adicionalmente, nós acreditamos que as pessoas são mais responsáveis quando elas compartilham esse tipo de comentário usando sua identidade autêntica".*

Um primeiro ponto a destacar é que essa política se baseia no conceito de "categorias protegidas" – um aspecto que Tarleton Gillespie e Kate Klonick relacionam

à influência jurídica das construções de leis antidiscriminação norte-americanas[183]. São as plataformas de internet que decidem quais sãos as "categorias protegidas" por cada uma delas, muito embora haja uma correspondência geral. Isso significa que, no caso do Facebook, por exemplo, a empresa proíbe e remove conteúdo que interpreta como um ataque à categoria de "orientação sexual" (homossexuais ou heterossexuais, indistintamente), mas não a "classes sociais" (ricos ou pobres) ou "ideologias políticas" (conservadores, progressistas, socialistas ou anarquistas, por exemplo). Ou seja: na prática, é proibido dizer que "gays são nojentos" ou "homens heterossexuais são nojentos", ao passo em que é permitido dizer que "ricos são nojentos", "pobres são nojentos" ou "socialistas são nojentos".

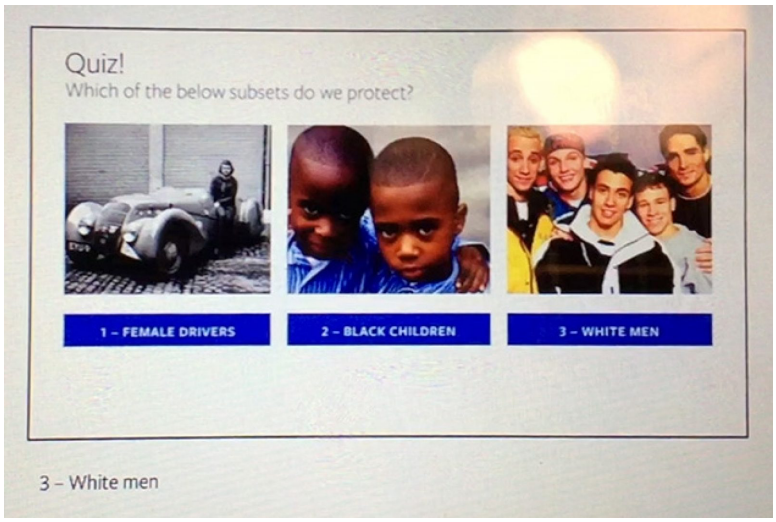
Nesse contexto, é de se reparar também que as "categorias protegidas" são consideradas pelo Facebook sem distinção dos grupos que a compõem; ou seja, a proteção não se destina exclusivamente a minorias ou grupos historicamente vulneráveis. Embora uma "categoria protegida" possa ser motivada pela existência de grupos vulneráveis (gays ou pessoas negras, por exemplo, que motivam as categorias de orientação sexual ou raça), a aplicação de seu critério no caso do Facebook vale para todas as pessoas que a integrem (ou seja, a regra vale igualmente para heterossexuais e pessoas brancas).

Em 2017, uma reportagem do portal ProPublica enfatizou supostas ou aparentes inconsistências da política da empresa para casos de discursos de ódio. Seu título de chamada era: "As regras secretas de censura do Facebook protegem homens brancos de

[183] "Na maior parte dos casos, as regras de discursos de ódio das plataformas ecoam a linguagem jurídica americana, particularmente pela listagem de grupos a quem tais regras se aplicam. Esta é a área mais óbvia onde a linguagem legal se sobressai ao tom mais casual das demais regras. Como no caso do Twitter, a maior parte [das plataformas] lista 'classes protegidas', grupos que gozam de proteção contra discriminação na legislação norte-americana, desenvolvida ao longo do tempo em uma série de leis antidiscriminação, do Título VII do *Civil Rights Act* of 1964 (que proíbe discriminação baseada em 'raça, cor, religião ou origem nacional'), o *Equal Pay Act* de 1963 (gênero), o *Age Discrimination in Employment Act de* 1967 (idade), o *Americans with Disabilites Act* de 1990 (deficiência), e etc" – Tarleton Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 59. Kate Klonick também caracteriza o uso de "categorais protegidas como uma influência da tradição legislativa norte-americana, em especial do *Civil Rights Act*" – "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1645.

discursos de ódio, mas não crianças negras". O texto apresentava a seguinte situação: logo após um atentado terrorista ocorrido em Londres, naquele ano, um congressista americano da Louisiana escreveu um post no Facebook no qual pedia pela morte de todos "muçulmanos radicalizados". "Cace-os, identifique-os e mate-os. Mate-os todos, em nome de tudo o que é justo e certo. Mate-os todos.", publicou o deputado Clay Higgins. O pedido de vingança violenta contra "muçulmanos radicalizados" foi mantido pela equipe de moderação de conteúdo do Facebook. Mas a reportagem contrapunha esse exemplo a um outro, com desfecho diverso. Cerca de um mês antes, Didi Delgado, um poeta residente em Boston e ativista do movimento Black Lives Matter, postou o seguinte texto: "Todas as pessoas brancas são racistas. Comece desse ponto de referência, ou você já terá falhado". O post foi removido pelo Facebook – e sua conta suspensa por uma semana[184].

A diferença de desfecho não era acidental ou fruto de julgamentos idiossincráticos. Nos casos acima, o discurso do parlamentar americano era permitido porque tinha como alvo um subgrupo específico de muçulmanos – "aqueles radicalizados". A postagem da ativista Delgado, por outro lado, mirava todas as "pessoas brancas" em geral – e por isso poderia ser considerada como discurso de ódio, segundo as regras da plataforma. É nesse exato sentido a instrução interna repassada a moderadores da plataforma como parte de seus treinamentos, válida em 2016 e relevada pelo jornal The Guardian no ano seguinte[185]:



[184] "Facebook's secret censorpship rules protect white men from hate speech, but not black children", reportagem publicada pelo portal *ProPublica*, em 28/06/2017.

[185] "Revealed: Facebook´s internal rulebook on sex, terrorism and violence", reportagem publicada pelo jornal *The Guardian* em 21/5/2017.

(Figura 1. Tradução: *Quiz! Qual desses subconjuntos abaixo nós protegemos? 1-Mulheres motoristas; 2- Crianças negras; 3- Homens brancos. Gabarito: 3 – Homens brancos*)

Diante de críticas de que essa aplicação "neutra" de categorias diminuía exatamente a proteção destinada às pessoas mais vulneráveis, a empresa replicou que isso era uma decorrência da necessidade que possui de aplicar critérios consistentes em uma escala global: "as políticas não levam sempre a resultados perfeitos. Essa é a realidade de ter políticas que se aplicam a uma comunidade global na qual pessoas ao redor do planeta terão ideias muito diferentes sobre o que é OK para compartilhar", disse na ocasião Monica Bickert, "head of global policy" do Facebook[186]. Depois da repercussão da reportagem, o Facebook alterou suas políticas para tornar *"*idade*"* uma característica protegida[187]. Como esse episódio bem demonstra, a construção de regras de discurso do ódio a partir das categorias protegidas leva a demandas sociais e reivindicações para expansão ou alteração dessas categorias.

Além do uso de "categorias protegidas", essa política foi construída a partir de um pilar inicial pelo qual ataques a instituições (como países, religiões ou líderes) seriam permitidos, mas não ataque a grupos ou indivíduos (pessoas de uma certa religião, raça ou país). Nesse sentido, integrar um grupo intitulado "eu odeio cristãos" é proibido; já integrar um grupo chamado "eu odeio o cristianismo" é permitido pela plataforma. A mesma regra se aplica ao exemplo dos grupos "usando minha liberdade de expressão para informar que acho que homossexuais são nojentos" (proibido) e "grupo anti-homossexualidade" (permitido).

Os exemplos até aqui mencionados constam do material de treinamento interno voltado a moderadores do Facebook, válido a partir do ano de 2016 e relevado pelo jornal

[186] "Facebook's secret censorpship rules protect white men from hate speech, but not black children", reportagem publicada pelo portal *ProPublica*, em 28/06/2017.

[187] "(…) para criar regras que funcionem e sejam consistentes, plataformas devem criar distinções sobre as quais pessoas razoáveis podem muitas vezes discordarem. Às vezes, no entanto, as distinções são muito difíceis de justificar, e é apenas por meio de trabalhos investigativos de jornalistas que essas regras vêm à público de modo que possam levar a um debate proveitoso (...). Depois da indignação pública que se seguiu à publicação da reportagem de Angwin [repórter da ProPublica], o Facebook alterou suas políticas para tornar *idade* uma característica protegida" – Nicolas P. Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p. 19.

The Guardian no ano seguinte. A reprodução desse material pode ser verificada no Anexo 1 desta tese.

O Facebook divide conteúdos de discursos de ódio em *três níveis de severidade* – do "Tier 1" (mais grave) ao "Tier 3" (menos grave). Destaco a seguir alguns exemplos representativos desse conjunto, de acordo com a própria empresa:

> No grupo do "Tier 1": "*qualquer discurso violento ou apoio de modo escrito ou visual"; "discurso desumanizador tal como uma referência ou comparação a insetos, animais que são culturalmente percebidos como intelectual ou fisicamente inferiores, sujeira, bactéria, doenças, fezes, predadores sexuais"; dizer que "toda as pessoas [com uma determinada característica protegida] são criminosos", "ridicularizar o conceito, eventos ou vítimas de crimes de ódio mesmo que nenhuma seja de fato retratada na imagem*".
>
> No grupo do "Tier 2": "*afirmações de inferioridade ou uma imagem que implique que uma pessoa ou um grupo possui deficiência física, mental ou moral"; "expressões de desprezo, incluindo 'eu odeio' e 'eu não gosto de', ou 'X são os piores'"; "expressões de desgosto ou seus equivalentes visuais, como 'nojentos'"; "xingamentos a uma pessoa ou grupo que compartilhe das características protegidas*".
>
> Por fim, no grupo "Tier 3": "*ataques que são um chamado para excluir ou segregar uma pessoa ou grupo de pessoas baseado nas características acima citadas. Nós permitimos críticas a políticas de imigração e argumentos para restringir essas políticas*"[188].

Em debate realizado no final de 2018 na Universidade de Harvard, Monica Bickert, "head of global policy" do Facebook, explicou que os esforços de moderação proativa pela empresa "são largamente focados [em conteúdos encontrados] no Tier 1": "se nós vamos treinar uma tecnologia para achar isso, vamos fazê-lo para o material mais grave. Por outro lado, se vamos ser menos rigorosos, isso vai ocorrer no Tier 3". Por "esforços de moderação proativa", compreenda-se: uso de tecnologias automatizadas para filtragem de conteúdo, já abordadas no Capítulo 2.

[188] Novamente, alguns exemplos de aplicações práticas dessas regras podem ser conferidos no Anexo 1 desta tese, que reproduz material de treinamento interno do Facebook voltado a seus moderadores e datado de 2016.

Essas exceções previstas às regras do Tier 3 – voltado a "pedidos de segregação ou de exclusão de grupos de pessoas" – são um bom exemplo para demonstrar como as nuances de discursos e de posições políticas revelam as dificuldades inerentes de tentativas de se traçar linhas distintivas entre discursos de ódio e discordâncias democráticas legítimas.

"Imigrantes" são tratados pelo Facebook como uma "categoria quase protegida" – ou seja, que não possuem a proteção integral destinada às categorias protegidas, em especial no contexto dos pedidos de "segregação ou exclusão" de grupo de pessoas. A esse respeito, novamente Bickert comenta:

> "*queremos que as pessoas possam ter uma conversa política sobre imigração, então tendemos a deixar que uma pessoa diga 'não quero essas pessoas em meu país', mas se uma pessoa diz 'não quero essas pessoas no meu país, porque elas são nojentas' [seria um caso diferente] (...), assim como 'não quero pessoas dessa religião na minha escola' também seria removido". Dizer 'queimem os imigrantes', 'imigrantes são horríveis' ('awful') ou 'imigrantes não merecem um lugar na Terra', isso é discurso de ódio. Se você disser 'não quero mais imigrantes neste país', nós vamos permitir"*[189].

O material de treinamento interno revela outra distinção relevante no trato destinado a imigrantes ou refugiados: as pessoas podem ser chamadas de "sujas" ("filthy"), mas não de "sujeira" ou "lixo" ("filth"). A lógica para essa regra é: embora uma adjetivação seja permitida, o tratamento de forma substantiva implicaria um grau de desumanização a ser vedado[190].

Na ocasião do debate em Harvard, Jonathan Zittrain indagou: "E se você disser 'eu quero apenas pessoas que sejam descendentes do Mayflower em minha escola?"[191]. A resposta de Bickert expôs com aparente sinceridade algumas limitações com que as redes sociais gigantes lidam pela escala e amplitude dos conteúdos com que lidam e dificuldades operacionais para aplicar definições em torno do discurso de ódio:

---

[189] Evento "The State of Online Speech and Governance", debate promovido pelo *Berkman Klein Center for Internet and Society* (*Harvard University*), entre Jonathan Zittrain e Monika Bickert, em 03/12/2018 – vídeo disponível em https://www.youtube.com/watch?v=lWkhFB0f2tw

[190] "Facebook's secret censorpship rules protect white men from hate speech, but not black children", reportagem publicada pelo portal *ProPublica*, em 28/06/2017. Ver, ainda, material no Anexo 1 desta tese.

[191] Navio britânico que transportou os primeiros peregrinos ingleses ao novo continente, em 1620; nesse contexto, uma referência à ascendência inglesa ou branca.

> "*acho que você poderia dizê-lo. A questão é como contextualizar isso para quinze mil moderadores – fazer com que todos entendam quais são as implicações de se dizer que alguém veio a bordo do Mayflower, isso seria algo muito difícil de se fazer. Tem de ser algo muito explícito [para ensejar a proibição]. Sou a primeira a dizer que isso [conjunto de regras] não é perfeito. Não é fácil para nós, quando pensamos na elaboração dessa política, não é fácil operacionaliza-la (...). Um exemplo: imagem de ódio ('hateful image'). Se alguém compartilha uma imagem de um campo de concentração e diz 'imigrantes deveriam ir para este local', significa que nós temos que providenciar o contexto* [de que é uma imagem de um campo de concentração] *para nossos moderadores. Isso pode ser mais fácil se for um conteúdo que está viralizando – e então nós ficamos sabendo desse conteúdo. (...) Creio que há restrições operacionais. Em um sistema que funciona nesse tamanho, não há como treinar moderadores para que eles façam 'saltos de compreensão' para entender algo diretamente. Mas se alguns conteúdos específicos, que viralizam, chegam à nossa atenção, então podemos apontar para aquele post específico – 'esta é uma imagem de Auschwitz', então é isso o que a pessoa quer dizer. Nós podemos até usar a tecnologia para evitar que isso seja postado novamente, ou para que fique em uma fila de espera antes de ser postado definitivamente, aguardando a revisão de um moderador de 'hate speech'*"[192].

Como mencionado no Capítulo 2-B, ao longo de 2018, ao menos metade do conteúdo removido por vedações a 'discursos de ódio' foi feita de modo proativo pelo Facebook com uso de tecnologia – o que inclui o monitoramento de imagens e de textos. Ainda que essa seja uma tendência crescente, problemas essencialmente contextuais das nuances de linguagem continuam sendo um desafio para uma avaliação acurada de conteúdos, uma questão que só se faz ainda mais presente sob a rubrica dos "discursos de ódio".

As considerações feitas até aqui podem levar à impressão de que essas restrições a discursos de ódio decorrem de uma motivação restrita à manutenção de um ambiente

[192] Ainda sobre a menção a campos de concentração, o Facebook estabeleceu uma exceção à sua regra de que é proibido defender o envio de pessoas a campos de concentração: é permitido dizer que "nazistas devem ser enviados a um campo de concentração", porque ele próprios são um grupo de ódio que praticavam essa conduta – Facebook's secret censorpship rules protect white men from hate speech, but not black children", reportagem publicada pelo portal *ProPublica*, em 28/06/2017.

discursivo "saudável" ou protetivo à imagem e dignidade de grupos vulneráveis alvos de práticas discriminatórias. Mas essa seria uma impressão equivocada: por vezes, essa moderação pode ocorrer em um cenário real de perseguição ou ataque físico a minorias.

Se o episódio do genocídio de tutsis perpetrado em Ruanda em 1994 ficou atrelado às transmissões de rádio que radicalizaram, em linha crescente, mensagens de ódio contra aquela parcela da população – repetidamente chamando-a de baratas, a serem esmagadas –, o Facebook viu-se diretamente atrelado a episódios de perseguição e ataques contra a minoria muçulmana rohingya em Myanmar, entre os anos de 2017 e 2018[193]. Em ambos os casos, por óbvio, as tensões étnicas estavam presentes desde há muito tempo nas estruturas históricas daquelas sociedades – contudo, também nos dois casos, essas plataformas de comunicação foram usadas de modo deliberado para instigar uma campanha agressiva e apta a criar um ambiente propício a ataques reais.

Por questões próprias àquele país – por décadas, uma ditadura extremamente fechada ao mundo e que, em anos recentes, popularizou o uso de internet para a população via telefonia celular – *o Facebook virou sinônimo da própria internet para boa parte da população*. Uma população que, em especial, não tinha histórico de acesso a fontes abertas de informações e opiniões, o que alguns apontaram como um fator adicional de vulnerabilidade a campanhas de ódio e desinformação. Com frequência, pelo menos desde 2013, diversos ativistas e pesquisadores tentaram alertar o Facebook sobre a escalada das postagens que incitavam violência. Uma das páginas de grupo existentes, por exemplo, chamava-se "Gangue de decapitação de Kalars" ("Kalar Beheading Gang*"*). As páginas chegariam a ter, somadas, milhões de seguidores – e algumas delas eram ligadas a integrantes do governo. O tema chegou a ser abordado em uma reunião com ativistas de direitos humanos na sede da empresa, na California[194]. Mas durante o período

[193] Estima-se que ondas de ataques contra aquela minoria muçulmana, por parte de militares e de grupos paramilitares de extremistas budistas, tenha levado a incontáveis mortes e cerca de 700 mil refugiados – de uma população original de cerca de 1,2 milhões de pessoas do estado de Rakhine. "Receio que o Facebook tenha se tornado uma besta, que não queria originalmente ser", avaliou a relatora especial da ONU para os direitos humanos no país. Ela se referia ao papel central que o Facebook teve ao se tornar uma plataforma para a massiva campanha de ódio contra os rohingya – "The country where Facebook posts whipped up hate", reportagem publicada pela *BBC*, em 12/09/2018; "UN human rights chief points to ´textbook example of ethnic cleansing´ in Myanmar, reportagem da *United Nations News*, publicada em 11/09/2017. Ver também: David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 27-30

[194] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 27-30; "The country where Facebook posts whipped up hate", reportagem publicada pela *BBC*, em 12/09/2018; "Hatebook: why is Facebook losing the war on hate speech in Myanmar", reportagem publicada pela *Reuters*, em 15/08/2018.

compreendido entre os anos de 2013 e 2018 a empresa demonstrou não ter capacidade de agir à altura da gravidade do problema. Primeiro, porque não possuía moderadores suficientes que fossem fluentes em birmanês. Consta que, até 2015, eles eram apenas duas pessoas. O resto da moderação era feita por pessoas fluentes em inglês, que além de não conhecer bem o contexto político local, dependiam de um serviço de tradução que frequentemente era falho. O termo altamente perjorativo "kalar" (um xingamento direcionado à etnia rohingya) também pode significar, literalmente, "grão de bico"[195].

Em agosto de 2018, a empresa publicou um "mea culpa*"* – enfatizando "apostar pesadamente em inteligência artificial que possa identificar de modo proativo quais postagens ferem nossas regras". O Facebook também prometeu melhorias no sistema de "flagging", além de incentivos para que a plataforma fosse usada no padrão de fontes mundial. Disse também que iria intensificar os esforços já em curso de ter mais moderadores fluentes em birmanês – esperava chegar a cerca de 100 pessoas, a partir dos 60 já em atividade. Também foram banidas diversas páginas de grupos e de indivíduos que foram designadas como sendo figuras de promoção ao ódio[196].

Se a situação era especialmente dramática em Myanmar, vale perceber também como essas medidas não destoavam da política ordinária e mundial da empresa para combate a discursos de ódio em demais países: dependência de revisores humanos, tentativa de maior automatização da revisão de conteúdo com uso de inteligência artificial, diálogo com organizações da sociedade civil e banimento de páginas consideradas efetivamente danosas. Myanmar, marcou, contudo, como a rede social poderia se transformar de modo efetivo em uma plataforma de campanhas de ódio em um país no qual tensões étnicas irrompiam em ações genocidas.

A experiência libertária fracassada do Twitter já deixava evidente que uma postura de não intervenção de discursos nas grandes redes sociais era inviável.

---

[195] Além disso, a maior parte dos telefones celulares em Myanmar exibia o alfabeto birmanês utilizando um padrão técnico de fontes chamado Zagwi, que funciona apenas na língua local e costuma ser incompatível com o padrão Unicode, utilizado por quase todo o mundo. Isso significava, primeiro, que embora o Facebook aceitasse postagens em birmanes (que o padrão Unicode também suporta), toda sua interface para auxílio e denúncias ('flag') não podia ser visualizado pela maior parte dos aparelhos de Myanmar. As pessoas podiam postar em birmanês, mas não tinham acesso às orientações e canais de ajuda ou de denúncia dispostas na parte fixa da plataforma. Essa incompatibilidade entre padrões técnicos também dificultava o monitoramento do conteúdo postado. Em birmanês, um post destacado pela Reuters dizia: "Mate todos os kalars que você achar em Myanmar, nenhum deles deve ficar vivo"; a tradução em inglês pelo Facebook era "Eu não deveria ter um arco-íris em Myanmar".

[196] "Update on Myanmar", nota publicada pelo *Facebook Newsroom*, em 15/08/2018.

Imperativos derivados da *economia*, das *normas sociais* e do próprio *direito* (especialmente de estados europeus, no caso) geraram uma pressão crescente para que as grandes redes sociais lidassem melhor com os problemas de discursos de ódio em suas plataformas. Com isso, o Facebook ampliou suas restrições e banimentos entre os anos de 2018 e 2019. Nos Estados Unidos, personalidades extremistas foram alvos das restrições: entre elas Alex Jones, o mais famoso difusor de teorias conspiratórias do país, e seu portal Infowars[197].

E isso faz sentido, já que esses ambientes (ou *contextos*) específicos criaram novas formas de divulgação e de massificação de discursos que amplificaram o potencial tóxico de mensagens de ódio, possibilitando que manifestações extremistas – antes filtradas pelos intermediários tradicionais e contrabalanceadas pela "desaprovação social aberta" – ganhassem maior repercussão social. De um modo geral, isso se traduz na *possibilidade de viralização de conteúdos* (uma nova dinâmica geral que incide no debate público); de modo mais específico, na *produção de discursos baratos como armas* ("weaponized speech"*)*[198], muitas vezes como táticas de campanhas que buscam silenciar grupos (minorias) ou mesmo incitar ataques reais a elas. *Esse contexto específico gera riscos e danos que devem ser gerenciados e enfrentados pelas empresas*[199].

As noções de risco ou de dano no ambiente das redes sociais não deve corresponder àquelas que um estado deve levar em conta quando regula ou potencialmente proíbe discursos. *Redes sociais não criminalizam condutas ou discursos – tampouco se confundem com a internet em si.* É natural que, em comparação com

---

197 Além de Alex Jones, o Facebook também baniu as seguintes personalidades: o polemista Milo Yiannopoulos, que fez fama à frente do site de extrema-direita Breibart News; Laura Loomer, famosa por sua retórica contra muçulmanos, bem como Louis Farrakhan, líder muçulmano conhecido por sua retórica antissemita. Parte deles já tinham sido alvo de derrubadas ou suspensões anteriores pelo Facebook, mas a empresa deixou claro em meados de 2019 sua decisão de aplicar as regras de sua política de modo mais consistente e abrangente, incluindo também sua plataforma Instagram, bem como a republicação de materiais por terceiros – "Instagram and Facebook ban far-right extremists", reportagem publicada por *The Atlantic*, em 02/05/2019.

198 Capítulo 2.F.

199 David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 13. Esse fator também causou a elevação da pressão de governos sobre as redes sociais ao longo dos últimos anos: "em uma conferência da ONU em setembro de 2015, [Angela] Merkel foi gravada contestando Mark Zuckerberg sobre isso abertamente. Skinheads e outros cheios de ódio racista agiam na Alemanha há tempos, mas suas mensagens podiam ser isoladas e suas táticas abordadas com ferramentas tradicionais de aplicação da lei. Mas havia algo diferente ali: o ódio se espalhava de modo viral, sem os fatores de moderação da mídia tradicional e sem a luz do sol da desaprovação social aberta" (p. 66). Ao final, a Alemanha promulgou em 2017 sua legislação NetzDG – largamente motivada pelo combate a discursos de ódio – e que será analisada no Capítulo 4-E.

estados, tenham uma maior margem para a restrição de discursos que causem esses riscos ou danos, assim considerados dentro de seus *contextos institucionais específicos*[200].

Por isso, essa seara convida a uma disputa permanente – conceitual e também a partir de casos concretos – sobre onde traçar as linhas *do que se pode ou não dizer*. Por sua natureza "limítrofe", o tema dos discursos de ódio demonstra como as redes sociais sofrem ao mesmo tempo pressões para serem permissivas e restritivas com relação a conteúdo sensíveis – independentemente de quais são as decisões tomadas, é certo que haverá descontentamentos e críticas, ora por partidários de restrições a conteúdos considerados danosos, ora de críticos que defendam restrições mínimas à liberdade de expressão dos usuários. Pessoas razoáveis podem discordar de qualquer critério adotado. *Daí a importância de que sejam públicos e submetidos a escrutínio, o que tem ocorrido, no caso do Facebook, apenas desde 2018.*

No mais, esse aspecto evidencia como *as grandes redes sociais devem tomar decisões de cunho editorial, articuladas em torno dos valores e objetivos que priorizam em seus ambientes*. Essas decisões estão entrelaçadas nas regras de discursos daquilo que é permitido ou proibido dizer. Se redes sociais nunca são neutras (Capítulo 2), suas regras substantivas, publicizadas ou não, contêm as escolhas valorativas que fazem. Como o exemplo da regra de "proteção parcial" a imigrantes deixa claro, a postura regulatória do Facebook de proibição a "discursos de ódio" implica encarar problemas espinhosos na seara dos discursos políticos – e que demandam decisões editoriais e valorativas. No próximo tópico, irei abordar como essa conexão delicada entre discursos de ódio e temas legítimos de debate público iria, no caso do Facebook, abrir caminho para que a empresa *assumisse de vez seu papel editorial*, reconhecendo uma maior liberdade discursiva a "figuras políticas" ou a postagens de "interesse noticioso".

3.C – Facebook e a proteção ao debate público: da regra da "figura pública" à regra do "interesse noticioso" ("newsworthy")

Outro aspecto central para compreender a atividade de moderação de conteúdos pelo Facebook é a consolidação de *conceitos que justificam a manutenção no ar de postagens que, a princípio, seriam retiradas por violação a seus termos de uso*, mas cujas publicações são ao fim justificadas pelo resguardo e fomento ao debate público. A esse

[200] Esse ponto será retomado no Capítulo 4-D.

respeito, este tópico vai elaborar como o Facebook passou a se utilizar dos conceitos de "figuras públicas", inicialmente, e de "interesse noticioso" ("newsworthy"), posteriormente, na busca de uma melhor calibragem na implementação de suas regras, traçando linhas para evitar a retirada de postagens conectadas a temas de interesse público – uma postura que remete a funções tradicionalmente ligadas à liberdade de imprensa.

O início dessa construção remete a 2009, quando o tema do "cyberbullying" começou a ser pautado com vigor, especialmente pela imprensa[201]. O Facebook – como uma plataforma emergente e em franca expansão – viu-se no centro de cobranças para exercer maior controle sobre postagens abusivas contra jovens. Mas a empresa não vislumbrava, a princípio, como formalizar regras objetivas que identificassem essas situações a partir apenas do conteúdo das postagens, que pouco diziam sobre o contexto social das pessoas envolvidas[202].

Kate Klonick e Thomas Kadri relatam que, diante desse problema, a plataforma tinha dois caminhos principais a trilhar: "errar pelo lado de deixar no ar conteúdo potencialmente nocivo, ou errar pelo lado de remover todos os potenciais atos de *bullying*, ainda que parte deste conteúdo fosse benigno". O Facebook optou pela última opção, mas incorporando *uma importante exceção à regra geral de remoção*: ela seria aplicada somente a pessoas não-públicas. Figuras públicas, assim, não contariam com essa proteção contra mensagens agressivas e abusivas[203].

Para esses autores, *esse foi um ponto de inflexão na política de conteúdo do Facebook*, que teria incorporado uma racionalidade predominante na jurisprudência da

---

[201] Por exemplo: "Social media brings bullying to light", reportagem publicada pela *CNN*, em 10/12/2009; "Tenageer is first to be jailed for Facebook bullying", reportagem publicada pelo *The Telegraph*, em 21/08/2009; "Prosecuting Cyber Bullies", reportagem publicada pela *National Public Radio*, em 5/4/2009.

[202] "Como você escreve uma regra sobre bullying? O que é bullying? O que você quer dizer com isso? Não é algo que apenas faz te sentir mal ('upsetting'); isso é definido como um padrão de comportamento não desejado, de abuso ou assédio, por um período de tempo, que ocorra entre alguém com maior poder sobre alguém com menor poder. Mas essa não é uma resposta ao problema que reside no conteúdo – você não pode determinar quem é um poder menor olhando para o conteúdo. Você muitas vezes não pode determinar isso sequer olhando para seus perfis ('profiles')", Dave Willner, funcionário do Facebook à época – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 60.

[203] "Veja, se você diz para nós [Facebook] que isso é sobre você, e que você não gosta [desse conteúdo] – e se você é um indivíduo privado – ou seja, você não é uma figura pública – então nós derrubamos esse conteúdo (...). Porque não temos como saber se todos os outros elementos [da definição] do bullying estão presentes, a gente simplesmente teve que tomar uma decisão para criar uma regra padrão para a remoção de bullying.", Jud Hoffman, funcionário do Facebook à época – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 61.

Primeira Emenda a partir do caso New York Times Co v. Sullivan*:* "de modo a preservar um discurso público robusto na plataforma, o time de Hoffman tomou a decisão consciente de tratar certos alvos de discursos supostamente nocivos de modo diferenciado com base em seus *status* social e o interesse público sobre suas ações"[204]. A proteção a um debate público "livre, robusto e amplo" remete a célebre passagem de New York Times Co. v Sullivan[205]; a decisão – considerada um marco para as liberdades de expressão e de imprensa[206] – definiu regras constitucionais mais rigorosas em ações de difamação contra servidores públicos. Foi essa razão de fundo – a proteção ao debate público "livre, robusto e amplo" – que iria fincar raízes no desenvolvimento da jurisprudência americana da liberdade de expressão. Sob essa racionalidade, casos subsequentes da Suprema Corte fariam com que a categoria de "figuras públicas" fosse ocupada não apenas por servidores públicos, sendo ampliada também para demais pessoas cujo caso sob análise demonstrasse conexão com temas de interesse público[207].

A decisão do Facebook, contudo, tinha implicações nada simples. Afinal, como definir se alguém era uma "figura pública" para fins de aplicação da política de conteúdo? Essa mesma questão havia consumido diversas decisões da Suprema Corte ao longo de anos, abrindo espaço para complicadas categorias jurisprudenciais. No Facebook, a solução encontrada foi verificar, em cada caso de conteúdo reportado ("flagged") como

---

[204] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 61. Os autores registram que Jud Hoffmann, "Global Policy Manager" na época, nega que a regra tenha sido expressamente retirada da doutrina da Primeira Emenda, postura justificada por ela pela missão da empresa de tornar o mundo "mais conectado" através de cada vez mais "compartilhamento de conteúdos", mas defendem que as razões invocadas remontam diretamente à linhagem jurisprudencial consolidada em *New York Times v. Sullivan*.

[205] "Avaliamos esse caso sob o pano de fundo do profundo compromisso nacional de que o debate sobre questões públicas deve ser livre, robusto e amplo, e de que ele pode incluir veementes, cáusticos e por vezes desagradáveis e enfáticos ataques contra o governo ou servidores públicos (...). Afirmações erradas são inevitáveis em um debate livre, e isso deve ser protegido para que a liberdade de expressão tenha o 'ar' ('breathing space') de que necessita.... para sobreviver" – New York Times Co. v. Sullivan (1964), Suprema Corte dos Estados Unidos.

[206] Embora muitas vezes associada diretamente à liberdade de imprensa, a decisão na verdade dá início à tradição da Suprema Corte americana de não fazer qualquer distinção entre a liberdade de expressão de cidadãos comuns e jornalistas. No caso, inclusive, o texto publicado que deu origem ao litígio não era uma reportagem – e sim um anúncio pago por ativistas do movimento antirracista em favor dos direitos civis, um aspecto muitas vezes negligenciado nos relatos produzidos a respeito. Para uma análise mais detida sobre o caso em geral e esse aspecto em especial: Rodrigo Vidal Nitrini, *Liberdade de informação e proteção ao sigilo de fonte: desafios constitucionais na era da informação digital*. Hucitec Editora, 2016, pp. 60-64; Anthony Lewis, *Make no law: the Sullivan Case and the First Amendment*. Vintage Books, 1992; Timothy Cook, "Introductory Essay", *in*: Timothy Cook (org.), *Freeing the presses: the First Amendment in action*, Louisiana State University Press, 2006.

[207] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 42 e seguintes.

bullying, se a suposta vítima aparecia no "Google News" – um site do Google que agrega inúmeras páginas de notícias. Se o nome dessa pessoa dava resultado positivo no Google News, o conteúdo não era derrubado, pois se presumia que se tratava de uma figura pública. Se não aparecia, era tratada como uma pessoa comum, resultando na forte presunção de regras a favor da imagem pessoal[208].

Klonick e Kadri ponderam como o uso do agregador do "Google News" era útil para o Facebook, que precisava de uma solução ágil e fácil de ser aplicada por seus moderadores – em um sistema que lidava com um problema intrínseco de escala. Mas essa solução também gerava sua própria sorte de problemas: uma delas é a questão da "figura pública involuntária" – que se mais rara no passado, aparecia com mais frequência em um mundo no qual a viralização de conteúdos na internet coloca pessoas inadvertidamente sob os holofotes[209]. Nesse novo mundo, viralizar no Facebook poderia ser uma maneira de entrar nos registros do Google News e, assim, perder a proteção que usuários em geral gozam no próprio Facebook contra mensagens e comentários agressivos ou abusivos.

Mas a questão principal decorrente da escolha da regra de exceção para "figuras públicas" era outra. Como já havia ocorrido anos antes no âmbito da Suprema Corte em julgamentos de ações de indenização por invasão de privacidade, o estabelecimento dos contornos das "figuras públicas" – já então definido pelo Facebook com o uso do Google News – em pouco tempo se transformou em um debate sobre se o conteúdo fazia referência a um *material de interesse noticios* ("newsworthy"). Como ressaltam Klonick e Kadri:

> *"Tudo o que uma busca no Google News poderia demonstrar a moderadores era que o nome de alguém havia aparecido em uma fonte de notícias – isso não poderia relevar de maneira consistente o verdadeiro status social de uma pessoa, a reputação da fonte de informação, a veracidade da história ou a*

[208] A designação de alguém como figura pública era um processo completamente dissociado da "verificação pública" da conta – que gera um "checkmark"/símbolo azul, demonstrando que a identidade da conta havia sido de algum modo verificado pela plataforma. Aquele sinal, visível a demais usuários, é feita por uma outra equipe da empresa – e não era utilizada pela equipe de moderação com a finalidade da aplicação da regra aqui descrita – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 62.

[209] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), pp. 80-88. Para uma história icônica de uma "pessoa pública involuntária" na era de viralizações em redes sociais, ver: "How one stupid Tweet blew up Justine Sacco's life", reportagem publicada pela *The New York Times Magazine*, em 12/02/2015.

> *gênese da controvérsia. Esse quadro significava que o Facebook encontraria muitas das questões que já ocupavam as cortes ao tentar definir os limites de conteúdos noticiosos ('newsworthiness')"*[210].

Como se não bastassem as dificuldades tradicionais de se definir o que é uma notícia[211], os profissionais da equipe responsável pela moderação de conteúdo já percebiam o Facebook como integrante de um ecossistema digital em que a facilidade de publicação proporcionada pela internet havia borrado as linhas entre o jornalismo profissional tradicional e novos veículos – estes pessoais, engajados, partidários ou comparativamente mais falhos. Por essas razões, a equipe de moderação era contrária ao estabelecimento de uma regra mais ampla que implicasse a manutenção de todo e qualquer conteúdo que fosse "newsworthy". Por alguns anos, eles mantiveram a decisão de manter essa exceção restrita aos casos de "bullying", conceituando como figuras públicas as pessoas cujas buscas retornavam resultados positivos pelo Google News. Uma decisão deliberada, que iria perdurar de modo consistente até 2013, ano do violento atentado a bomba cometido na maratona de Boston.

Uma das fotos que circulou logo na esteira da tragédia de Boston retratava uma das vítimas – um homem em uma cadeira de rodas – com a perna destroçada abaixo do joelho e parte de seus ossos expostos. Havia três versões daquela foto: uma delas editada, que ocultava a parte explícita de sua perna arrancada; uma segunda, tirada de um ângulo mais amplo, no qual esse ferimento era um detalhe menos perceptível; e uma terceira, mais explícita, que mostrava claramente uma forte imagem violenta. Todas elas tinham sido publicadas por veículos de imprensa durante a cobertura do atentado. A princípio, o Facebook baniu a terceira versão, determinando a remoção das imagens em todas as suas postagens, em virtude da clara exibição explícita de um corpo humano destroçado. A

---

[210] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 63. Os autores avaliam o Facebook se deparou com o mesmo dilema que cortes judiciais tinham enfrentado no julgamento de casos em décadas anteriores, qual seja, uma escolha fundamental sobre duas abordagens possíveis para o critério do interesse noticioso: um normativo, que se propõe a determinar o que constitui uma notícia; e outro descritivo, que toma a publicação de uma notícia como evidência de que se trata de um assunto merecedor de uma notícia, tendo por base uma alta deferência aos julgamentos editoriais de veículos jornalísticos, ainda que com eventuais reservas.

[211] Quem define o que é notícia? Notícia para quem? Para uma análise sobre essa questão, a partir das abordagens *institucional* e f*uncional* da liberdade de informação: Rodrigo Vidal Nitrini, *Liberdade de informação e proteção ao sigilo de fonte: desafios constitucionais na era da informação digital*, Hucitec Editora, 2016, pp. 79-124.

política então vigente tinha sido aplicada de modo consistente – até que, pouco tempo depois, altos executivos da empresa determinaram a restituição da foto e de todas as suas postagens, sob o argumento de que ela era dotada de claro interesse noticioso.

Pela primeira vez, *a exceção da regra de interesse noticioso prevaleceu sobre outro tipo de proibição* – no caso, a vedação de imagens de violência explícita. Como resultado, parte da equipe de moderação de conteúdo discordou da imposição da exceção "ad hoc*"*, que colocava pressão no que viam como uma política que vinha sendo construída sobre procedimentos e critérios consistentes[212]. A regra de manutenção de conteúdo "newsworthy" foi aplicada algumas outras poucas vezes nos anos seguintes, de modo claudicante, em discussões que não ultrapassavam o âmbito interno do Facebook. Mas isso mudaria a partir de uma polêmica que ultrapassaria os muros da empresa e que ganharia manchetes por todo o mundo, nascida na Noruega.

Em meados de 2016, o Facebook viu-se em meio a uma polêmica global, que envolveu órgãos de imprensa e especialistas em mídia, por conta da seguinte decisão: a plataforma havia suprimido a publicação de uma postagem do escritor norueguês Tom Egeland porque continha a imagem da icônica fotografia "The terror of war", de Nick Ult[213]. Popularmente conhecida como "Napalm Girl photo", a imagem ganhadora do prêmio Pulitzer em 1973 registra crianças vietnamitas correndo em fuga de um ataque com bombas napalm a seu vilarejo durante a Guerra do Vietnã. Entre elas, uma garota nua, que se tornou sua figura símbolo. A princípio, a postagem foi suprimida em razão da exibição de nudez infantil. O conteúdo original publicado por Egeland pautava "sete fotografias que mudaram a história das guerras".

O caso acendeu de vez a fagulha no debate público a respeito da moderação de conteúdo da rede social: o maior jornal norueguês, Afterposten, publicou em sua primeira página uma reprodução da imagem, ao lado de uma carta aberta a Mark Zuckerberg, acusando-o de "abuso de poder". O editor-chefe do jornal registrou que "estava chateado, desapontado – e, de fato, até com medo" sobre o que Zuckerberg poderia fazer pela

---

[212] "Filosoficamente, se nós fossemos assumir a posição de que [partes internas para fora do corpo] era nossa própria definição de sanguinolência explícita ('gore') e de que nós não permitíamos isso, então apenas porque isso aconteceu em Boston não alterava esse fato", funcionário anônimo do Facebook à época – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 65.

[213] "Fury over Facebook 'Napalm Girl' censorship", reportagem publicada pelo portal *BBC News,* em 09/09/2016.

manutenção de uma sociedade democrática. A então primeira-ministra norueguesa, Erna Solberg, também realizou uma postagem no Facebook exibindo a imagem e pedindo para que a empresa revisasse seus critérios com relação ao episódio. A postagem de Solberg igualmente foi apagada; quando ela reclamou publicamente a respeito dessa decisão, sua conta foi suspensa. Metade de seu gabinete de governo também publicou a foto, sob o argumento de defesa da liberdade de expressão e de preservação da memória histórica. A querela ganhou repercussão em órgãos de imprensa de diversos países, que ilustravam suas reportagens, publicadas no próprio Facebook, com a foto "proibida"; outros usuários da plataforma em todo o mundo demonstravam seu descontentamento igualmente publicando-a em suas páginas[214].

No início da polêmica, o Facebook defendeu publicamente sua posição de supressão da fotografia; no entanto, após um dano considerável de imagem pública, a empresa voltou atrás:

> "*Depois de ouvir nossa comunidade, nós revisamos como nossos Padrões de Comunidade foram aplicados nesse caso. Uma imagem de uma criança nua normalmente geraria a presunção de violação de nossos Padrões de Comunidade e em alguns países isso poderia inclusive caracterizar pornografia infantil. Neste caso, nós reconhecemos que a importância global e histórica dessa imagem ao documentar um momento particular*"[215].

Era a crise de imagem que faltava para justificar a adoção plena do critério de interesse noticioso. Cerca de seis semanas depois, um novo comunicado da empresa o formalizou:

> *"nós vamos começar a permitir mais itens que as pessoas achem que possui interesse noticioso ('newsworthy'), significativo ou importante para o interesse público – ainda que eles possam por outro lado violar nossos padrões [termos de uso]. Nós vamos trabalhar com nossa comunidade e parceiros para explorar exatamente como fazer isso, tanto por meio de novas*

[214] "Facebook deletes Norwegian PM´s post as 'napalm girl' row *escalates*", reportagem publicada pelo jornal *The Guardian* em 09/09/2016; "Facebook backs down from 'napalm girl' censorship and reinstates photo", reportagem publicada pelo jornal *The Guardian* em 09/09/2016.

[215] A manifestação finalizava concluindo que "em razão de seu status como uma imagem de importância histórica, o valor de se permitir se permitir seu compartilhamento supera o valor de proteger nossa comunidade por sua remoção, então decidimos reinstalar a imagem no Facebook" – "Facebook backs down from 'napalm girl' censorship and reinstates photo*"*, reportagem publicada pelo jornal *The Guardian* em 09/09/2016.

> *ferramentas quanto por abordagens na aplicação de nossas regras ('enforcement'). Nosso objetivo é permitir mais imagens e histórias sem que isso traga riscos de segurança ou mostre imagens fortes a menores ou outros que não queriam vê-las"*[216].

A fissura inicial provocada pela regra da "figura pública" para casos de "bullying*"* ou assédio havia assim *progredido para uma regra geral de manutenção de conteúdo de interesse noticioso*.

Desde então, como ocorre com toda a política de moderação, as regras têm sido constantemente reformuladas em busca de aprimoramentos[217]. Figuras públicas, por exemplo, passaram a gozar de um maior grau de proteção contra mensagens agressivas[218]. Um documento interno que continha diretrizes a moderadores, publicado pelo jornal The Guardian em 2017, por exemplo, explicava: "Rihana é famosa, por seus próprios esforços, por ser uma cantora. Ela também é uma vítima de violência doméstica. Você pode ridicularizá-la em função de seu canto, mas não por ser uma vítima de violência doméstica"[219].

Além disso, o processo para determinar se alguém é uma figura pública tornou-se mais complexo. Menções recorrentes no Google News durante um determinado período de tempo continuam a ser um critério importante, mas também são categorizadas

---

216 "Input from Community and Partners on Our Community Standards", comunicado publicado pelo *Facebook Newsroom,* em 21/10/2016.

217 Para um relato sobre a reunião quinzenal que o Facebook promove para a revisão de suas regras de moderação e refinamento de critérios, ver: David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 53 e seguintes. Para um estudo sistematizado, porém ainda inicial, sobre esse processo de revisão e produção de normas pela empresa, ver: Mathias C. Kettemann e Wolfgang Schulz, "Setting rules for 2.7 billion: a (first) look into Facebook's norm-making system – results of a pilot study", *Working Papers of the Hans-Bredow-Institut/ Leibniz Institute for Media Research (Hamburg) – Works in Progress #1* (2020).

218 "'Nossa nova política não permite alguns tipos de ataques de alta intensidade, como ameaças de morte, direcionadas a uma determinada figura pública', disseram membros da equipe de política de conteúdo do Facebook em ligação recente. No passado, eles explicaram, uma frase como 'Kim Kardashian' é uma vagabunda" nunca seria removida por *bullying* ou assédio (enquanto chamar uma pessoa privada disso seria). Mas agora, o Facebook permite que alguns discursos voltados a figuras públicas, quando postados em suas próprias páginas ou contas, sejam removidos dependendo da severidade da linguagem" – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 67.

219 "Contudo, o documento diz a moderadores que eles não devem derrubar automaticamente um post, se reportado ('flagged'), que diz: "Rihana, por que você continua trabalhando com Chris Brown? Essa pergunta me dá um tapa ('Beats me')". A ironia nesse caso, de acordo com a instrução, deve ser mantida como expressão legítima – "How Facebook allow users to post footage of children being bullied", reportagem publicada pelo *The Guardian*, em 22/05/2017.

assim "pessoas eleitas ou designadas por meio de um processo político a uma posição governamental; pessoas com centenas de milhares de fãs ou seguidores em alguma conta de rede social e pessoas que trabalham para empresas de mídia ou que costumam falar publicamente"[220].

A regra do interesse noticioso, no geral, é aplicada de modo menos mecânico – e mais contextual. Revisores da empresa avaliam a incidência dessa regra caso a caso:

> *"Ao decidirem isso, os empregados do Facebook enfatizam que eles ponderam ('weight the value') a 'voz' contra um risco de dano. As avaliações de dano são feitas com base na natureza, bem como na substância do conteúdo sob revisão. O discurso de ódio por si só, por exemplo, pode ser visto como menos danoso do que um apelo direto à violência. Mas os profissionais mantêm que, na maior parte dos casos, essas decisões sobre o caráter noticioso se relacionam com [casos de] nudez. Decisões difíceis incluem o que fazer no caso de nudez em protestos"*[221].

Ao trilhar esse caminho, o Facebook incorporou assumidamente um papel tradicionalmente atribuído a veículos de imprensa: definir para seu público "o que é notícia" ou "de interesse público" – *função que, de um ponto de vista jurídico, parece intimamente conectada a uma ideia de "liberdade editorial", aspecto que será desenvolvido no próximo tópico deste capítulo*.

Como visto, a utilização a título "normativo" pelo Facebook de resultados do Google News demonstra, em importante medida, a deferência que a plataforma concede a decisões editoriais e de publicações de outras páginas e veículos (o que implica sua própria sorte de problemas). Mas é importante destacar que essa decisão não é tomada apenas com base no "ecossistema externo" de notícias por outras mídias. *O Facebook também se investe nesse papel tradicionalmente desempenhado por editores ao definir quais postagens em sua própria plataforma merecem especial proteção pelo caráter*

---

[220] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), pp. 67-68.

[221] Sobre nudez em protestos, um dos responsáveis explicou que "há apenas alguns anos nós derrubávamos conteúdos assim. Mas é importante que deixemos esse conteúdo no ar, algo consistente com nossos princípios de dar voz. Isso levou a uma mudança da política, que agora opera em escala para toda a plataforma" – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), pp. 68-69.

noticioso. E, inexoravelmente, essa avaliação editorial é sempre contextual, além de fruto de uma atividade valorativa. Dois casos ilustram bem esse ponto.

Em 2017, um morador de Cleveland matou a tiros um idoso negro que andava perto de sua casa, na calçada. Ele disse que decidiu matar alguém porque estava bravo com sua ex-namorada – e o vídeo do crime, postado por ele, ficou no ar por duas horas antes de ser removido pelo Facebook. Compare essa primeira situação com a seguinte: no ano anterior, Philando Castile, um cidadão também negro, foi baleado enquanto era abordado por policiais dentro de seu carro. Sua namorada transmitiu ao vivo um vídeo no qual ele agonizava, até morrer. O vídeo chegou a ser retirado da plataforma, mas foi restabelecido posteriormente sob o aviso prévio de que se tratava de material com cenas fortes. A transmissão ao vivo, sem edições, deu repercussão nacional ao episódio sob o tema da violência policial direcionada a pessoas negras, na esteira do movimento Black Lives Matter:

> "*Para moderadores assistindo a ambos, os vídeos podem parecer similares – um horrível vídeo de uma pessoa negra sendo baleada nos Estados Unidos – mas a empresa eventualmente determinou que as intenções por detrás de cada um lhes davam significados distintos: deixar o segundo no ar chamava atenção ao racismo institucional do sistema de justiça criminal, enquanto derrubar o primeiro silenciava uma homenagem insana de um assassino a sua ex-namorada*"[222].

Além disso, o percurso que começou com uma preocupação pontual para evitar que as restrições ao "cyberbullying*"* interferissem de maneira desproporcional em debates públicos levou o Facebook à *posição especialmente difícil de árbitro na aplicação de suas próprias regras na seara de debates políticos, especialmente quando protagonizados por candidatos ou figuras partidárias.*

Durante a campanha presidencial americana de 2016, o Facebook já havia enfrentado uma discussão interna sobre se permitia ou proibia as postagens do então candidato Donald Trump que propunham o banimento à entrada de muçulmanos no País, sob argumentos de que "grande parte dos muçulmanos odeia os americanos". As postagens foram reportadas como inapropriadas por diversos usuários e parte

[222] Kate Klonick, "Inside the team at Facebook that dealt with the Christchurch shooting", *The New Yorker*, artigo publicado em 25/04/2019.

considerável da equipe de moderação de conteúdo entendia que seu teor violava as regras da empresa que vedam discursos de ódio – no caso, a proibição contra os pedidos de exclusão de um grupo religioso. A celeuma subiu ao alto escalão da empresa. Ao final, as postagens de Trump foram permitidas. Mas as razões para essa decisão nunca foram articuladas publicamente e, por isso, não restaram esclarecidas[223].

A falta de clareza ou de transparência quanto às razões da decisão da empresa, portanto, era um problema em si. Isso ficava cada vez mais evidente por conta de decisões tomadas em aparente sentido contrário em outros países: no final de 2018, por exemplo, a conta do filho do primeiro ministro de Israel foi temporariamente suspensa depois que o rapaz de 27 anos publicou que desejava que "todos os muçulmanos saíssem da terra de Israel". A remoção foi determinada com base nas regras de vedação ao discurso do ódio da plataforma[224].

O Brasil também já teve um caso semelhante. Também em 2018, durante período prévio à campanha presidencial, o então pré-candidato Geraldo Alckmin moveu uma ação na justiça estadual contra o Facebook pleiteando liminarmente a retirada do ar de um vídeo postado pelo perfil do vereador carioca Carlos Bolsonaro[225]. A mensagem da montagem editada era clara: identificar seu adversário como um apoiador do movimento LGBT, caricaturando-o por isso. A ação de Alckmin não encontrou guarida no judiciário:

---

[223] O caso veio à tona logo depois que a empresa publicou a posição pública de adotar a regra do interesse noticioso, na esteira da polêmica da "Napalm Girl photo". Prevaleceu a versão de que Mark Zuckerberg decidiu pessoalmente que as postagens deveriam ser permitidas, por se tratar de notícia de interesse público conectada ao debate político presidencial, já que era uma proposta proveniente de um dos candidatos. Por outro lado, também foi divulgado que Trump talvez tenha se beneficiado da exceção para a categoria de subgrupos: "um banimento a muçulmanos poderia ser interpretado como direcionado contra um subgrupo, imigrantes muçulmanos, e por isso não configuraria discurso de ódio contra uma determinada categoria – "Facebook employees wanted to block Donald Trump for hate speech, but Mark Zuckerberg said no", reportagem publicada pelo *Business Insider*, em 21/10/2016; "Facebook employees wanted Trump post removed", reportagem publicada por *The Hill*, em 21/10/2016; "Facebook's secret censorpship rules protect white men from hate speech, but not black children", reportagem publicada pelo portal *ProPublica*, em 28/06/2017; Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), pp. 89-91.

[224] "Facebook temporarily bans Israeli PM's son over post", reportagem publicada pela BBC News, em 17/12/2018; Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 92.

[225] O vídeo, postado originalmente em 2017, era uma montagem na qual imagens supostamente controversas – como representações transexuais de Jesus Cristo, cenas de sexo explícito entre homens em vias públicas, além de presença de crianças em eventos que pareciam ser mobilizações públicas de orgulho LGBT – foram apostas a um vídeo original de Alckmin, no qual ele se dirigia ao grupo "Diversidade Tucana", parabenizando-o e empenhando seu apoio pessoal pelo fato de o trabalho daquele coletivo ter alcançado o status de secretariado dentro do PSDB. A ação pedia também a identificação do responsável pela postagem, visando eventual futura de indenização por danos à honra e imagem de Alckmin.

a liminar pleiteada para retirada do vídeo foi negada em primeira instância. Em um trecho de sua decisão, a juíza responsável considerou que "o direito à liberdade de expressão e de manifestação, assim como o direito à honra, devem coexistir harmoniosamente", além de que "somente em situações excepcionais" poderia haver a retirada de conteúdo do ar, "sob pena de ofensa aos princípios democráticos já elencados e caracterização de censura". Ainda assim, com a divulgação na imprensa do ajuizamento da ação, o Facebook decidiu *por conta própria* derrubar a postagem de sua plataforma. A empresa argumentou que, "após ter recebido denúncias de usuários", verificou que o material "violava seus termos de uso". *Novamente, a menção genérica ocultava quais seriam especificamente as regras violadas e as razões da decisão*[226]. Alckmin teve seu pedido de derrubada negado pelo judiciário, mas acolhido pelo sistema normativo autônomo do Facebook[227].

É possível debater se esses três últimos casos são completamente análogos e se há razões consistentes para que o Facebook tenha chegado a decisões opostas em cada um deles. Mas a contraposição entre eles demonstra como *as decisões editoriais do Facebook passaram a ter importância crescente a respeito de quais conteúdos podem ou não circular no debate público – e no político e eleitoral, em especial.*

Em sua movimentação mais recente, as regras do Facebook foram alteradas para *garantir a figuras públicas um grau mais favorável de liberdade discursiva* – uma espécie de reviravolta, já que o sentido original das regras era prover a elas um menor grau de proteção contra os ataques de terceiros.

No final de 2019 – ano anterior à eleição presidencial americana – o Facebook anunciou que políticos, de um modo geral, teriam uma maior imunidade relativa com relação às regras de política de conteúdo, de duas maneiras distintas. Primeiro, suas postagens seriam isentas da política de "checagem de fatos" – incluindo aí suas

[226] É possível cogitar que o problema seria a estereotipação destinada a inferiorizar gays ou ridicularizar os crimes de ódio contra esse grupo. Também há a possibilidade de que tenha havido o juízo de que o vídeo alterado poderia induzir o público a erro, de modo a acreditar que era o vídeo original e oficial de Alckmin a abarcar aquelas imagens. O fato de o Facebook não ter explicado publicamente os motivos de sua decisão impede de saber quais foram; como ré, a empresa tampouco informou nos autos as razões pela derrubada autônoma do vídeo. Logo após o vídeo ter sido apagado, o autor desistiu da ação, que terminou arquivada – "Facebook tira do ar vídeo sobre Alckmin postado por filho de Bolsonaro", reportagem publicada pelo portal *Universo Online* em 06/03/2018; acesso aos autos, Tribunal de Justiça de São Paulo, processo nº 1017321-95.2018.8.26.0100, 38ª Vara Cível da Capital.

[227] A concorrência entre as decisões autônomas de moderação pelas redes sociais e as decisões judiciais será desenvolvida no Capítulo 4-D.

propagandas (postagens pagas). Além disso, suas postagens seriam tratadas como sendo de "interesse noticioso" de partida, com a derrubada da publicação somente em casos de riscos reais e excepcionais de danos. Esta segunda regra não abarca, por outro lado, os anúncios pagos. Assim, anúncios pagos feitos por políticos não passam pela política de "checagem de fatos", mas devem observar as regras de vedação de conteúdo discriminatório[228].

Novamente, as regras do Facebook implicam um novo flanco de problemas: não apenas para realizar avaliações "holísticas e abrangentes" em episódios políticos potencialmente delicados, mas também para determinar quem será considerado um "político" para fins de aplicação da regra: se uma pessoa previamente banida do Facebook (um supremacista branco, por exemplo) decide concorrer a um cargo eletivo (governo, prefeitura ou conselho tutelar), ele terá direito de voltar a usar a plataforma?

A nova postura da empresa para regras de discursos políticos pode, na melhor hipótese, ser justificada como uma proteção ao debate democrático "livre, robusto e amplo" – mas também logo atraiu críticas de que essa deferência para com a classe política era fruto de interesses comerciais estratégicos do Facebook[229] ou, pior, a subestimação do risco de a plataforma contribuir decisivamente para um debate público no qual a verdade e os fatos importam cada vez menos[230]. Outros argumentam que a tarefa

[228] Segundo Nick Clegg, "vice-president of global affairs and communications" da empresa (e ex-vice primeiro ministro do Reino Unido):"Nós não acreditamos (...) que seja um papel apropriado para nós o de árbitros de debates políticos, de modo a evitar que o discurso de um político alcance sua audiência e se torne sujeito ao escrutínio e ao debate público (...) Nós iremos tratar os discursos de políticos como sendo de interesse noticioso que devem, como regra geral, serem vistos e ouvidos (...). Quando fazemos uma determinação quanto ao "interesse noticioso", nós avaliamos o valor do interesse público do discurso contra o risco de dano. Ao ponderarmos esses interesses, levamos em conta um número de fatores, incluindo circunstâncias específicas de alguns países, como se há uma eleição em curso ou se o país está em guerra, a natureza do discurso, incluindo se está relacionado à governança ou política; e a estrutura política do país, incluindo se há uma imprensa livre. Ao avaliar o risco de danos, nós iremos considerar a severidade do dano. Conteúdo que possui o potencial de gerar violência, por exemplo, pode implicar um risco de segurança que se sobreponha ao valor do interesse público. Cada uma dessa avaliações será holística e abrangente em sua natureza, e irá levar em conta padrões internacionais de direitos humanos" – "Facebook, Elections and Political Speech", artigo publicado em *Facebook Newsroom*, em 24/09/2019.

[229] Dave Willner, primeiro responsável pela construção dos padrões de comunidade do Facebook, classificou a postura da empresa como "covarde", pelo risco que traz a "grupos vulneráveis" – "The guy who wrote Facebook's content rules says its politician hate speech exemption is 'cowardice'" – reportagem publicada por *Wired*, em 30/09/2019.

[230] Cass Sunstein, "Facebook can fight lies in political ads", artigo publicado em *Bloomberg*, em 09/10/2019. Para Sunstein, "com alguma urgência, [o Facebook] deveria estar buscando novas maneiras de reduzir o risco de mentiras e falsidades minarem o processo democrático (...). É fácil entender a relutância do Facebook em operar como um ministério orwelliano da verdade. Mas 1984 é uma coisa; 2019 é outra. Contra suas vontades, o Facebook e outras plataformas de redes sociais estão contribuindo para uma

de “separar as falsidades”, bem como a “identificação de quais discursos são políticos”, é tão complexa que é impossível aplica-la de modo consistente na prática[231]. Nesse contexto, as redes sociais enfrentam sob novas circunstâncias alguns velhos dilemas, antes típicos de redações jornalísticas[232].

Ao mesmo tempo, ao enveredar por esse – inevitável – caminho, o Facebook se abre cada vez mais ao escrutínio público e a eventuais críticas sobre a consistência de seus posicionamentos – e das razões políticas que os movem. Nesse sentido, o fato de essas decisões serem tomadas em âmbito interno da empresa, sem um grau relevante de transparência quanto aos processos decisórios e seus critérios, passou a gerar críticas crescentes. Evelyn Douek ressalta que essa é uma das questões mais espinhosas da moderação de conteúdo. Na medida em que redes sociais se tornam fóruns importantes para que governos e políticos se comuniquem com o público, as empresas se vêm na posição de tomar decisões difíceis quando postagens desses atores violam suas regras.

> “*No mínimo, é muito problemático para uma empresa privada (e, na maior parte das jurisdições, uma empresa privada* estrangeira*) decidir que os cidadãos não possuem esse direito [de acesso ao discurso de governantes]. Não há maneiras fáceis de tomar decisões sobre como equilibrar os diferentes interesses que surgem nesses casos – por exemplo, quando um político usa linguagem odiosa, entram em conflito [o objetivo de] prevenção ao assédio ou ao discurso de ódio e o interesse do público de acesso a informações relevantes. Mas as decisões de plataformas nesses casos, até hoje, têm sido tomadas de modo* ad hoc *e frequentemente refletem as dinâmicas de poder do mundo real*”[233].

situação que diminui o poder da verdade no debate democrático todos os dias. Isso coloca em risco a própria democracia. A pergunta permanece: o que vamos fazer a esse respeito?”.

231 Siva Vaidhyanathan, “The Real Reason Facebook Won’t Fact-Check Political Ads”, artigo publicado pelo *The New York Times*, em 02/02/2019. O autor entende que, além de impraticável, essa conduta iria ferir as relações institucionais do Facebook com diversos governos pelo mundo, e defende que um ataque ao modelo de negócios da plataforma (com a proibição legal de microdirecionamento de anúncios) seria uma medida efetiva.

232 Thomas Kadri & Kate Klonick, “Facebook v. Sullivan: public figures and newsworthiness in online speech”, *Southern California Law Review Volume 93* (2019), p. 94.

233 Evelyn Douek, “Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation”, *Hoover Working Group on National Security, Technology, and Law – Aegis Series Paper No. 1903* (2019)

A opacidade de todo esse processo acaba dando margem para que essas decisões sejam percebidas como arbitrárias ou enviesadas. Soma-se a isso o fato de haver um histórico comprovado de inconsistências das decisões, além de idas e vindas nas formulações de regras, fatores que contribuíram para um problema de legitimidade. Um cenário que mina a autoridade das decisões, sejam elas quais forem.

3.D – A nova governança de discursos como um novo tipo de liberdade editorial

É possível refinar uma conceituação sobre *o que são* e *o que fazem* as redes sociais a partir de informações e das conclusões parciais já apresentadas. Abandonada qualquer ideia de que as plataformas sejam neutras, um grau de autonomia – que pode ser compreendida como uma *nova espécie de liberdade editorial* – vem à tona.

Essa é uma chave consistente para compreender esse (ainda em construção) novo papel estrutural que regula discursos entre os campos do permitido/proibido e do visível/invisível. Em sua afirmação de que as grandes redes sociais são "os novos governantes de discursos" ("the new governors") da era digital, Kate Klonick e Thomas Kadri traçam analogias com funções tradicionais – públicas e privadas – para caracterizar os papéis desses novos intermediários:

> *"Na época dominada pelos velhos governantes, a governança de discursos era essencialmente dividida entre o legislativo, o executivo, o judiciário e a imprensa (...). As decisões da imprensa a respeito de o que publicar – o que possuía interesse noticioso – eram feitas por conselhos editoriais a quem era dada alguma deferência pelas cortes (...). Hoje, na era dos novos governantes, nós podemos ver sombras desses vários papéis, mas em uma construção razoavelmente diversa. Grande parte da governança do discurso online é feita por plataformas privadas que exercem todos esses papéis – legislativo, executivo, judiciário e de imprensa – todos de uma vez"*[234].

[234] Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), p. 94. Ver também: Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), pp. 1662-1669. Destacam-se os seguintes trechos deste último artigo: "analisar as plataformas online pela perspectiva da governança é mais acurado do ponto de vista descritivo e também mais útil do ponto de vista normativo para abordar essa infraestrutura de espaço privado em constante evolução" (p. 1662); "essas plataformas tem um corpo centralizado, um conjunto de regra ou de leis estabelecidas, procedimentos *ex ante* ou *ex post* para adjudicação do conteúdo em face de suas regras, cultura e valores democráticos; suas políticas e regras são modificadas e atualizadas a partir de *inputs* externos; as plataformas são economicamente sujeitas à influência normativas de cidadãos-usuários e também são colaborativas com redes externas como governos e outras organizações (...) Talvez de modo mais significativo, a ideia de governança captura o poder e o escopo que essas plataformas privadas exercem por

De fato, uma compreensão sobre o estabelecimento de regras de moderação de conteúdo só faz sentido quando essa atividade é conjugada com *demais aspectos* da governança privada, como o estabelecimento das *condições de publicação* ou da *curadoria algorítmica* que define quais conteúdos serão exibidos com prioridade – e para quem. As próprias regras de permissão ou proibição de conteúdo levam em conta (e *devem* levar em conta) seus impactos na manutenção de um ambiente de livre debates, como demonstra a experiência do Twitter com regras de discursos de ódio (Capítulo 3-A), do Facebook com a consolidação da exceção do "interesse noticioso" (Capítulo 3-C) ou do Youtube sobre o valor informativo de vídeos violentos (Capítulo 3-A). As decisões que determinam a manutenção no ar de postagens em razão do interesse público em sua veiculação ou que as derrubam (ou invisibilizam) quando em casos de informações falsas ou outros tipos de riscos sociais próprios de ambientes digitais evidenciam esse aspecto de *novo tipo de julgamento editorial*[235].

Há um *novo tipo* de liberdade editorial[236] porque a analogia com as funções de "velhos governantes" só funciona até certo ponto. Pode haver semelhanças, mas também há diferenças significativas. Redes sociais são novos tipos de editores porque: a) ao contrário de editores tradicionais, não realizam uma curadoria específica a respeito do

---

seus sistemas de moderação de conteúdo e fornece a profundidade ('*gravitas*') de seus papéis em uma cultura democrática" (p. 1663).

[235] Randall Bezanson, no contexto da liberdade de imprensa tradicional, afirma que um "julgamento editorial" é "a atividade que reflete uma escolha independente de informações e opiniões atualmente relevantes, direcionada ao interesse público e motivada por propósitos que não sejam interesses próprios" – "The developing law of editorial judgment", *Nebraska Law Review Volume 78 (1999)*, pp. 857 e ss.

[236] Kate Klonick inicialmente defendeu que a categoria de "editores" no direito americano não seria a mais adequada a se aplicar ao Facebook, Twitter ou Youtube. Na ocasião, ela argumentava a partir do contexto específico de aplicação dessa categoria diante da jurisprudência da Suprema Corte americana sobre a Primeira Emenda, que garante um grau *praticamente absoluto* de independência editorial a veículos de imprensa, colocando-os em posição análoga ao de "speakers" em geral, a partir de Miami Herald v. Tornillo (1974*)* – decisão que rechaçou por inconstitucional a possibilidade de se determinar um *direito de reposta* frente àquele jornal. O fato de as plataformas não produzirem conteúdo em nome próprio seria determinante para essa posição inicial da autora, embora ela tenha consignado que esse entendimento pudesse avançar na medida em que "a presença das redes se expandisse no discurso online" – "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1660. Posteriormente, em artigo publicado em coautoria com Thomas Kadri, esse paralelo parece ter sido incorporado sim à descrição da função estrutural/institucional das redes sociais, por meio de uma referência esclarecedora com base no caso New York Times Co v. Sullivan: "plataformas são ao mesmo tempo governantes [de discursos], definindo políticas de discursos e adjudicando disputas em torno deles, e publishers, controlando o acesso a discursos em nome de quem pretende falar e de seu público. Eles são [ao mesmo tempo] a corte de *Sullivan*, e eles são o *New York Times*" – Thomas Kadri & Kate Klonick, "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019), pp. 94-95.

que será publicado, pois seu funcionamento opera pela lógica geral e inversa de livre postagens; b) a intervenção editorial de políticas de moderação costuma ser feita posteriormente, levando em conta a avaliação de riscos ou problemas que surjam, ou a partir de reclamações. Há intervenção editorial, mas sob uma lógica mais excepcional, que não determina de partida qual conteúdo "entra" na plataforma.

Se um artigo favorável ao nazismo ou a favor da "teoria" de que o planeta é plano é enviado a um jornal, entende-se que o veículo possui a liberdade editorial de não o publicar, pois realiza a curadoria específica sobre "o que entra". No caso de uma rede social, ao permitir um espaço de publicação livre, por sua própria vontade e liberdade empresarial, a empresa exerce um novo tipo de controle editorial: por meio ele, ela terá que justificar eventualmente quais os motivos que justificam uma restrição a (ou distribuição reduzida de) um determinado tipo de conteúdo. Nesse caso, o processo editorial funciona por outra lógica: ao não escolher especificamente "o que entra", a plataforma não exerce uma "coautoria" sobre o conteúdo publicado. Mas ao trabalhar com a premissa diversa de livre publicação, terá de fornecer as razões que impedem a postagem de conteúdo que considera problemático – conceito este que não necessariamente irá coincidir com critérios de legalidade ou ilegalidade[237].

As diferenças entre a *função de editores tradicionais* e essa *nova função editorial das grandes redes sociais* justificam porque os direitos e deveres decorrentes dessas atividades também devem ser distintos – possibilitando (e demandando) diferenças de tratamento jurídico, notadamente pelo legislativo e pelo judiciário. Elas ajudam a explicar, dadas essas *especificidades institucionais*, porque está correto um modelo mais favorável de responsabilização civil de intermediários digitais, tal como será abordado no próximo capítulo. E iluminam também porque é falsa uma oposição dual entre "controle editorial tradicional" e "neutralidade", que não possui lastro na realidade.

Uma analogia absoluta com editores tradicionais, portanto, carece de sentido. No entanto, no contexto específico do direito americano, empresas de tecnologia invocam a categoria de "publishers" ou "editores" (tradicionais) como parte de uma estratégia para menor responsabilização jurídica por suas atividades. Google e Facebook argumentam com forte grau de sucesso ao judiciário que suas práticas comerciais integram sua liberdade editorial. Esse argumento adquire um *sentido específico naquele país*, em razão

[237] Essa ideia em especial será desenvolvida a partir do Capítulo 4-C.

de sua jurisprudência constitucional: ali, a categoria de "editores" evoca proteção da Primeira Emenda e garante uma forte liberdade que serve de escudo contra regulações estatais, levando a um ponto no qual a liberdade de expressão se confunde ampla liberdade empresarial, reforçada pela lógica da "state action doctrine"[238]. No Brasil, a posição análoga a um editor tradicional, em sentido contrário, implicaria uma maior responsabilização jurídica do intermediário pelo conteúdo publicado[239].

A noção de liberdade editorial, mesmo em sua nova forma, implica necessariamente escolhas valorativas, fruto de algum grau de autonomia. No jornalismo tradicional (que produz conteúdo em nome próprio), isso significa que uma noção absoluta de objetividade jornalística não é possível, pois a própria seleção de fatos, de termos utilizados e de recortes de análise incorporam vieses valorativos, decorrentes daquela própria liberdade editorial[240]. No caso das grandes redes sociais, também é possível identificar como valores norteiam suas governanças de discursos, compondo esse novo tipo de exercício de liberdade editorial. O Twitter, em seus anos iniciais, incorporava um "ethos" libertário de forte tradição nos Estados Unidos, historicamente oposto ao controle governamental: a experiência fracassou e revelou como vetores de regulação da própria infraestrutura de expressão na internet (mercado e normas sociais, especialmente) impuseram uma guinada de rumo. O Facebook tampouco conseguiu escapar da necessidade de diminuir o alcance ou de vedar a publicação de discursos com base em vieses valorativos que atendam a uma concepção de interesse público: a ocultação de páginas com informações falsas e potencialmente danosas à saúde pública a respeito de vacinas (Capítulo 2-F), por exemplo, pode ser comparada à hipótese de um jornal que se recusa a publicar um artigo que veicule essas opiniões.

Para além disso, a ideia de liberdade editorial ajuda a esclarecer porque decisões autônomas desses entes privados permitem abordar e atacar alguns problemas reais e concretos. Ao se deparar com um problema relevante de desinformação (postagens que

---

[238] A esse respeito, ver: Heather Whitney, "Search engines, social media, and the editorial analogy", artigo publicado por *Knight First Amendment Institute – Columbia University*, em 27/02/2018, no qual autora descarta a analogia com editores tradicionais para mecanismos de busca ou redes sociais. O Facebook chegou a invocar para si, em 2018, a posição de "publisher" em um caso que não tratava propriamente de discursos, mas sim do nível de acesso a dados de usuários que são fornecidos a desenvolvedores de aplicativos para sua plataforma – "Is Facebook a publisher? In public it says no, but in court it says yes", reportagem publicada por *The Guardian*, em 03/07/2018.

[239] Capítulo 4-C.

[240] Rodrigo Vidal Nitrini, *Liberdade de informação e proteção ao sigilo de fonte: desafios constitucionais na era da informação digital*, Hucitec Editora, 2016, pp. 117-119.

associam falsamente vacinas a autismo, por exemplo), uma plataforma *não precisa se engajar em uma postura muito abrangente que tenha a pretensão de sempre separar verdades e falsidades em distintas situações*. As eventuais restrições podem ser plenamente justificadas por conta de riscos concretos que se apresentam em cada plataforma: uma empresa pode vedar postagens com risco de viralização e desinformação na área de saúde pública, sem que isso signifique que irá tomar uma medida análoga em todas as postagens que contenham falsidades – como em postagens sobre astrologia, por exemplo[241].

É nesse exato sentido que deve ser compreendida a derrubada de postagens do presidente da república brasileiro em meio à pandemia do novo coronavírus, cujo teor foi considerado como desinformação apta a causar "danos reais" no bojo da crise sanitária. Nas postagens, Bolsonaro defendia o uso do medicamento cloroquina (então, sem comprovação científica de sua eficácia para tratamento da doença Covid-19) e criticava medidas de isolamento social, indo naquele momento de encontro ao consenso estabelecido por autoridades de saúde internacionais e também brasileiras. Dias antes, postagens do presidente venezuelano Nicolás Maduro que indicavam uma receita caseira para combate à doença também foram apagadas sob os mesmos fundamentos. Essas derrubadas foram feitas pelas grandes plataformas que são objeto deste estudo, em meio a mais uma ação conjunta e coordenada por essas grandes empresas de tecnologia para aplicação de critérios de moderação de conteúdo[242]. Essas derrubadas, no entanto, não significam um silenciamento completo dos discursos políticos desses dois presidentes, que continuam possuindo amplos meios, inclusive oficiais, para divulgarem suas ideias. Ainda assim, caracterizam essa disposição das grandes redes sociais de realizarem um nível de controle editorial *dentro de seus ambientes*, que não se confundem com a internet em geral, diante de riscos que são próprios a suas atividades (tal como a viralização).

A própria noção de liberdade editorial implica que existe uma *margem discricionária*: é possível argumentar que uma rede social possui a liberdade de *manter ou não* páginas que disseminem a "tese" de que o planeta é plano – e que qualquer das decisões possíveis pode mudar a partir de novos cenários referentes a riscos ou percepções

[241] Os exemplos de *desinformações sobre vacinas* e de *postagens sobre astrologia* foram retirados de "thread" no Twitter publicada por Kate Klonick (@Klonick) em 23/01/2020.

[242] "Depois do Twitter, Facebook e Instagram também apagam post de Bolsonaro", reportagem publicada pelo jornal *Folha de S. Paulo*, em 31/03/20; "Covid-19 misinformation remains difficult to stop on social media", reportagem publicada por *Forbes*, em 17/04/20.

sociais sobre esse problema[243]. *Mas é justamente aí que reside uma das consequências jurídicas mais relevantes dessa concepção: junto ao exercício desse novo tipo de liberdade editorial, emerge o ônus de demonstrar as razões pelas quais eventuais restrições de publicação são impostas*. O reconhecimento da liberdade editorial traz consigo a necessidade de motivar que as restrições são razoáveis, dando-se a devida consideração ao direito de liberdade de expressão dos usuários das plataformas[244].

Por isso, o que está em jogo não uma definição não sobre se as plataformas possuem um algum grau de autonomia e de liberdade editorial, mas sim sobre os limites dessas liberdades e até onde elas devem ir em problemas e contextos específicos – como no caso de anúncios políticos, de campanhas de desinformação em áreas como saúde pública ou em searas como de discursos de ódio. Essa configuração de novas funções é causa de um período de estranhamento com os papéis dos novos intermediários digitais do debate público. O ecossistema da livre expressão foi profundamente alterado e está em curso um processo de readequação de normas, padrões, expectativas e responsabilidades em tornos desses novos atores.

As posturas e iniciativas das plataformas acabam sendo naturalmente objetos de críticas e cobranças por parte do público – *muitas vezes contraditórias entre si, como em casos de oposição entre pedidos de mais controle ou de menos controle sobre a publicação de discursos*[245]. Como diz Nicolas Suzor: "não há resposta fácil ainda sobre o que diferentes sociedades esperam da parte das plataformas de mídia digitais"[246]. Não

[243] Páginas que advogam que o planeta é plano podem ficar circunscritas à esfera do ridículo das inevitáveis franjas da internet. Mas considere-se por um momento a hipótese de que a ideia começa a ganhar tração entre crianças e jovens; uma situação do tipo poderia motivar uma nova postura editorial mais restritiva quanto a esse tema.

[244] Ver Capítulo 4.

[245] Como mostra a polêmica em torno da decisão do Facebook de não checar a veracidade de anúncios políticos pagos nos Estados Unidos: a American Civil Liberties Union (ACLU) concordou com a postura, que por sua vez atraiu críticas da Electronic Frontier Foundation (EFF). "Essa não será uma opinião popular. Mas no geral, achamos que o Facebook acertou (...). Não acredito que sejam capazes de fazer uma checagem (´fact checking´) efetivo, e não acho que a sociedade em geral deva querer que o Facebook seja a entidade fazendo esses tipos de distinção", disse Ben Wizner (ACLU). Para Genie Gebhart (EFF), "o Facebook rápida e amplamente se coloca no papel de árbitro de discursos para o público em geral, mas, preocupado com jogos políticos, aplica um grupo de regras mais permissivo para grupos políticos poderosos que reclamam de seus vieses" – "FEC Comissioner rips Facebook over Political Ad Policy: 'This will not do'", reportagem publicada pelo portal da *National Public Radio*, em 09/01/20.

[246] Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p. 163. Ainda: "Há muitos grupos de interesse distintos com diferentes prioridades e agendas, e os intermediários frequentemente se vêem sujeitos a demandas conflitantes por variados atores" (p. 124).

se trata de um processo inédito: o histórico da liberdade de imprensa no direito constitucional fornece paralelos relevantes.

À época de sua inscrição aos textos fundantes do constitucionalismo moderno, nos séculos XVIII e XIX, a liberdade de imprensa era compreendida como a liberdade de imprimir qualquer texto sem a necessidade de licença prévia do governo. Em uma época em que estados monárquicos começavam a perder seu poder absoluto para definir o que seria publicado, cidadãos (na esfera privada) passavam a ter o direito de usar da tecnologia de impressão popularizada por Gutenberg. O alvo de John Milton, ao escrever "Aeropagitica*"* em 1644, era o Parlamento inglês, que no ano anterior havia restituído um sistema de licença prévia. Essa compreensão inicial foi sendo alterada e expandida com a evolução no século XIX da infraestrutura de expressão: com a popularização de jornais, ainda marcadamente politizados e panfletários, esse conceito passou a incorporar também uma dimensão substantiva de resguardo da livre discussão democrática, especialmente a partir da experiência republicana norte-americana. Quando a imprensa profissional se consolida ao longo do século XX – entre outros motivos, graças a inovações como telégrafos, telefones, malhas viárias e um modelo de negócios calcado na publicidade – a ideia de liberdade de imprensa adquire um sentido distinto e ligado às funções do jornalismo moderno, que à medida em que adota padrões profissionais e éticos próprios, especialmente em torno da ideia de objetividade jornalística, passa a reivindicar uma identidade institucional específica e prerrogativas próprias, tal como o sigilo de fonte[247]. Para Horwitz, "jornais evoluíram (...) por um complexo e constante sistema de interações com o mundo social à sua volta. Eles se tornaram importantes para o discurso público precisamente porque eles evoluíram para se tornarem importantes para isso – e o discurso público, por sua vez, evoluiu para acomodar esse papel evolutivo"[248].

Balkin também sublinha como uma comparação com o processo de consolidação da imprensa profissional no século XX, cuja "visão de objetividade jornalística (...) não surgiu de uma noite para outra", ajuda a compreender essa *definição conceitual ainda em curso sobre as funções e responsabilidades públicas das grandes redes sociais*. Para ele, assim como os jornais, as redes sociais se apresentam como mais do que empresas voltadas exclusivamente a empreitadas lucrativas:

---

[247] Rodrigo Vidal Nitrini, *Liberdade de informação e proteção ao sigilo de fonte: desafios constitucionais na era da informação digital*. Hucitec Editora, 2016, pp. 42-56.

[248] Paul Horwitz, *First Amendment Institutions*, Harvard University Press, 2012, pp. 77-78

> "*elas explicam que usam sua expertise em tecnologias especiais para promover objetivos de interesse público, como acesso ao conhecimento, liberdade de expressão e construção de comunidades (..). Nesse sentido, encorajam a ideia de que agem e deveriam agir de acordo com normas profissionais que levam em conta o interesse público. Além disso, empresas de redes sociais e de mecanismos de busca invocam essas normas profissionais e de interesse público para justificar suas decisões para organizar resultados de mecanismos de busca, para realizar a curadoria de discursos públicos e para aplicar (ou às vezes se abster de aplicar) normas de civilidade (...) Estamos começando a ver uma lenta e hesitante evolução da auto-compreensão das plataformas exatamente nessas linhas* "[249].

A imprensa profissional do século passado se consolidou após um longo processo, no qual suas prerrogativas e responsabilidades foram definidas também a partir das funções a que se propôs exercer. As grandes redes sociais ainda articulam suas respostas aos problemas que lhes causaram graves crises de imagens nos anos recentes, influenciadas pelas novas expectativas e cobranças sociais que comumente destacam a alta concentração de poder em suas mãos sobre os debates públicos na esfera digital.

3.E – Considerações finais do capítulo

Feita a defesa de uma conceituação da nova governança de discursos das grandes redes sociais como um novo tipo de liberdade editorial, torna-se importante sublinhar que *nenhum dos argumentos apresentados até aqui deve ser interpretado como uma carta branca para a definição privada sobre regras de discursos públicos*, mesmo em ambientes corporativos. Ao delinear essa governança privada de intermediários digitais que se desenvolveram com relativa autonomia, colocam-se também importantes desafios sobre o poder dessas plataformas, riscos para a liberdade de expressão e, principalmente, indagações sobre as possibilidades existentes para a tutela e preservação de direitos fundamentais. Ainda neste tópico, já foi destacada a noção de que o exercício desse novo

[249] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118, n. 7* (2018), pp. 2042-2043. Para ele, "assim como no caso da mídia de massa do século XX (...) o estado não pode impor constitucionalmente essas normas – ou seus equivalentes do século XXI – nos curadores digitais. Isso não significa que o público não possa ou que não deva demandar essas normas. Estamos começando a ver uma lenta e hesitante evolução da auto-compreensão das plataformas exatamente nessas linhas. Esse processo de aprendizagem é o resultado de onda após onda de pressão pública em companhias como Google, Facebook e Twitter, normalmente facilitadas por jornalistas que aplicam eles próprios as normas profissionais de reportagem desenvolvidas no século anterior".

tipo de liberdade editorial deve implicar também o ônus argumentativo de apresentação das razões pelas quais se impõe restrições ao direito de publicação de usuários.

Portanto, como o direito pode responder à emergência dos "novos governantes" (transnacionais) de discursos? Quais são as possibilidades e desafios que se apresentam para resguardar direitos fundamentais, notadamente a liberdade de expressão? Redes sociais devem ser completamente livres para definirem suas regras de discursos, desde que respeitem eventual limite de ilicitude de um determinado país? Ou, ao contrário, devem permitir todo e qualquer conteúdo que não seja ilícito?

O próximo capítulo pretende avançar frente a essas questões, abordando-as tanto no contexto transnacional quanto do direito brasileiro, a partir do marco teórico do constitucionalismo digital.

Capítulo 4 – Constitucionalismo digital e as perspectivas para políticas de moderação de conteúdo de redes sociais pautadas por direitos fundamentais

As regras de discursos das grandes redes sociais são estabelecidas em um plano autônomo e transnacional, mas que responde a variados incentivos: mercado, normas sociais, imperativos de códigos e, também, leis e regras jurídicas dos variados países. Nesse dinâmico e complexo cenário normativo, coloca-se a pergunta sobre como o direito pode responder aos variados desafios apresentados até aqui.

Para responde-la, este capítulo apresenta o marco teórico do *constitucionalismo digital*, que pretende renovar para o ambiente da internet duas preocupações centrais do constitucionalismo tradicional: limitação de poderes e garantias de direitos. Seu foco, em especial, é a elaboração de processos de governança legítimos, que respondam aos imperativos de um devido processo digital e de restrições justificadas a direitos fundamentais.

Em seguida, serão apresentados argumentos que articulam parâmetros normativos do constitucionalismo digital em torno das políticas de moderação de conteúdos das grandes redes sociais em duas searas: a) no âmbito do "direito das plataformas" das redes sociais, englobando regras, processos e desenhos institucionais para a aplicação dessas regras, e; b) no âmbito do direito brasileiro, apontando critérios de atuação do poder judiciário e de atualização das regras legislativas, a partir dos dispositivos vigentes do Marco Civil da Internet.

4.A – Constitucionalismo digital: conjugando os planos nacionais e transnacional a partir da lógica de direitos

A emergência dos grandes atores digitais – privados e transnacionais – acompanha um ceticismo acentuado quanto à capacidade de regras jurídicas tradicionais (criadas em torno da figura dos estados) de lidarem, por si só, com os *riscos e ameaças a direitos fundamentais* derivados desses "novos poderes". Por isso, é tão apenas lógico que esses atores privados se tornem eles próprios alvos de pressões para que tomem suas decisões autônomas que impactam direitos fundamentais *levando-os em conta e*

assegurando a eles um nível adequado de proteção, por meio de *processos decisórios legítimos*[250].

Essencialmente, é essa preocupação com os novos poderes privados da rede mundial que tem caracterizado a consolidação do conceito de *constitucionalismo digital:* a defesa de um marco teórico que pretende renovar dois objetivos tradicionais do constitucionalismo – garantia de direitos e limitação de poderes – diante de uma realidade cujo palco é agora compartilhado por estados e também grandes empresas de tecnologia (Capítulo 1). Como explica Edoardo Celeste:

> "*Constitucionalismo digital é um conceito que se refere a um contexto específico, o ambiente digital, no qual atores privados emergem ao lado de estados-nação como potenciais violadores de direitos fundamentais. Essa particularidade do ambiente digital requer que o conceito de constitucionalismo se liberte da dimensão estatal de modo a dimensionar adequadamente a emergência dos poderes de atores privados*"[251].

Nesse cenário, o constitucionalismo digital enfrenta o desafio particular de:

> "*regular poderes que estão distribuídos entre muitos atores, dentro de sistemas complexos, com diversos componentes de interação. Para governos, isso significa repensar de modo radical como a regulação pode operar em um ambiente descentralizado – onde o estado não é o único, ou mesmo o mais poderoso, ator que procura regular comportamentos*"[252].

---

[250] Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 107-108.

[251] Edoardo Celeste, "Digital Constitutionalism: mapping the constitutional responses to digital techonology´s challenges", *HIIG Discussion Paper Series No. 2018-02 (2018)*. Nesse texto, Celeste apresenta e analisa as diferentes propostas de diversos autores em torno do conceito de constitucionalismo digital, antes de apresentar sua própria definição. De um modo geral, todas elas convergem com os elementos apresentados acima, salvo poucas exceções laterais. Essa referência bibliográfica serve, de todo modo, para um mapeamento geral do uso do termo na literatura.

[252] Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 164-166. Ainda: "mesmo quando obrigações legais são introduzidas, iniciativas especializadas da indústria tornam-se provavelmente necessárias para monitorar uma adequação às regras ('compliance') e definir boas práticas que não são sempre possíveis ou factíveis por meio de um processo regulatório governamental. Há muito espaço para intermediários tomarem diferentes decisões sobre como eles arquitetam seus sistemas, quais regras escolhem e como determinam suas aplicações, mas o requisito mais importante é de que se tornem de alguma maneira responsabilizáveis por suas escolhas e que forneçam as justificativas devidas" (p. 124).

Por isso, para além de respostas normativas tradicionais, são necessários "instrumentos [jurídicos] inovadores que emergem nesse contexto transnacional"[253]. Essas reações normativas podem se dar em diferentes níveis: em dimensões "tradicionais", construídas em torno dos estados (incluindo ordenamentos jurídicos nacionais ou organizações regionais/internacionais), *mas também em dimensões desatreladas desse paradigma*, como em sistemas de governança ou organizações transnacionais de resoluções de disputas digitais (como o ICANN[254]) ou *no âmbito das regras autônomas dos atores privados globais*[255].

Um modo consistente (e talvez especialmente esclarecedor) de apresentação do conceito de constitucionalismo digital é: uma atualização da noção de eficácia horizontal de direitos fundamentais entre particulares para o ambiente digital. Como essas novas relações jurídicas – que implicam ameaças a direitos fundamentais e a necessidade de limitação de poderes – não ocorrem mais apenas no interior de um ordenamento jurídico nacional, *o constitucionalismo digital enfatiza a necessidade da eficácia de direitos fundamentais nas relações jurídicas entre usuários de internet e esses atores privados transnacionais, que possuem relevante capacidade normativa autônoma*.

---

253 Edoardo Celeste, "Digital Constitutionalism: mapping the constitutional responses to digital techonology´s challenges", *HIIG Discussion Paper Series No. 2018-02* (2018), pp. 17-19. Em sentido complementar: "há um papel para cortes e legislativos aqui, mas também há a necessidade para novas instituições que possam canalizar as pressões sociais na governança feita no dia-a-dia, para a qual o sistema legal é muito rígido. Essas novas instituições requerem alguma imaginação – nós teremos que inventá-las" – Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p 171.

254 Em português, "Corporação da internet para atribuição de nomes e números" – uma entidade sem fins lucrativos sediada nos Estados Unidos que é tecnicamente responsável pela alocação e coordenação do espaço de endereços de protocolo de internet (IP) na internet, possibilitando também o uso unificado de DNS (sistema de nomes de domínio, em português), desde os anos 1990. O ICANN costuma ser descrito com um regulador transnacional da internet porque desenvolveu regras autônomas, desatreladas da base legal americana, com efeitos transnacionais, para resolver disputas em torno de sua área de competência. O fato de historicamente ter relações formais com órgãos do governo americano foi causa de críticas e pressões para que sua gestão fosse multilateral e independente daquele país – www.icann.org.

255 Edoardo Celeste, "Digital Constitutionalism: mapping the constitutional responses to digital techonology´s challenges", *HIIG Discussion Paper Series No. 2018-02* (2018), pp. 17-19. Para uma referência teórica mais abrangente – que também considera a coexistência de ordens sociais e constitucionais autônomas no nível transnacional, que desenvolvem parâmetros normativos próprios com impactos sobre direitos fundamentais e são desatreladas de estados nacionais, ver: Gunther Teubner, *Constitutional Fragments: societal constitutionalism and globalization*, Oxford University Press, 2012. Teubner reconhece no constitucionalismo digital uma resposta à digitalização, privatização e globalização que marcam o ambiente digital – embora enfatize que o direito oficial do estado sempre conviveu com atores sociais inseridos em sistemas normativos autônomos, cujas regras reagem de modo dinâmico com o direito tradicional, em processo de influência mútua.

Nicolas Suzor, também a partir da ideia de constitucionalismo digital, é o autor que melhor especificou propostas normativas que ajudam a fazer frente ao problema de pesquisa. Ele evoca *duas linhas principais* para um *processo de constitucionalização* das grandes plataformas digitais.

A primeira delas decorre do diagnóstico de que essas plataformas operam atualmente em um terreno em larga medida *à margem do direito* ("lawless way"), porque as empresas possuem uma larga discricionariedade para criarem e aplicarem regras da maneira que acharem melhor, com uma quase inexistente possibilidade de responsabilização ("accountability") por suas decisões. Isso permite que sejam "arbitrárias, caprichosas, imprevisíveis e inconsistentes". Para Suzor, essa é a "antítese da maneira jurídica de se tomar decisões", pois a ideia central do "rule of law" é que as pessoas saibam as razões pelas quais as decisões que as afetam são tomadas, por meio de regras previamente definidas e aplicadas de modo isonômico[256].

As garantias do "rule of law" evoluíram em uma época no qual o estado-nação era o ator mais poderoso para governar as vidas de cidadãos. Como essa premissa não corresponde mais aos fatos no ambiente digital global, Suzor enfatiza a necessidade de aplicar garantias análogas frente às grandes plataformas[257]. A constitucionalização surge atrelada ao ideário do "estado de direito" – um processo que se caracteriza pela "internalização auto-reflexiva de limites aos poderes de um corpo autônomo":

> *"Constitucionalização em última instância envolve um paradoxo pelo qual o poder de governança se torna legítimo porque aqueles que exercem o poder acreditam que estão eles próprios vinculados a suas próprias regras"*[258].

---

256 Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 106-110. Ainda: "os princípios de direito mais ampla e comumente reconhecidos sob o 'rule of law' são proteções procedimentais. Isso essencialmente requer que as regras sejam aplicadas de maneira isonômica e previsivelmente. No mínimo, isso significa que as pessoas devem estar cientes das regras e saber as razões pelas quais uma decisão que as afeta foi tomada"; "legitimidade, quando falamos de decisões por aqueles que têm poder sobre nós, significa que decisões sejam feitas de acordo com um conjunto de regras e que quem toma a decisão possa ser responsabilizado".

257 "Elas [as empresas] não nos governam da maneira que governos o fazem – não criam impostos ou nos aprisionam – mas elas determinam sim as regras sobre como conversamos entre nós, quais informações estão disponíveis para vermos, o que é removido ou interessado, com quem podemos conversar, ou quanto podemos ser ouvidos. Elas arbitram disputas entre nós e, quando nos punem, nos isolam de amigos, familiares e nossas audiências. As maneiras como elas nos influenciam, por sua vez, é influenciada por forças de mercado, por governos ao redor do mundo, por cada uma delas e por seus usuários, além de pressão social pela mídia, grupos da sociedade civil e o público em geral" – Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p. 111.

258 Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 110-120. Ainda: "Quando falamos sobre a constitucionalização de empresas de tecnologia, elas

Para o âmbito do problema de pesquisa (decisões de moderação de conteúdo por redes sociais), esse enfoque se traduz imediatamente na necessidade de regras claras e pré-definidas, que decorram de decisões motivadas. A partir dele, é possível pensar em um *devido processo digital*. Por ora, é importante sublinhar que garantias procedimentais compõem uma abordagem crucial para conciliar as preocupações inerentes ao constitucionalismo digital com a existência de algum grau necessário de autonomia que as plataformas devem ter para a formulação de suas regras de atuação.

Como já apontado no Capítulo 3-C, uma maior legitimação dos processos decisórios e de aplicação de regras é especialmente importante para casos que envolvem dilemas em torno da liberdade de expressão, que surgem a partir da necessidade de traçar linhas de permissão ou proibição de discursos e inevitavelmente dão margem a discordâncias legítimas e inconciliáveis. Mencionando o caso Facebook/Myanmar (Capítulo 3-B), Suzor aponta:

> "*Podemos dizer que o Facebook poderia ter feito um trabalho melhor identificando e moderando discursos de ódio em Myanmar, mas discordar sobre onde exatamente deveria ser traçada uma linha entre expressões políticas válidas e discursos que promovem o genocídio. Alguns pensam que a linha deveria ser traçada pela incitação – que o Facebook deveria apenas proibir postagens que diretamente encorajam a violência. Outros pensam que postagens que desumanizam uma minoria étnica ou religiosa são precursoras de atos violentos e deveriam ser proibidos. (...) A história é a mesma para outras linhas traçadas na areia*"[259].

Essa posição normativa pela garantia de um processo de constitucionalização não se confunde, portanto, com posições substantivas rígidas[260]. Mas garantias

não são totalmente autônomas nem são sistemas independentes; seus limites internos de poder não são autossuficientes, mas são reforçados e baseados em leis externas. Mas, porque esses sistemas têm de fato um grau substancial de autonomia, para que limites sejam efetivos eles devem também partir de dentro".

[259] "(...) Podemos debater se é melhor para o Google prover resultados de busca censurados na China ou não prover serviço nenhum" – Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 129-130.

[260] A ideia de devido processo digital como fundamento do constitucionalismo digital aproxima-se bastante também de perspectiva que busca superar os impasses resultantes da governança de discursos pelos intermediários digitais por meio de *soluções de processo*. John Bowers e Jonathan Zittrain argumentam que a governança da internet está entrando em uma "terceira era". Para os autores, entre os anos 1990 até cerca de 2010, as preocupações à frente de uma regulamentação do espaço online eram calcadas pela "era dos

processuais e procedimentais, embora cruciais para limitação de poderes, são inúteis se completamente desacompanhadas de direitos materiais.

Por isso, Suzor também argumenta pela incorporação, por esse processo de constitucionalização, de uma lógica de proteção aos direitos humanos. Ao fazê-lo, porém, ele novamente não advoga pela adoção de critérios substantivos, mas sim por *uma lógica de respeito a direitos fundamentais, a partir da gramática dos direitos humanos*. Isso significa que restrições à liberdade de expressão devem ser *devidamente motivadas* e que a articulação dessas razões pode se valer do repertório jurídico dos direitos humanos:

> "*Proteger direitos humanos não significa que plataformas não devam definir regras ou que elas devam todas definir as mesmas regras, mas sim que as regras que aplicam sejam defensáveis. Logo, enquanto a liberdade de expressão é um direito humano, isso não significa que empresas de tecnologia devam sempre permitir usuários postarem o que quiserem. Empresas podem legitimamente definir políticas que restringem a liberdade de expressão para fins válidos*"[261].

direitos" – quando proteções ao discurso online eram construídas para resguardar esses nascentes espaços de riscos de coerção ou interferência, especialmente por governos. A partir de 2010, começaram a se avolumar visões críticas contra as poucas empresas gigantes que dominam a internet e que desenvolvem novas formas de controle de discursos – filtragens e curadoria algorítmicas – e seus efeitos problemáticos – viralização de discursos como armas ("weaponized speech"), desinformação. Esses últimos anos marcaram a "era da saúde pública", na qual as novas sensibilidades passaram a demandar que proteções de discursos da "era dos direitos" cedam a considerações feitas em nome de riscos e danos causados a outras pessoas, instituições e à sociedade geral. Desse contraste, começa a nascer uma terceira era de governança da internet – a "era do processo", que foca no "desenvolvimento de mecanismos legitimados de modo amplo e que alcancem comprometimentos factíveis entre direitos, saúde pública e um número qualquer de outras considerações emergentes (...). Afastar-se de um conflito improdutivo e frustrante entre esses dois discursos aparentemente incompatíveis – direitos e saúde pública – significa, para começar, adotar modelos de governança que incorporem um novo tipo de 'profissionalismo' na governança de conteúdo". Isso pode significar uma reorientação dos processos decisórios de moderação de conteúdo, que leve em conta os interesses de usuários, a delegação de certas decisões para além das plataformas, ou novas possibilidades que conjuguem esses dois princípios, pois isso "representa nossa melhor chance de construir legitimidade sobre como conteúdos são distribuídos, filtrados e ordenados" – John Bowers & Jonathan Zittrain, "Answering impossible questions: content governance in an age of disinformation", artigo publicado por *The Harvard Kennedy School (HKS) Misinformation Review, em 14/01/20*. Acesso em https://doi.org/10.37016/mr-2020-005

[261] **"**Porque direitos humanos são sempre contingentes e dependentes de contexto, eles não tentam ditar uma resposta certa sobre quais regras podem ser definidas. Pelo contrário, direitos humanos fornecem uma ferramenta conceitual para garantir que regras que impactam direitos sejam bem desenhadas para o contexto particular no qual se aplicam – Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 130-131. E também: "os padrões fornecidos pelo direito internacional dos direitos humanos não são perfeitos. Eles não abrangem todas as questões para as quais diferentes grupos possam querer melhorias; eles normalmente não ditam um resultado substancial em particular; e não vinculam as empresas sob qualquer consequência legal. Mas são o mais perto do que temos como um denominador comum universal de princípios com os quais quase todos nós podemos concordar" (p. 147).

Para Suzor, um benefício chave dessa abordagem seria a incorporação pelas regras e políticas das plataformas de maiores graus de proteção a minorias e grupos vulneráveis. Ele considera que as regras de discursos do ambiente digital normalmente partem de uma mentalidade liberal valorizadora de uma concepção de igualdade formal, que ignora impactos desiguais provocados em diferentes contextos sociais; "(...) a regra que proíbe nudez feminina no Instagram e no Facebook, quando aplicada a imagens de cerimônias culturais indígenas, trabalha para silenciar um grupo já marginalizado", diz. A base conceitual do direito dos direitos humanos, construída sobre instrumentos e tratados internacionais que protegem grupos vulneráveis, também poderia justificar as razões para se tratar discursos supremacistas brancos de modo distinto de uma postagem que diz que "todos os brancos são racistas" como forma de problematização do racismo[262].

David Kaye, relator especial da Organizações das Nações Unidas sobre as liberdades de expressão e de opinião, também advoga pela vinculação de plataformas digitais globais a parâmetros de direitos humanos. Kaye defende que as empresas tornem o direito internacional dos direitos humanos o padrão explícito sobre o qual definem suas regras de moderação de conteúdo, especificamente. A esse respeito, menciona que o Facebook e o Twitter, ao menos desde 2018, publicamente declaram que usam essa base normativa como referência para suas decisões. Para ele, a incorporação dessa referência pode não apenas dar mais legitimidade a suas decisões diante da sociedade, mas também servir como um argumento público poderoso contra demandas autoritárias de governos pela derrubada de materiais, que poderiam acabar sendo publicamente constrangidos por normalmente serem signatários de tratados internacionais[263].

Para Kaye, a lógica dos direitos humanos obrigaria empresas de tecnologia a justificar a adoção de critérios ou regras que restrinjam de maneira intensa a liberdade de

---

[262] Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 134-135; "Um comprometimento sério com os direitos humanos reconheceria que as regras devem levar em conta essas diferenças estruturais e fornecer às plataformas uma base para articular e justificar políticas feitas para diminuir os riscos de que continuem, reflitam ou amplifiquem a desigualdade social existente".

[263] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 118-121; "O direito internacional dos direitos humanos garante a todas as pessoas o direito de procurar, receber e transmitir informações e ideias de todos os tipos, independente de fronteiras. Protege o direito de todas as pessoas a opiniões sem interferência. Tão importante quanto, o direito dos direitos humanos dá às empresas a linguagem para articular suas posições globalmente em maneiras que respeitam as normas democráticas e contrariam demandas autoritárias".

expressão, explicitando as razões da não adoção de alternativas menos invasivas. Por fim, ele aponta que as empresas podem se valer de fontes formais do direito internacional dos direitos humanos, com sua vasta jurisprudência em diversos sistemas regionais e órgãos internacionais, para a tomada de decisões[264].

Tomadas em conjunto, essas propostas praticamente convergentes de Suzor e Kaye em favor de uma incorporação dos parâmetros de direitos humanos pelas plataformas são pertinentes, mas devem ser exploradas e qualificadas. Vale mencionar que, embora seus argumentos deem uma ênfase central a decisões sobre moderação de conteúdo, suas considerações por vezes abarcam sugestões muito mais abrangentes – como a adoção pelas plataformas de mecanismos internos para avaliação de impactos de suas políticas sobre os direitos humanos nos mercados em que atuam[265]. A análise a seguir, contudo, mantém o enfoque no objeto de pesquisa deste trabalho. E, a respeito das decisões de derrubada de postagens, *a ideia central dos argumentos de ambos – e que deve ser mantida – é a incorporação de uma lógica de direitos, pela qual as restrições a direitos fundamentais devem ser motivadas*.

Incorporar a gramática dos direitos significa reconhecer que a regulação de discursos em redes sociais não se esgota no simples exercício de uma liberdade empresarial, pois essa atividade impacta de modo relevante os direitos fundamentais, notadamente a liberdade de expressão. Ao articular problemas e decisões em torno da lógica de direitos, *a decorrente necessidade de motivação das regras estabelecidas e das decisões construídas a partir delas permite não apenas um maior controle social e intersubjetivo, mas também viabiliza uma responsabilização jurídica pelas decisões tomadas – esta última, especialmente, pelo judiciário*[266]. Torna-se possível debater alternativas decisórias (mais ou menos restritivas) e cobrar posturas sólidas, coerentes e isonômicas.

No mais, os parâmetros do direito internacional dos direitos humanos não fazem sumir os dilemas inerentes aos grandes sistemas de moderação. Primeiro, porque entre os diversos órgãos internacionais há divergências interpretativas sobre a aplicação de

---

[264] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 118-121.

[265] Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, p. 132 e seguintes.

[266] Esse aspecto será desenvolvido em tópico a seguir, Capítulo 4-E.

direitos envolvidos. A própria natureza interpretativa dos argumentos de direitos fundamentais, especialmente em casos difíceis, leva a divergências razoáveis em torno de suas aplicações concretas – como bem exemplifica a "margem de apreciação" que a Corte Europeia de Direitos Humanos reconhece aos países sob sua jurisdição[267]. São parâmetros, ademais, voltados a estados – que não podem ser simplesmente transferidos às plataformas, como entes privados que são.

Por fim, há uma diferença importante entre argumentar por uma vinculação jurídica cogente das plataformas a parâmetros do direito internacional dos direitos humanos e argumentar a favor de que elas adotem critérios, a partir de sua própria autonomia, que correspondam a parâmetros daquele sistema. No plano transnacional, tratados internacionais são oponíveis a estados, bem como decisões de cortes dos sistemas de direitos humanos são proferidas diante das obrigações por eles assumidas. Não há mecanismos cogentes de vinculação de empresas a esses parâmetros[268]. Em sentido contrário, uma capacidade concreta de vinculação dessas empresas a parâmetros de direitos fundamentais encontra-se presente nos ordenamentos jurídicos nacionais[269].

Permanece então, uma *espécie de paradoxo*: parâmetros substantivos de direitos fundamentais (sejam eles provenientes do direito internacional dos direitos humanos, como defendem Suzor e Kaye) podem ser incorporados pelas plataformas dentro de sua esfera de autonomia normativa no plano transnacional, por uma espécie de exercício de autolimitação, porque nesse plano não há mecanismos cogentes de vinculação jurídica. Esses mecanismos vinculantes, porém, estão presentes nos âmbitos das ordens nacionais,

---

[267] David Kaye, como relator especial da ONU sobre liberdade de opinião e expressão, explicita *divergências interpretativas sobre problemas que envolvem a liberdade de expressão entre o sistema internacional e o sistema regional europeu de direitos humanos,* por exemplo. Enquanto o Comitê de Direitos Humanos da ONU se posiciona contrário a leis que criminalizam blasfêmias ou a negação de genocídios, a tendência inversa é observada na Corte Europeia de Direitos Humanos, que tem sido deferente com criminalizações promovidas nesse sentido por estados-membros. Entre essas esferas, os parâmetros interpretativos do sistema internacional tendem, assim, a ser mais protetivos à liberdade de expressão, de um modo geral. A própria ideia de "margem de apreciação" é rejeitada pelo Comitê de Direitos Humanos da ONU. Ver: *Relatório Especial da ONU sobre a promoção e proteção da liberdade de opinião e expressão A/74/486*, publicado em 09/10/19.

[268] Tanto Suzor quanto Kaye contextualizam suas propostas dentro do movimento crescente e mais amplo que busca direcionar a eficácia de direitos humanos a empresas multinacionais com alto impacto em sistemas de governança global – iniciativa exemplificada pela "UN Guiding Principles on Business and Human Rights", instrumento de "soft law" voltado àqueles entes privados.

[269] Que, por sua vez, também devem levar em conta os parâmetros internacionais de direitos humanos aos quais estão vinculados. A respeito da possibilidade de divergências interpretativas sobre direitos fundamentais entre ordens nacionais e transnacionais, ver: Virgílio Afonso da Silva, "Colisões de direitos fundamentais entre ordem nacional e ordem transnacional", *in:* Marcelo Neves (org.), *Transnacionalidade do direito: novas perspectivas dos conflitos entre lógicas jurídicas*, Quartier Latin, 2010.

especialmente por decisões legislativas e judiciais – mas que são limitados a seus próprios territórios. Essa é uma observação importante, pois deve, através de um olhar cosmopolita, *orientar as propostas que conjugam iniciativas nos planos transnacional e nacional para uma implementação do ideário do constitucionalismo digital*.

4.B – Constitucionalismo digital, moderação de conteúdo e a perspectiva transnacional do direito das plataformas

Não sem motivo, porque o "direito das plataformas" das redes sociais possui um grau de autonomia para a definição de suas próprias regras com relação aos ordenamentos jurídicos nacionais, tem sido ele também um campo singular no qual são feitas demandas em torno de *argumentos de direitos* e de reclamos em torno da noção de um *devido processo digital*, as duas linhas que caracterizam o *marco do constitucionalismo digital*.

A governança privada de discursos pelas redes sociais já deixou de ser socialmente compreendida como um simples fruto de decisões empresariais internas e sujeita-se cada vez mais a escrutínio público e demandas por *direitos digitais*[270], que problematizam as escolhas ou restrições existentes. Como a moderação de conteúdo não coincide com critérios de licitude ou ilicitude, as disputas em torno dessas regras sob a lógica do permitido/proibido miram as próprias condições de discursos e de visibilidade no meio digital – a definição das *normas sociais*[271] prevalecentes naquele ambiente.

A questão da nudez pode ser exemplificativa dessa dinâmica. Desde 2010, uma crescente mobilização de usuárias do Facebook começou a problematizar – dentro da plataforma, mas também em manifestações públicas que atingiram a mídia em geral – a restrição da empresa para a exibição de seios femininos, quando aplicada em fotos que retratavam a amamentação de bebês. Grupos contra a "censura" a fotos de amamentação foram criados no próprio Facebook, com centenas de milhares de integrantes. Petições também foram assinadas por dezenas de milhares de pessoas, sob o argumento de que a política era sexista, equiparava o ato de amamentação a um ato sexual e impedia mulheres

[270] Do ponto de vista de atores que articulam essas demandas, pode ser muito mais efetivo direcionar propostas de reformas a uma plataforma global, em comparação a múltiplos governos nacionais ou órgãos internacionais – que além de serem potencialmente menos responsivos a pequenos grupos organizados, encaram um "gap*"* de capacidade técnica de regulação. Uma coalização de entidades do terceiro setor pode ter um impacto mais certeiro sobre o Facebook do que sobre o processo legislativo nacional de um ou diversos países.

[271] Lawrence Lessig, *Code: version 2.0*, Basic Books, 2006; ver Capítulo 2.

de se manifestarem e trocarem suas experiências pessoais sobre o tema. Lideranças da campanha postavam imagens proibidas para denunciar a consequente suspensão temporária de seus perfis. A pauta reverberava em veículos de imprensa, uma entre várias outras que também apontavam desafios e problemas que recaiam sobre as mulheres no ambiente digital, como assédios, piadas sobre estupros, entre outros. Ainda em uma época em que a empresa era mais fechada a demandas de seu público e menos transparente com relação a suas regras, o Facebook manteve a proibição – até 2014, quando finalmente cedeu para criar exceções que permitem imagens de amamentação e também de cicatrizes de cirurgias de mastectomia, divulgadas como forma de conscientização para prevenção de câncer e apoio mútuo entre pacientes[272].

Atualmente, o Instagram é o alvo preferencial da campanha #Liberteomamilo ("#Freethenipple") que denuncia um sexismo inerente a essa proibição, considerando que é permitida a exibição de mamilos masculinos. A plataforma se defende dizendo que precisa se adequar a padrões sociais e culturais dominantes, inclusive sob uma perspectiva global[273].

Demandas do gênero são articuladas em torno de argumentos substantivos de direitos fundamentais e problematizam a razoabilidade das restrições impostas à liberdade de expressão de usuários. Levam em conta as especificidades do ambiente digital e geram uma pressão externa que, no limite, pode levar empresas a motivarem publicamente as razões de suas regras ou eventualmente revisa-las. Além disso, como aponta Gillespie, as regras de moderação de conteúdo de redes sociais também constituem uma arena de disputas políticas:

> *"na medida em que a expressão pública tem cada vez mais se transferido para essas plataformas, e a atenção ao fato de que elas moderam conteúdo tem aumentado, contestar essas plataformas, suas regras e seu direito de intervir tem se tornado parte da estratégia política sobre os assuntos em si"*[274].

---

[272] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, pp. 141-172. Recentemente, a empresa também flexibilizou suas regras de proibição a nudez quando no contexto de protestos e manifestações públicas (Capítulo 3-C).

[273] Will Instagram ever 'Free the Nipple'", reportagem publicada pelo jornal *The New York Times*, em 22/11/2019.

[274] Tarleton Gillespie, *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 170.

Como os parâmetros substantivos de direitos fundamentais estão sempre em aberto, pois não são definidos de antemão (vide tópico anterior), ganha força e importância a linha de consolidação de um devido processo digital junto às plataformas. Em 2018, entidades da sociedade civil e acadêmicos publicaram em conjunto os "Princípios de Santa Clara para transparência e prestação de contas em moderação de conteúdo"[275]. Esse documento propõe três eixos de propostas a serem incorporadas pelas políticas de moderação de conteúdo das redes sociais:

a) "Números": *como princípio geral, estatísticas sobre a remoção de postagens devem estar disponíveis ao público*. A manifestação pede ainda uma especificação mínima desse mecanismo de transparência, sugerindo a publicação do número de postagens reportadas ("flagged") e suas autorias (se governos, usuários comuns, entidades credenciadas, etc), de dados que identifiquem cada regra infringida, os formatos das postagens (audio, vídeo, texto, etc) e a localidade de suas autorias. Pede, também, que os respectivos relatórios sejam publicados trimestralmente;

b) "Notificação": *como princípio geral, usuários que tiveram suas postagens derrubadas ou contas suspensas devem ser notificados sobre os respectivos motivos*. O documento sugere que cada notificação deve conter a identificação precisa de URL, a cláusula de moderação de conteúdo supostamente infringida, como se chegou a essa tomada de decisão (se processo automático, ordem judicial, reclamação de outro usuário, pedido de governo, etc) e explicação do processo de apelação contra a decisão.

c) "Apelação": *como princípio geral, as empresas devem prover uma oportunidade efetiva de recurso em tempo razoável nos casos de derrubada de postagem ou suspensão de perfil*. Isso deveria incluir: revisão humana de pessoas não envolvidas na decisão inicial (especialmente quando esta foi feita por sistemas automatizados),

[275] No original, "Santa Clara Principles: on transparency and accountability in content moderation"; ver: www.santaclaraprinciples.org. Os signatários originais são: Electronic Frontier Foundation, Center for Democracy & Technology, ACLU Foundation of Northern California, New America´s Open Technology Institute e os pesquisadores acadêmicos Irina Raicu, Nicolas Suzor, Sarah T. Roberts, Sarah Myers West. Em sentido semelhante e baseado na primeira, há proposta preliminar e em construção por parte de uma coalizão de entidades da sociedade civil na América Latina: "Contribuições para uma regulação democrática das grandes plataformas que garanta a liberdade de expressão na internet", disponível em www.observacom.org, elaborada por Coletivo Intervozes, Desarrolo Digital, Instituto Brasileiro de Defesa do Consumidor (Idec) e Observatório Latinoamericano de Regulación Médios y Convergencias (Observacom).

oportunidade de apresentar informações adicionais para serem consideradas e notificação dos resultados do processo de revisão[276].

É possível observar uma correlação entre o amadurecimento de demandas da sociedade civil e especialistas no sentido de um devido processo digital e medidas concretas tomadas nesse mesmo sentido pelas grandes plataformas de redes sociais[277]. Por volta da mesma época – cerca de um mês antes da publicação dos "Princípios de Santa Clara" – o Facebook tornou públicas suas diretrizes e parte de seus critérios de moderação, até então restritos a seus ambientes internos, além de ampliar os processos de recursos contra suas decisões[278] – uma tendência também observada, de modo geral e desde então, no Twitter e Youtube. As práticas da indústria avançaram nesse sentido, ainda que nem todas as minúcias de padrões dos "Princípios de Santa Clara" tenham sido implementadas pelas plataformas.

No âmbito transnacional do "direito das plataformas", os mecanismos de devido processo digital têm sido, sobretudo, resultados de iniciativas de autolimitação, transparência e prestação de contas pelas empresas, que buscam aumentar a legitimidade de suas decisões junto ao público. A mais ambiciosa dessas iniciativas, porém, ainda está atualmente em fase de implementação pelo Facebook – e parece

[276] Diversos casos demonstram a necessidade de procedimentos efetivos de recursos. O Youtube permite exceções para materiais a princípio proibidos quando eles possuem valor "educacional, documental, científico ou artístico". Ainda assim, seus sistemas automatizados derrubavam muitos vídeos de protestos civis e episódios de repressão feitos por cidadãos sírios contra o governo autoritário do país. Integrantes do governo reportavam os vídeos como "violentos" para fomentar suas derrubadas. Quando elas ocorriam, ativistas tinham dificuldades para recorrer e oferecer os contextos factuais sobre as imagens ao Youtube para que, segundo as regras dos próprios termos de uso, eles fossem mantidos no ar – David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 22-27. Uma situação análoga ocorreu no Twitter com relação a postagens sobre manifestações e protestos na Caxemira, derrubadas a pedidos do governo indiano, sem o fornecimento das razões específicas ou de direito de recurso dado aos usuários daquele território (pp. 31-33).

[277] A necessidade de processos de apelação ou de revisão em favor de usuários no âmbito das políticas de conteúdo das próprias plataformas, de relatórios de transparência sobre essas decisões, a publicação de regras prévias e claras, bem como o fato de que elas respeitassem os 'direitos humanos' em sua formulação, já constavam dos "Princípios de Manila para a responsabilidade de intermediários", editados em 2015 (ver Capítulo 4-C). O foco desse primeiro documento era majoritariamente voltado, contudo, à regulamentação legal da responsabilidade de intermediários – um vetor independente e complementar de governança do ambiente digital.

[278] "Nós decidimos publicar essas diretrizes internas por duas razões. Primeiro, as diretrizes irão ajudar as pessoas a entenderem onde nós traçamos as linhas em questões nuançadas. Em segundo lugar, fornecer esses detalhes torna mais fácil para que todas as pessoas, incluindo especialistas em diferentes áreas, possam dar para nós um retorno sobre o que podemos melhorar nas regras – e também nas decisões que tomamos – ao longo do tempo" – "Publishing our internal enforcement guidelines and expanding our appeals process", *Facebook Newsroom*, publicado em 24/04/2018.

aprofundar esse processo de constitucionalização das grandes plataformas digitais, exatamente na seara de moderação de conteúdo.

O vasto poder de regulação de discursos no ambiente digital que o Facebook possui não corresponde apenas à estrutura de sua pessoa jurídica: como fundador, CEO e acionista majoritário, Mark Zuckerberg concentra em si a capacidade (e a prática) de tomar todas as decisões finais sobre as políticas de sua empresa. Isso torna ainda mais extraordinária sua decisão, anunciada inicialmente em 2018, de criar um *órgão independente, formado por pessoas externas à empresa, para tomar decisões finais, transparentes e vinculantes no campo da moderação de conteúdo*. O projeto chegou a ser chamado preliminarmente de "Suprema Corte do Facebook"[279]; ao final, foi batizado como "Conselho Supervisor" ("Oversight Board"). O fato de esse órgão ainda não ter sido instalado em definitivo – a previsão é de que comece a julgar casos em meados de 2020 – impede uma análise exaustiva e muito aprofundada sobre essa iniciativa; no entanto, as informações já divulgadas permitem uma apresentação consistente com os objetivos deste tópico e capítulo.

A ideia de estabelecer uma "corte independente" no sistema de governança privada de conteúdo do Facebook remete imediatamente a um sistema de separação de poderes, no qual um órgão julgador supervisiona as demais funções, como a "legislativa" (que cria as regras) e a "executiva" (que as implementa). "Essa é uma estrutura de governança sem precedentes para uma empresa privada. Mas é, claro, a forma dominante de governança em estados-nações"[280].

Os casos enviados ao Conselho poderão ser formulados *pelos usuários* (a partir de procedimentos de recursos, por formulários digitais na plataforma) *ou pelo próprio Facebook*. O órgão terá ampla liberdade para decidir *quantos e quais casos serão analisados*, já que a empresa recebe cerca de um milhão de denúncias de violações de termos de uso por dia[281]– as escolhas serão feitas por um comitê de seleção, composto de

[279] "Zuckerberg sugere Suprema Corte para julgar conteúdo no Facebook", reportagem publicada pelo jornal *Folha de S. Paulo* em 02/04/18; Thomas Kadri e Kate Klonick, "How to make Facebook's Supreme Court work", artigo publicado no jornal The *New York Times* em 17/11/2018.

[280] "Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, p. 16. Douek avalia que esse é um dos "projetos constitucionais mais ambiciosos da era moderna".

[281] "Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, p. 11. Como aponta a autora, a liberdade

modo rotativo por uma parcela dos integrantes do Conselho. O objetivo é de que se dedique aos casos mais relevantes e paradigmáticos; suas decisões também irão valer para a plataforma Instagram. O regime interno já publicado prevê prazos de julgamento, que podem ser abreviados quando houver uma espécie de regime de urgência a pedido da empresa. Para garantir a autonomia do Conselho, o Facebook montou um fundo independente dotado de cerca de 130 milhões de dólares que irá financiar suas atividades pelos próximos seis anos, incluindo uma equipe própria de funcionários[282].

As decisões do Conselho serão vinculantes ao respectivo caso concreto. A competência do órgão abarca diversas áreas de moderação de conteúdo, incluindo políticas da empresa para anúncios ou mesmo conteúdos rotulados como "falsos" por parceiros que realizam checagem de fatos ("fact checking"). Como não poderia deixar de ser, não serão acatadas decisões apenas nos casos ou situações em que isso implicaria a violação pelo Facebook de alguma legislação.

Algumas dúvidas ainda permanecem sobre competência do órgão: seu regimento interno aponta que ela irá se consolidar "gradualmente", sujeita também às possibilidades técnicas do Facebook. A principal indagação por ora é se o Conselho poderá rever as decisões que determinam a diminuição da visibilidade de postagens, medida que pode ser tomada como parte da política de moderação de conteúdo, com menor grau de intervenção em comparação a uma simples derrubada (Capítulo 2-F)[283]. Além disso, está claro que o Conselho pode reverter as decisões de derrubada – mas permanece incerto se poderá rever decisões que *decidem manter no ar* um determinado conteúdo[284].

---

de escolha de casos é essencial para o Conselho, caso contrário, seria fácil manipular sua pauta pelo excesso de casos submetidos a análise.

282 "Why Mark Zuckerberg´s Oversight Board may kill his political ad policy", reportagem publicada por *Wired*, em 28/01/2020.

283 Douek comenta como é importante que essas decisões também sejam passíveis de revisão pelo Conselho, pois caso contrário o Facebook poderia toma-las, ao invés de simples derrubadas, exatamente para evitar essa supervisão externa. Nesse sentido, as atribuições do órgão deveriam estar conectadas a "algum grau de transparência algorítmica" – Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, pp. 42-43; "Facebook´s Oversight Board Bylaws: for once, moving slowly", artigo publicado em *Lawfare Blog*, em 28/01/20.

284 "'Derrubadas' podem parecer o tipo de decisão que mais ameaçam a liberdade de expressão. Mas uma parte significativa das decisões de moderação de conteúdo controversas feitas pelo Facebook nos últimos anos têm sido decisões para deixar no ar um conteúdo, e não derrubá-lo". Ela menciona os casos de discursos de ódio em Myanmar ou de conteúdos extremistas e de teorias conspiratórias, como o do americano Alex Jones – Evelyn Douek, "Facebook´s Oversight Board Bylaws: for once, moving slowly", artigo publicado em *Lawfare Blog*, em 28/01/20.

Serão 40 conselheiros ao total: o Facebook escolheu as quatro pessoas iniciais, responsáveis a partir de então pela seleção de seus novos colegas em conjunto com a empresa. Além deles, foi apontado o primeiro diretor do Conselho, Thomas Hughes – que atuava como diretor da Article 19, entidade voltada a direitos humanos e à liberdade de expressão. Poucos meses depois, foram anunciados 16 integrantes adicionais, entre eles o advogado e professor brasileiro especialista em direito digital Ronaldo Lemos. Cada membro serve no máximo três mandatos de três anos – e os conselheiros devem se reunir pessoalmente pelo menos uma vez ao ano; na maior parte do tempo, os julgamentos serão virtuais[285].

Os casos serão julgados inicialmente por painéis compostos por cinco integrantes, que recebem informações da empresa sobre as decisões previamente tomadas e que podem também solicitar opiniões externas de especialistas sobre os assuntos abordados. Esses painéis serão designados aleatoriamente entre os integrantes do Conselho, mantendo pelo menos uma pessoa que seja da região do mundo ligada diretamente ao conteúdo. A identidade dos julgadores será preservada para protegê-los de eventuais pressões. Todos os integrantes do painel devem votar; não é possível se abster. O painel é responsável pela minuta de decisão, que é submetida ao plenário do Conselho Supervisor para uma revisão final. O órgão pode fazer recomendações antes da publicação da versão final, que devem ser analisadas pelo painel responsável. O plenário também pode, por maioria de votos, "decidir se a minuta de decisão será adotada ou se é necessária uma nova revisão" – nesta última hipótese, o plenário envia o caso para um novo painel[286].

---

[285] "Human Rights expert to keep Zuckerberg in check", reportagem publicada pela *BBC News*, em 28/01/2020; "Announcing the first members of the Oversight Board", publicação por *Oversightboard.com*, em 06/05/2020. O Conselho destacou uma representatividade global desse grupo inicial de dez homens e dez mulheres, com pessoas provenientes de todos os continentes, a maioria delas fluentes em mais de uma língua e com experiência de ter vivido em países estrangeiros – 25% vêm dos Estados Unidos e Canadá, 20% da Europa, 15% do Sudeste Asiático e Oceania e 10%, respectivamente, da América Latina e Caribe, do Oriente Médio e Norte da África, do Sul da África e da Ásia Sul e Central. Há de fato uma distribuição predominante entre pessoas com histórico de atuação acadêmica e em entidades da sociedade civil, com destaque para experiência em direitos humanos. Três conselheiros são juízes aposentados de cortes nacionais ou internacionais. Dois integrantes possuem conhecimento em linguagem de programação. Seis atuam ou atuaram como jornalistas, entre eles um ex-editor-chefe do jornal britânico The Guardian, Alan Rusbridger. Nicolas Suzor, professor de direito australiano citado algumas vezes neste trabalho, também integra o colegiado.

[286] Regimento Interno do Conselho Supervisor; "Preparing the way forward for Facebook´s Oversight Board", publicação pelo *Facebook Newsroom*, em 28/01/20.

O processo que levou a esse desenho institucional envolveu uma ampla consulta pública em nível global, documentada pela empresa. Ao total, foram ouvidas cerca de 2 mil pessoas em 88 países[287]. Durante ele, o principal ponto de sugestão acatado foi a *atribuição do Conselho Supervisor para sugerir mudanças das regras da política de conteúdo*. Como suas decisões são vinculantes apenas ao caso concreto, quando for feita uma recomendação de alteração de regra geral para a empresa, *o Facebook fica obrigado a se manifestar sobre ela – acatando-a ou fornecendo as razões para uma rejeição*[288]. "Isso é precisamente o que o sistema é desenhado para fazer – colocar uma fortíssima pressão sobre nós para que mantenhamos uma política apenas se estivermos absolutamente certos de que isso é a coisa certa a fazer", explicou Nick Clegg, diretor global de política da empresa[289].

Em setembro de 2019, o Facebook havia estabelecido de modo mais claro os *valores que devem guiar não apenas as atividades direta da empresa, mas também as decisões de seu Conselho Supervisor*. O primeiro e mais abrangente deles é a "voz" – dar voz às pessoas, o que corresponde a uma liberdade de expressão (publicação). Esse direito, porém, pode ser restringido quando em conflito com demais valores: autenticidade, segurança, privacidade e dignidade[290]. Espera-se que o Conselho Superior fundamente suas decisões em torno da ponderação desses valores a partir das regras estabelecidas e dos casos concretos[291].

Caso seja bem-sucedida, essa dinâmica irá emular a hermenêutica de adjudicação de direitos fundamentais. Se em um estágio anterior as normas do Facebook

---

287 "Global feedback and input on the Facebook oversight board for content decisions", publicado pelo *Facebook Newsroom*, em 27/06/2019.

288 Douek ressalta a importância dessa regra e a compara a uma espécie de "revisão judicial fraca", voltada à imposição de um ônus argumentativo, ao invés de uma instância deliberativa "judicial" responsável por uma última palavra – Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, p. 54 e seguintes.

289 "Why Mark Zuckerberg´s Oversight Board may kill his political ad policy", reportagem publicada por *Wired*, em 28/01/2020. Para a publicação, o Conselho Supervisor parece ter sido feito sob medida para que sejam revistas, por essa instância "externa", as regras do Facebook para propagandas políticas.

290 "Updating the values that inform our community standards", publicação pelo *Facebook Newsroom*, em 12/09/20.

291 Conforme prevê a "Carta" do Conselho Supervisor, que dispõe também quais critérios devem ser analisados para a seleção de casos, incluindo aspectos como severidade ("o conteúdo alcança ou afeta a voz, privacidade ou dignidade de alguém"), discurso público ("o conteúdo levanta significativo debate público e/ou discursos sociais e políticos importantes") e dificuldade ("há desacordo sobre a decisão do Facebook sobre o conteúdo e/ou sobre as regras ou políticas aplicáveis") – "Establishing Structure and Governance for an independent oversight board", publicação do *Facebook Newsroom*, em 17/09/19.

evoluíram de "standards*"* para regras (Capítulo 3-A), a implementação de seu Conselho Supervisor parece colocar claramente sua política de moderação de conteúdo no bojo de um *sistema de regras e princípios*, no qual as colisões entre direitos e valores subjacentes são reconhecidas como inevitáveis e os conflitos são resolvidos por meio de ponderações e apresentações de argumentos[292].

Para Douek, o Conselho não servirá para corrigir todas as decisões individuais de moderação de conteúdo e nem, por outro lado, para definir regras globais de discursos. Seu "verdadeiro valor" decorre de duas "mais modestas" funções:

> "*Primeiro, pode ajudar a iluminar defeitos na formulação de políticas pelo Facebook, removendo bloqueios (como pontos-cego ou inércia) no 'processo legislativo' que leva à formulação dos 'padrões de comunidade'. Em segundo lugar, ao possibilitar um fórum independente para discussões de decisões polêmicas sobre moderação de conteúdo, o Conselho Supervisor pode ser um importante fórum para o processo de razão pública necessário para que pessoas em uma comunidade plural aceitem as regras que as governam, mesmo que elas discordem da substância dessas regras*"[293].

Sua principal virtude, portanto, ainda a ser testada e comprovada, seria a garantia de um foro público para apresentações de razões e argumentos que sustentem publicamente uma determinada decisão, aumentando a possibilidade de controle intersubjetivo sobre as regras de moderação de conteúdo do Facebook – e, por consequência, sua legitimidade.

---

[292] "Esses valores vão cumprir um papel importante no trabalho do Conselho Supervisor, mas há maneiras cruciais pelas quais eles não se comparam a uma constituição. O estabelecimento de valores pode desempenhar o mesmo papel que constituições ao expressar a visão fundamental do Facebook para sua plataforma, o que o Conselho pode usar para resolver as ambiguidades decorrentes de regras da comunidade (...). Mas porque esses valores são do Facebook e não da comunidade de usuários, eles não podem desempenhar o mesmo papel legitimador que constituições, porque não expressam a delegação de autoridades por parte de usuários ou um conjunto de regras amplamente aceitas" – "Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, p. 49.

[293] "Evelyn Douek, "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21 (2019)*, p. 7. Para a autora, há quatro principais incentivos que motivam o Facebook a adotar a iniciativa: "(1) envolver as decisões de moderação de conteúdo com uma aura de legitimidade, auxiliando na relação com seus usuários; (2) evitar ou guiar o rumo de regulações governamentais mais extensas; (3) terceirizar as decisões controvertidas para além da empresa; (4) facilitar uma melhor implementação das regras existentes" (p. 17).

Por ora restrito à maior e mais importante rede social, é possível que o Conselho Supervisor, se bem-sucedido, passe a abranger também outras plataformas globais que decidam aderir à sua "jurisdição". Ou que demais plataformas repliquem esse modelo, cada qual com seu conselho ou grupo de conselhos. De um modo geral, a ampliação de um modelo que garanta mais transparência e controle social sobre as decisões de moderação de conteúdo pelas demais grandes plataformas globais também seria especialmente relevante porque elas possuem iniciativas conjuntas e centralizadoras para a definição de conteúdos a serem derrubados, como nos casos de propagandas de terrorismo (Capítulo 2-A) e, mais recentemente, campanhas de desinformação – o que Douek define como "um cartel de conteúdos" ("content cartel")[294]. Ao mesmo tempo, há em paralelo outras propostas para a instituição de órgãos de supervisão ("accountability") formado por empresas, entidades e especialistas que buscam, em um nível global ou regional, criar instâncias regulatórias sobre moderação de conteúdo que estejam localizadas entre a regulação estatal e a auto regulação por empresas[295].

Sem dúvida, trata-se de uma experimentação institucional ousada. Os sucessos e as falhas dessa empreitada ainda estão para serem vistos e debatidos. As próprias independência e efetividade do Conselho Superior estão postas para serem testadas. Ainda assim, inclusive diante da magnitude da empreitada, ela revela – especialmente no contexto deste tópico – como há um espaço abrangente para a construção de formas institucionais novas e criativas que contemplem o resguardo ou avanço de direitos fundamentais, em um plano normativo transnacional (e privado), a partir de uma perspectiva do *constitucionalismo digital*.

---

[294] Evelyn Douek, "The rise of content cartels", artigo publicado por *Knight First Amendment Institute – Columbia University*, em 11/02/20.

[295] Como a proposta da entidade Article 19 para a instituição de um "Conselho de Redes Sociais" ("Social Media Council"), "um modelo de mecanismo de supervisão por diversos atores que garanta que isso seja feito de modo aberto, transparente, independente ao abordar problemas de moderação de conteúdo nas plataformas de redes sociais, com base nos padrões internacionais de direitos humanos". Para debates em torno dessa proposta, bem como referências a outras ideias aproximadas, ver: Pierre François Docquir, "The Social Media Council: bringing human rights standards to content moderation on social media", artigo publicado por *Centre of International Governance Innovation*, em 28/10/2019. Para uma proposta mais tradicional de órgão regulador estatal dotado de independência e autonomia, ver: "Contribuições para uma regulação democrática das grandes plataformas que garanta a liberdade de expressão na internet", disponível em www.observacom.org. Jonathan Zittrain, por sua vez, chega a propor que as decisões a respeito de conteúdos sob disputa sejam feitas por júris aleatórios de bibliotecários, professores de ensino médio ou estudantes, evocando as supostas propriedades socialmente legitimadoras de júris – "A jury of random people can do wonders for Facebook", artigo publicado em *The Atlantic*, em 14/11/19.

Esse olhar global e cosmopolita do constitucionalismo digital, contudo, não deve abandonar uma perspectiva para o direito nacional. Pelo contrário, deve orientá-la. A necessária construção de mecanismos que implementem um ideário constitucionalista no ambiente digital transnacional pode (e deve) ser reforçada e complementada por iniciativas do direito estatal. Por isso, os tópicos a seguir se voltam à propositura de parâmetros normativos no âmbito do direito brasileiro, a partir dos dispositivos em vigor do Marco Civil da Internet.

4.C – Contextualizando a moderação de conteúdo das redes sociais no direito brasileiro a partir do Marco Civil da Internet e suas regras de responsabilização civil de intermediários

O problema jurídico da derrubada de conteúdos por decisões das grandes redes sociais adquire contornos próprios diante do direito brasileiro, especialmente em função das regras previstas pelo Marco Civil da Internet (lei federal nº 12.965/14). A lei, considerada um estatuto de "princípios, direitos, deveres e responsabilidades" ligados à governança da internet[296], dedica sua importante Seção III (artigos 18 a 21) a *regras que criam incentivos para evitar a derrubada de postagens na rede mundial.*

Isso é feito a partir de um regime de *quase isenção de responsabilização civil a intermediários digitais*. Com os objetivos declarados na própria lei de "assegurar a liberdade de expressão e impedir a censura", determina-se que a responsabilidade civil de provedores de aplicações em razão de conteúdos gerados por terceiros ocorrerá "*somente após ordem judicial específica*" de remoção de conteúdo (artigo 19)[297]. Esse regime de responsabilidade civil *favorável a intermediários* traça uma linha que separa "o autor da

[296] Para um histórico sobre o cenário legal brasileiro pré-Marco Civil da Internet, um retrato sobre o amplo processo de consulta e deliberação que levou à sua formulação, os principais aspectos de seus dispositivos legais e sua aplicação em casos judiciais, ver: Luiz Fernando Marrey Moncau e Diego Werneck Arguelles, "The Marco Civil da Internet and Digital Constitutionalism": *in* Giacarlo Frosio (ed.), *The Oxford Handbook of Online Intermediary Liability*, Oxford University Press, 2020; Carlos Affonso Souza e Ronaldo Lemos, *Marco Civil da Internet: construção e aplicação*, Editar Editora, 2016. Para uma pesquisa mais detida sobre o singular processo de consulta pública, tramitação legislativa e principais políticas definidas pela lei: Francisco Carvalho de Brito Cruz, *Direito, democracia e cultura digital: a experiência de elaboração legislativa do Marco Civil da Internet*, dissertação de mestrado apresentada à Faculdade de Direito da Universidade de São Paulo, 2015.

[297] Art. 19, *caput*: "Com o intuito de assegurar a liberdade de expressão e impedir a censura, 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, no âmbito e nos limites técnicos do seu serviço e dentro do prazo assinalado, tornar indisponível o conteúdo apontado como infringente, ressalvadas as disposições legais em contrário".

mensagem e seu mensageiro", partindo da premissa de que plataformas não costumam moldar, participar da elaboração ou monitorar previamente os conteúdos nela publicados. Diferenciam-se, nesse aspecto, de editores tradicionais – que se caracterizam exatamente por decidirem aquilo que será publicado, participando da produção do conteúdo ou anuindo previamente com seu teor.

A legislação é direcionada a qualquer tipo de intermediário digital – o que inclui também páginas de mecanismos de buscas, portais de notícias ou mesmo de compras que permitem comentários e resenhas por seus usuários. Este trabalho, contudo, foca em seus impactos para as grandes redes sociais – o que pode motivar a modulação de uma atualização legislativa para lidar de maneira especializada com essa espécie de atores, conforme argumentos a seguir.

O modelo brasileiro segue, em linhas gerais, as diretrizes que se consolidaram como propostas de entidades internacionais que atuam pela proteção de direitos digitais na democracia (especialmente, a liberdade de expressão), condensadas nos chamados "Princípios de Manila para a responsabilidade de intermediários"[298]. Esse escudo jurídico tem a virtude de conter a ameaça de "censura colateral"[299], pela qual o direito estatal (por suas leis ou decisões judiciais) pode ser usado para suprimir discursos por meio da responsabilização jurídica de "mensageiros". Ainda sob esse raciocínio, se intermediários tiverem que *necessariamente* avaliar a licitude ou ilicitude de postagens, eles teriam um forte incentivo para derrubar conteúdos (mesmo que eventualmente lícitos) que lhes possam acarretar ônus em suas operações (como eventuais custosas discussões em processos judiciais), bem como para criar amplos sistemas de censura prévia ou de amplo

[298] O primeiro deles sendo: "os intermediários devem ser protegidos por lei da responsabilização por conteúdos produzidos por terceiros", com as seguintes recomendações mais específicas: "quaisquer regras que disponham sobre a responsabilidade dos intermediários devem ser previstas em leis que sejam precisas, claras e acessíveis"; além disso, eles "devem ser imunes de responsabilização por conteúdos de terceiros sempre que não tenham realizado quaisquer modificações", "não devem ser responsabilização por não restringir conteúdos legais" e "nunca devem ser estritamente responsabilizados por hospedar conteúdos ilegais de terceiros, nem devem ser obrigados a monitorar conteúdos de maneira proativa como parte de um regime de responsabilidade de intermediários". Os princípios foram declarados como um mapa de boas práticas em 2015 por uma coalização de entidades internacionais do terceiro setor que se ocupam de direitos digitais na democracia, como Article 19, Eletronic Frontier Foundations, Derechos Digitales, The Centre for Internet & Society, entre outros – www.manilaprinciples.org

[299] Vide argumentos de Jack Balkin em Capítulo 1.B.

monitoramento. Daí que, ainda de acordo com os "Princípios de Manila", a *obrigatoriedade* de ordens de remoção de conteúdo deve decorrer de decisões judiciais[300].

A constitucionalidade desse modelo de responsabilidade civil previsto pelo artigo 19 do Marco Civil da Internet está em pauta para julgamento pelo Supremo Tribunal Federal, a partir de dois recursos extraordinários com reconhecimento de repercussão geral[301]. A princípio, o modelo alternativo – e que era vigente antes da atual legislação – era exatamente o de *responsabilidade civil a partir do aviso ou notificação sobre o conteúdo potencialmente ilícito* pela parte prejudicada, tal como então sedimentado pela jurisprudência do Superior Tribunal de Justiça[302]. Nesse caso, para evitar a responsabilização, intermediários devem agir por meio de um procedimento de "notificação e retirada" ("notice and takedown")[303]. Em manifestação feita ao STF a título de "amicus curiae*"* em um dos recursos extraordinários, o centro de pesquisas Internetlab apresentou relatório no qual classificou o sistema de responsabilização por mera notificação (em regra extrajudicial e feita por interessados, por vezes a cargo de órgãos

[300] Para Nicolas Suzor, "intermediários não estão bem posicionados para tomarem decisões sobre se um material ou conduta é legal ou não. Se eles devem tomar uma decisão sobre se algo é legal, eles não possuem as proteções de independência que cortes têm para que o público possa confiar que suas decisões são legítimas. Se é difícil resolver quando um reclamo é valido ou não com relação a um problema que afete o direito de alguém, esse é o estágio que devemos envolver o sistema judicial" - *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 160-161.

[301] Supremo Tribunal Federal, recursos extraordinários nº 1.057.258/MG e nº 1.037/396/SP. No segundo deles, decisão inferior do judiciário paulista que ascendeu ao STF baseou-se na ideia de que o regime do art. 19 do Marco Civil da Internet fere as garantias constitucionais de proteção a consumidores, a direitos de personalidade e de acesso ao judiciário. O primeiro caso é anterior à vigência daquela legislação.

[302] Ver, por exemplo: Recurso Especial nº 1.193.764/SP, julgado em 2011.

[303] Para Luiz Fernando Marrey Moncau, "cada modelo possui vantagens e desvantagens. O modelo de notificação e retirada facilita a remoção de conteúdo da rede, na medida em que não exige a movimentação do aparato burocrático do Estado e depende única e exclusivamente da movimentação da parte lesada. Entretanto, tal modelo possui problemas, especialmente nos casos em que os conteúdos apontados como infringentes não são facilmente reconhecidos como ilícitos (casos em que a crítica se confunde com a difamação, ou em que a violação do direito de autor está na linha tênue dos usos permitidos pela lei autoral). Nessas situações, aponta-se que o intermediário de internet não está na melhor posição para analisar se o conteúdo é ou não ilícito, seja porque este papel deve ser reservado aos tribunais ou porque o risco de ser responsabilizado pode levar o intermediário a remover o conteúdo para evitar uma possível obrigação de indenizar. A alternativa da ordem judicial, por outro lado, aloca no Poder Judiciário a responsabilidade por definir se determinado conteúdo é ou não ilegal. Apesar de mais lenta (devido a necessidade de manifestação de um juiz no curso de um processo), evita o problema da adjudicação de direitos por uma entidade privada que tem, entre seus interesses, aquele de evitar condenações e indenizações. Tendo em vista o risco de remoção de conteúdos perfeitamente lícitos, o modelo de retirada após ordem judicial foi apontado por diversas organizações da sociedade civil como o mais adequado à proteção da liberdade de expressão" – *Direito ao esquecimento: entre a liberdade de expressão, a privacidade e a proteção de dados pessoais*, Editora RT, 2020, p. 295.

governamentais) como sendo típico de regimes autoritários que buscam controlar o fluxo de informações em suas sociedades[304].

Esse regime geral do Marco Civil da Internet comporta exceções: seu artigo 21 prevê que a responsabilidade civil da plataforma será subsidiária pela não remoção, depois de simples notificação, em casos de violação da intimidade por materiais contendo imagens não autorizadas de nudez, atos sexuais privados e afins. Para casos do tipo, a responsabilidade civil decorre de mero aviso, que gera o dever jurídico de a empresa verificar a postagem e realizar a derrubada do conteúdo – uma exceção para a qual vige, portanto, o procedimento de notificação e retirada[305].

Outra exceção àquela regra geral envolve casos que tratam de direitos autorais (art. 19, §2º). Nesses casos, a lei determina que as regras do artigo 19 só serão aplicáveis após a promulgação de lei específica, o que até hoje nunca ocorreu. Por isso, nessa seara, a princípio vige o regime legal anterior, de responsabilização civil a partir de notificação. No entanto, a falta de uma legislação específica não impede que plataformas realizem, mesmo no Brasil, um controle autônomo ainda mais rigoroso do que o procedimento de notificação e retirada para materiais supostamente protegidos por direitos autorais, implementando varreduras automatizadas que levam em conta as determinações e incentivos de todo seu mercado global, *como ilustra o caso do Youtube e seu sistema de Content ID, apresentados no Capítulo 2-A.*

Neste ponto em especial, novamente transparece como pode haver uma *dinâmica potencialmente competitiva entre as regras legais de um determinado país e aquelas de plataformas transnacionais.* Essa tensão na área dos direitos autorais já chegou ao judiciário brasileiro: em 2018, o tribunal de justiça de Santa Catarina, por decisão unânime, impôs uma condenação cível ao Google depois de considerar que a exclusão de um vídeo satírico do Youtube, após a reclamação ("notice*"*) de titulares de direitos autorais da obra original, foi ilegal, porque feria a proteção dada às sátiras pela legislação autoral brasileira[306].

---

[304] A manifestação traz os exemplos da China, Venezuela, Irã, Rússia e Ruanda – parecer do centro de pesquisa Internetlab nos autos do recurso extraordinário nº 1.037.396/SP.

[305] De maneira análoga, o artigo 241-A do Estatuto da Criança e do Adolescente prevê a responsabilização criminal e administrativa para quem possibilita o acesso a imagens de pornografia infantil por acesso à "rede de computadores" apenas se, após "notificação oficial", deixa de retirar o conteúdo do ar

[306] O caso envolvia uma sátira à música "10%", de Maiara e Maraisa. Segundo consta nos autos, a autora do vídeo satírico negou inicialmente a proposta de repartir a monetização decorrente das visualizações com

Não há dúvidas de que o modelo legal de responsabilização civil de intermediários digitais que for adotado em um determinado país tem impactos para muito além de meros interesses patrimoniais privados, *sendo um importante vetor de regulação jurídica para a liberdade de expressão na internet e, por essa razão, gerador de consequências diretas para os exercícios de direitos fundamentais*[307].

Mas, ainda sob uma perspectiva do direito brasileiro, um enfoque exclusivo nas regras do Marco Civil da Internet, ainda que extremamente relevantes, oculta os aspectos da governança privada de discursos – que é normativa e também transnacional – feita pelas redes sociais, cuja amplitude foi apresentada nos dois capítulos anteriores. Embora o direito de cada país possa influenciar essa esfera normativa privada (e potencialmente condicioná-la em vários aspectos, em seu próprio território), as políticas de moderação de cada plataforma são de fato resultado da influência (com pesos variados) de diversos sistemas jurídicos, bem como de pressões derivadas de normas sociais e também de mercado[308], em suas arquiteturas e em seus modelos de negócios.

---

os detentores de direitos da música original. O próprio vídeo continha o termo "sátira" em seu título e uma referência, em seus primeiros segundos, sobre o dispositivo legal que protege esse tipo de uso. Quando avisada da derrubada iminente do vídeo, a autora da sátira enviou contra-argumentos pela manutenção do conteúdo – que foi derrubado mesmo assim. Para mais detalhes sobre o sistema de proteção de direitos autorais do Youtube, ver Capítulo 2-A. Para o desembargador relator, "o Youtube disponibiliza uma ferramenta denominada Content ID, o qual, de forma automatizada, mediante leitura audiovisual dos conteúdos postados, acusa supostas infringências aos direitos autorais de terceiros, que restam notificados pela plataforma para tomarem as providências cabíveis. Nessa toada, não se observa o suposto distanciamento por parte da segunda ré em relação às contendas administrativas de seus usuários envolvendo direitos autorais. Ao que parece, a plataforma age de forma a identificar conteúdos potencialmente violadores e efetivar sua remoção do ar, ainda que provisória. E, ao assim proceder, acaba assumindo o risco de que esses conteúdos sejam legítimos, tal como ocorre no caso em exame" – Tribunal de Justiça de Santa Catarina, apelação cível nº 0000447-46.2016.8.24.0175, 3ª Câmara de Direito Civil, votação unânime, julgamento em 06/02/18.

[307] Para um panorama sobre características de modelos regulatórios em todo o mundo, ver: Luiz Fernando Marrey Moncau, "Intermediários de Internet e Liberdade de Expressão: o mapa da busca de um delicado equilíbrio regulatório", artigo publicado no portal *Dissenso.org*, em 06/06/2018, com detalhes do projeto "World Intermediary Liability Map" do *Center for Internet and Society* da *Stanford Law School*.

[308] Como apontado no início do Capítulo 2, Lawrence Lessig destaca há anos que a governança da internet se sustenta sobre *quatro diferentes modais reguladores: o direito, o mercado, as normas sociais e o código – Code: version 2.0*, Basic Books, 2006. As crises de imagem recentes das redes sociais não ficaram restritas apenas à percepção geral do público: *as maiores empresas anunciantes do mercado publicitário mundial também exercem pressão direta para que suas propagandas não apareçam ao lado de mensagens extremistas ou problemáticas em plataformas digitais*. A Procter & Gamble chegou a suspender durante um ano todos seus anúncios no Youtube – atualmente, eles são veiculados apenas em canais selecionados. Depois de denúncias de que pedófilos estariam compartilhando trechos de vídeos infantis regulares entre si na plataforma, o próprio Youtube tomo uma medida "drástica" de impedir comentários em vídeos de crianças, buscando afastar de si esse problema, além de manter a plataforma viável a anunciantes. Em 2019, foi lançada a Global Alliance for Responsible Media, que reúne anunciantes, agências de publicidade, companhias de mídia, plataformas e demais atores com o objetivo de "abordar ambientes de mídia danosos ou desinformadores; e para desenvolver e entregar um conjunto concreto de ações, processos e protocolos para proteger marcas". Ver: "These brands spend nearly $100 billion on ads. They want Facebook and

Atualmente, *o Marco Civil da Internet silencia sobre essa governança privada* – que também gera óbvias e importantes implicações ao exercício de direitos fundamentais, notadamente à liberdade de expressão na internet. De fato, quando da promulgação da legislação, sequer havia informações públicas robustas sobre a extensão da moderação de conteúdo realizada pelas grandes redes sociais disponíveis ao público.

Não é assim na correspondente legislação americana, que exerceu um papel fundamental na modelagem da internet comercial que, a partir daquele país, dominou o mercado mundial.

A seção 230 do "Communications Decency Act*"* de 1996 garante hoje uma blindagem de responsabilidade civil praticamente absoluta às plataformas. A lei já foi chamada de "*as 26 palavras que criaram a internet*", por possibilitar, a partir do mercado americano, a expansão de um modelo de internet comercial que passou a prevalecer globalmente em meados dos anos 2000: a "Web 2.0*"*, na qual plataformas operam a partir de conteúdos criados por usuários (Capítulo 1-A)[309]. Assim como o Marco Civil da Internet, a lei americana também prevê exceções a essa regra geral – que abarcam principalmente crimes federais (na prática, especialmente, pornografia infantil) e direitos autorais[310]. Mas, ao mesmo tempo em que cria essa imunidade em sua versão mais forte, prevê expressamente a chamada regra do "bom samaritano" – que permite que esses intermediários restrinjam em boa-fé a publicação de materiais "obscenos, lascivos ou de modo geral condenável", "seja ou não esse material constitucionalmente protegido", sem que isso implique a perda daquela imunidade. A regra do "bom samaritano" *incentiva* a

---

Google to raise their game", reportagem publicada por *CNN Business*, em 23/01/20; Global Alliance for Responsible Media, www.wfanet.org.

[309] Em tradução livre dos dispositivos legais relevantes: "(c) Proteção para bloqueio e filtragem 'bom samaritano' de materiais ofensivos: (1) Tratamento de 'publisher' ou 'speaker': nenhum provedor ou usuário de um serviço interativo de computador será tratado como responsável ('publisher' ou 'speaker') de qualquer informação produzida por outro provedor de conteúdo de informação; (2) Responsabilidade civil: nenhum provedor ou usuário de um serviço de computador interativo pode ser responsável por (A) qualquer ação para voluntariamente restringir acesso ou disponibilidade de material que o provedor ou usuário considere obsceno, libidinoso, lascivo, nojento, excessivamente violento, assediador, ou de modo geral condenável, seja ou não esse material constitucionalmente protegido; (B) qualquer ação tomada para possibilitar ou tornar disponível para provedores de conteúdo de informação ou outros os meios técnicos para restringir o acesso a materiais descritos no parágrafo (1)". Para uma "biografia" atualizada da lei, ver: Jeff Kossef, *The Twenty-Six Words That Created the Internet*, Cornell University Press, 2019.

[310] Tal como abordado no Capítulo 2-A. Mais recentemente, a legislação Fosta/Sesta, promulgada em 2018, incluiu exceções para crimes federais e estaduais relativos ao tráfico de pessoas, o que levantou críticas de algumas entidades que militam em favor de direitos digitais por considerarem um precedente perigoso de relativização da regra geral da Seção 230 da "Communications Decency Act" – "How Congress censored the internet", editorial da *Electronic Frontier Foundation*, publicado em 21/03/2018.

implementação de uma política de moderação de conteúdo pelas empresas (o que afasta qualquer ideia de neutralidade desses intermediários), garantindo que isso não implique uma responsabilidade civil comum a editores tradicionais e mantendo, portanto, aquela completa isenção de "civil liability"[311].

Ainda assim, mesmo atualmente nos Estados Unidos, é comum que sejam feitas avaliações no sentido de que, se plataformas (especialmente as redes sociais) não forem "neutras", elas perdem o direito à imunidade de responsabilização civil prevista pelo "Communications Decency Act". Essa ideia contraria *expressamente* a legislação americana, que desde seu início teve o objetivo claro de *permitir e fomentar a construção de políticas de moderação nos ambientes privados digitais*[312].

É também essa confusão sobre a relação entre "isenção de responsabilidade civil" e "neutralidade" que motiva a recente "ordem executiva para a prevenção da censura online" promulgada pelo presidente americano Donald Trump em maio do ano eleitoral de 2020, sob a alegada defesa da liberdade de expressão. A normativa busca impor a órgãos do governo federal a interpretação de que, se a ação de moderação prevista a "bons samaritanos" deixar de ser exercida com "boa-fé", em havendo uma atuação para silenciar os pontos de vista com os quais as empresas discordam, aquela imunidade cível deixa de ser aplicável, devendo a plataforma ser tratada como um editor ou "publisher" qualquer[313]. Em seu aspecto central, a ordem executiva contraria frontalmente a

---

311 A regra foi criada inicialmente como uma resposta legislativa a decisões judiciais anteriores que impunham uma responsabilização cível a plataformas pelas postagens de terceiros nos anos iniciais da internet comercial. A regra do "bom samaritano" foi crucial para o desenvolvimento de suas políticas de moderação de conteúdo – processo que, no caso das redes sociais, foi descrito no Capítulo 3-A. Além dessa importante lei, essa liberdade de criação de regras por plataformas americanas foi influenciada também pela doutrina de direito constitucional naquele país, no qual direitos constitucionais não são oponíveis a entes privados ("state action docrine"). Sob essa cultura constitucional, entes privados tem ampla liberdade para criar regras, que são interpretadas como adstritas ao nível da liberdade contratual. A jurisprudência que se desenvolveu a partir da aplicação da lei, além disso, em geral reforçou e ampliou essa forte imunização de responsabilidade mesmo em casos limítrofes surgidos décadas após sua promulgação, em contextos tecnológicos inteiramente novos – o que tem levado a debates recentes sobre eventuais reformas legislativas. Ver: Jeff Kossef, *The Twenty-Six Words That Created the Internet*, Cornell University Press, 2019. Para referências de algumas propostas de reforma desse dispositivo legal, ver nota de rodapé nº 360, neste capítulo.

312 Sarah Jeong, "Politicans want to change the internet´s most important law. They should read it first", artigo publicado no jornal *The New York Times*, em 26/07/2019. É claro que, quando a lei foi promulgada, em 1996, sequer era imaginável a existência das redes sociais e de grandes plataformas tais como são hoje, incluindo o escopo de suas governanças privadas de discursos; ainda assim, a previsão dessa função de moderação pelas plataformas, com intervenção restritiva sobre as postagens de usuários, estava presente desde o início da primeira, mais decisiva e mais influente legislação sobre responsabilidade civil de intermediários digitais.

313 Acesso à íntegra da ordem executiva em: https://www.whitehouse.gov/presidential-actions/executive-order-preventing-online-censorship/. O primeiro dispositivo da normativa enuncia que "Twitter, Facebook,

legislação federal em vigor e a jurisprudência dela decorrente; por essa razão, é seguro antecipar que deve encontrar dificuldades substanciais após impugnações judiciais[314].

Em sentido semelhante, a linguagem adotada por precedentes recentes do Superior Tribunal de Justiça também pode dar causa a algumas confusões do tipo: a não realização de um controle editorial tradicional não implica dizer que há um dever ou posição de neutralidade (ou ausência completa de intervenção editorial)[315] – conforme

Instagram e Youtube exercem imenso, senão sem precedentes, poder para modelar a interpretação de eventos públicos; e para censurar, apagar ou fazer sumir informações; e controlar o que pessoas podem ou não podem ver". Seu texto expressamente enquadra plataformas de redes sociais como "praças públicas do século XXI" – uma categorização jurídica que visa a implicação, à luz da jurisprudência constitucional daquele país, de um dever de neutralidade com relação a pontos de vista de agentes discursivos ("speakers"), pois tende a impor a esses "fóruns" obrigações análogas às do poder público (estado). Nesse ponto, a ordem executiva também evoca o precedente Packingham v. North Carolina (2017) da Suprema Corte americana. Nele, a corte julgou inconstitucional a vedação criminal pelo respectivo estado do acesso a redes sociais (vagamente definidas) da internet para réus condenados anteriormente por crimes sexuais, sob o argumento de que tais plataformas são "as praças públicas do século XXI" e que, portanto, essa vedação por uma lei estadual contrariava a Primeira Emenda. A vedação do caso possuía caráter criminal e o réu havia sido preso por realizar postagens no Facebook. A esse respeito, é possível destacar que há uma grande distância entre julgar inconstitucional a criminalização de acesso a plataformas de internet – nesse sentido específico consideradas como "praças públicas" – e concluir, a partir disso, que elas não podem realizar uma moderação de conteúdo de cunho editorial porque essa prerrogativa deve sofrer limitações frente ao direito de liberdade de expressão que são análogas àquelas aplicáveis ao poder público. Para um panorama sobre as implicações do julgamento de Packingham v. North Carolina e as possíveis categorizações de redes sociais sob conceitos tradicionais da jurisprudência constitucional da Primeira Emenda no direito americano, ver: Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1609 e seguintes. Para além disso, por meio da ordem executiva, a presidência exorta a "Federal Communications Comission" (FCC) a promulgar uma regulação sobre as circunstâncias em que a moderação seria considerada como de "boa-fé", de acordo com essa leitura da seção 230 da "Communications Decency Act". Também indica ao "Department of Justice" que avalie e supervisione as condutas de moderação de conteúdo por essas plataformas, entre outros aspectos.

[314] Sua edição ocorreu poucos dias depois de uma postagem de Trump que colocava em xeque a segurança do próximo processo eleitoral presidencial ser marcada pelo Twitter para incluir conteúdos jornalísticos de "fact-checking" que o desmentiam. Politicamente, a ordem executiva simboliza uma reação incisiva da Casa Branca às tentativas do Twitter de exercer maior moderação sobre conteúdos postados pelo presidente em sua conta pessoal, em oposição à postura menos intervencionista, nesse particular, adotada pelo Facebook. Para além dessa conjuntura, contudo, paira também sobre a iniciativa uma forte ironia: uma diminuição do escopo da proteção cível conferida a essas plataformas criaria um incentivo para que elas cerceassem conteúdos controversos em seus ambientes, por receio de futuras responsabilizações judiciais – o que, potencialmente, prejudicaria a ampla liberdade de publicação de perfis exatamente como de Donald Trump. Ver, nesse sentido: "Trump's order on social media could harm one person in particular: Donald Trump", reportagem publicada pelo jornal *The New York Times*, em 28/05/2020; Charles Duan e Jeffrey Westling, "Will Trump's executive order harm online speech? It already did.", artigo publicado no *Lawfare Blog*, em 01/06/20.

[315] "A responsabilidade dos provedores de conteúdo de internet em geral depende da existência ou não do controle editorial do material disponibilizado na rede. Não havendo esse controle, a responsabilização somente é devida se, após notificação judicial para a retirada do material, mantiver-se inerte. Se houver o controle, o provedor de conteúdo torna-se responsável pelo material publicado independentemente de notificação" – Superior Tribunal de Justiça, Recurso Especial nº 1.568.935/RJ, ministro relator Ricardo Villas Bôas Cueva, 3ª Turma, votação unânime e julgamento em 05/04/2016. Vale destacar, porém, que a decisão aborda caso da extinta plataforma Orkut, de propriedade do Google, que realmente não implementava mecanismos de controle ou de curadoria algorítmica comparáveis às redes sociais atuais.

abordado no Capítulo 3-D. Do mesmo modo, não realizar a filtragem ou monitoramento prévio de *todas* as postagens não significa dizer que não se *faça exatamente esse tipo de controle "ex ante" para uma parcela dos conteúdos* (Capítulo 2-A, abordando casos de pornografia infantil, direitos autorais e terrorismo)[316].

Esse silêncio do "estatuto da internet brasileiro" sobre a governança privada de discursos por intermediários parece causar algumas confusões e inconsistências em argumentos recorrentes sobre a regulação de discursos no meio digital, que podem ser observadas inclusive nas discussões em andamento sobre as próprias regras do Marco Civil da Internet. Elas serão expostas a seguir, identificadas por meio de perguntas que serão respondidas no próximo tópico.

A primeira delas é relativa à proposição de que *apenas* o poder judiciário poderia determinar a derrubada de postagens. Em parecer apresentado ao STF no qual defende a constitucionalidade do art. 19 do Marco Civil da Internet, a procuradoria-geral da República argumenta que redes sociais não podem ter a prerrogativa retirar conteúdo do ar, pois essa seria exclusiva do poder judiciário, que teria um monopólio sobre a resolução de conflitos entre a liberdade de expressão e direitos da personalidade. Segundo a manifestação, a legislação vigente prevê "a necessidade de intermediação judicial para a superação do conflito concreto surgido entre esse direito e outros também revestidos de fundamentalidade dentro da ordem constitucional vigente"[317].

---

316 "Não há no ordenamento jurídico qualquer dispositivo legal que obrigue o recorrente a realizar um 'monitoramento' das informações e conteúdos que serão disponibilizados pelo extinto ORKUT ou por qualquer outra aplicação oferecida pelo recorrente. Aliás, na hipótese dos autos, esse chamado monitoramento nada mais é que a imposição de censura prévia à livre manifestação em redes sociais. Conforme entendimento desta Corte, o controle editorial prévio do conteúdo das informações se equipara à quebra do sigilo da correspondência e das comunicações, vedada pelo art. 5º, XII, da CF/88. Não bastasse isso, a avaliação prévia do conteúdo de todas as informações inseridas na web eliminaria um dos maiores atrativos da internet, que é a transmissão de dados em tempo real" – Superior Tribunal de Justiça, Recurso Especial nº 1.342.640-SP, ministra relatora Nancy Andrighi, 3ª Turma, votação unânime e julgamento em 07/02/2017. Aqui, assim como na rota de rodapé anterior, a decisão refere-se ao Orkut, desprovido de mecanismos de controle ou de curadoria algorítmica comparáveis às redes sociais atuais.

317 Ao defender a constitucionalidade do art. 19 do Marco Civil da Internet, a procuradoria-geral argumenta que um sistema alternativo de "notificação e retirada" "acabaria na prática por transferir àqueles entes privados o poder de decidir as colisões eventualmente surgidas entre direitos fundamentais de usuários da rede mundial de computadores, poder este que, se mal exercido, poderia ter evidente impacto na liberdade de expressão, abrindo-se espaço à prática de monitoramento e censura das publicações efetuadas no espaço cibernético. Haveria, em outras palavras, a transferência de um poder de decisão que, no Estado de Direito brasileiro, é típico do Poder Judiciário, para as empresas gestoras de aplicações da internet, as quais, em última análise, receberiam as demandas de seus usuários e julgariam se o conteúdo contestado violaria direitos da personalidade, atentaria contra a honra de alguém ou descumpriria algum mandamento constitucional, concluindo, ao final, pela manutenção ou remoção desse conteúdo do ambiente virtual" – Parecer da procuradoria-geral da República, recurso extraordinário n° 1.037.396/SP

Essa proposição de que a derrubada de conteúdos deve ser uma prerrogativa exclusiva do poder judiciário já ecoou também, sem maiores considerações, em decisão do Superior Tribunal de Justiça, ao se afirmar que "não há respaldo na legislação ou na jurisprudência que permitam atribuir a um particular a prerrogativa de determinar a exclusão de um conteúdo. (...) a ordem que determina a retirada de um conteúdo deve ser proveniente do Poder Judiciário e, como requisito de validade, ser identificada claramente"[318]. Aqui, novamente, a afirmação parte de uma perspectiva que considera somente a aplicação das regras do art. 19 do Marco Civil da Internet – que, afinal, termina por ser a base normativa para as controvérsias que chegam àquela corte. Essa linha de raciocínio parece pressupor que, com a vigência dessa lei a partir de 2014, as derrubadas de postagens seriam feitas apenas por ordens judiciais, já que o abandono do sistema de "notificação e retirada" pelo Marco Civil, supostamente, acabaria por completo com a necessidade de as plataformas tomarem decisões autônomas de derrubadas.

A primeira pergunta, então, permanece: faz sentido argumentar por uma prerrogativa exclusiva de derrubada de postagens pelo judiciário, inclusive diante dos dispositivos do Marco Civil da Internet?

Uma certa perplexidade sobre como harmonizar a convivência dos dispositivos do Marco Civil da Internet com a moderação de conteúdo implementada pelas grandes redes sociais também caracteriza outra controvérsia judicial concreta: em 2019, um órgão do Ministério Público Federal em Goiás ajuizou uma ação civil pública em face do Facebook para contestar a possibilidade jurídica de a plataforma remover conteúdos por conta própria, sem que tivesse havido prévio pedido de interessado nesse sentido[319].

---

[318] Nesse caso, o Superior Tribunal de Justiça, aplicando corretamente os dispositivos do Marco Civil da Internet, derrubou decisão do tribunal estadual paulista que determinava a derrubada de vídeo do Youtube sem indicação precisa de URL, ao contrário do que determina a lei – Recurso especial nº 1.698.647/SP, ministra relatora Nancy Andrighi, 3ª Turma, votação unânime e julgamento em 06/02/2018.

[319] Tribunal Regional da 1ª Região, ação civil pública n. 1005155-11.2019.4.01.3500, 8ª Vara da Justiça Federal de Goiânia. Antes do ajuizamento da ação, uma proposta de termo de ajustamento de conduta em sentido semelhante foi enviada à empresa – e por ela negada. Além disso, os argumentos dessa ação ecoam também no parecer apresentado em 2019 pela procuradoria-geral da república ao STF nos autos do recurso extraordinário nº 1.057.258/MG, ainda pendente de julgamento e também com repercussão geral reconhecida. O caso é anterior à vigência do Marco Civil da Internet e nele a PGR defende a conveniência de um sistema de "notificação e retirada" (operante antes da atual legislação), pois a derrubada sem notificação prévia "poderia esbarrar no direito à liberdade de expressão e de opinião dos usuários, quando, por juízo próprio e sem provocação de qualquer interessado, o gestor de hospedagem excluir dados ou censurar manifestações legítimas dos usuários. É de se perceber que essa autorização poderia redundar em clara censura à liberdade de pensamento e de expressão, bem como no cerceamento unilateral de ideias ou críticas contrárias a certas pessoas ou posições políticas sem a necessária e idônea motivação" – Parecer da procuradoria-geral da República, recurso extraordinário n° 1.057.258/MG.

Para o procurador da República responsável, o Facebook promove uma "censura ilícita" que contraria, em resumo, o direito fundamental de liberdade de expressão tal como previsto na constituição e legislação brasileiras, além de tratados internacionais. A ação mirou especialmente as derrubadas (ou diminuição de alcance de postagens) e as suspensões de perfis decorrentes das regras que combatem notícias falsas ou discursos de ódio. Também enfatizou que as decisões desse tipo eram tomadas sem contraditório prévio ou sem exposição das razões ou motivos aos usuários afetados, "de forma unilateral e não transparente"[320].

Sobre os temas das notícias falsas e de discursos de ódio, a ação argumenta que são categorias inexistentes no ordenamento jurídico brasileiro. Ao registrar que, durante procedimento administrativo investigatório prévio, o Facebook informou que a análise de aplicação das regras de vedação a discursos de ódio não era feita apenas por tecnologias automatizadas, sendo necessária uma avaliação contextualizada por equipes de revisão, o MPF argumentou que a plataforma não realiza julgamentos "puramente objetivos", valendo-se de "juízo subjetivos de seres humanos" – o que implicaria um caráter de arbitrariedade.

Embora sem desenvolver esse ponto com mais detalhes, a procuradoria também afirmou que o Marco Civil da Internet "não permite que os provedores de aplicações realizem diretamente controle relativamente ao conteúdo publicado por terceiros, à medida em que condiciona a sua indisponibilidade ao cumprimento de ordem judicial específica; em contrapartida, isenta os mesmos provedores de responsabilidade civil pelo que publicam terceiros"[321]. A medida dá a entender que, salvo ordem judicial, o Facebook deveria remover apenas conteúdos considerados ilegais, nos termos da legislação brasileira, e em casos de "provocação prévia de sujeito de direitos eventualmente prejudicado". O estabelecimento de um requisito de existência de um terceiro prejudicado remete à ideia de dano ou infração específica a direitos da personalidade.

[320] Esse ponto específico a respeito de direitos procedimentais – ou do *devido processo digital* – está de acordo, por sua vez, com os predicados normativos do constitucionalismo digital, conforme demais tópicos deste capítulo.

[321] Além disso, a *neutralidade de rede* prevista pelo art. 9º do Marco Civil da Internet foi invocada – de modo equivocado – como dispositivo que supostamente impõe a provedores de aplicações uma postura de neutralidade com relação ao conteúdo que transitam em suas plataformas. Esse argumento, contudo, não faz sentido algum, pois essa neutralidade, conforme expressa disposição legal e entendimento comum, aborda a questão de distribuição isonômica de pacotes de dados.

Ao final, a ação foi malsucedida, pois o processo foi extinto sem julgamento de mérito, sob o argumento processual de falta de interesse de agir do Ministério Público. Mas uma análise de seus argumentos levanta também mais duas perguntas importantes: apenas conteúdos ilícitos podem ser derrubados das plataformas? E faz sentido que essa derrubada ocorra apenas quando haja infração ao direito de uma terceira parte prejudicada?

4.D – A concorrência entre as decisões autônomas de moderação pelas redes sociais e as decisões judiciais

Fomentada pela regra do "bom samaritano" da legislação americana, as grandes redes sociais desenvolveram sistemas de moderação de conteúdo com uma abrangência inédita, que são na realidade *constitutivos de suas próprias atividades-fim* (Capítulos 2 e 3)[322]. De certo modo, legislações como a brasileira – que, ao criarem regimes favoráveis de responsabilidade civil a intermediários digitais, *isenta-os de formularem juízos sobre a licitude ou a ilicitude de um conteúdo que lhes tragam consequências jurídicas diretas* – colaboram com esse processo de governança privada autônoma, pois diminuem os custos e riscos legais de suas decisões.

E há *de fato* uma convivência de esferas decisórias distintas. Em um contexto mais geral, esse pano de fundo traz uma *dinâmica potencialmente competitiva entre os sistemas normativos das plataformas e de ordens jurídicas tradicionais*: a) estados continuam determinando por suas leis ou decisões judiciais quais conteúdos ou discursos são proibidos em seus territórios e, nesses casos, as plataformas, enquanto empresas, naturalmente costumam se sujeitar a tais decisões no âmbito daquele território; b) podem ocorrer situações mais excepcionais em que uma plataforma transnacional ignora a existência de uma proibição jurídica nacional, porque não visualiza um risco jurídico prático e concreto às suas atividades, ante a falta de interesse do respectivo governo de aplicação daquela regra, como exemplifica o caso da permissão de negação de Holocausto pelo Facebook em países onde isso é ilegal[323]; c) por fim, restam as situações – essas

[322] Como aponta Tarleton Gillespie, "moderação é, em muitos aspectos, a commodity que plataformas oferecem" – *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*, Yale University Press, 2018, p. 13.

[323] Capítulo 2.C; diante dessa *dinâmica potencialmente competitiva*, os estados nacionais possuem uma alavancagem para fazer valer seus ordenamentos tão mais alta quanto mais importante for o mercado constituído por sua população.

sim, mais comuns – em que as plataformas realizam a curadoria de seus conteúdos, retirando do ar ou invisibilizando conteúdos que podem ser inclusive perfeitamente lícitos, mas contrários a seus termos de uso, segundo suas próprias interpretações.

Ao judiciário brasileiro certamente cabe a palavra final sobre juízos de legalidade ou ilegalidade de um determinado conteúdo, incluindo decisões cogentes sobre sua eventual derrubada. Esse lado da moeda é dotado de sua própria complexidade, pois processos judiciais podem veicular casos difíceis sobre liberdade de expressão, além do sempre presente risco de más decisões judiciais determinarem incorretamente a derrubada de postagens. Ainda assim, à luz do que já foi exposto, *não faz sentido argumentar por uma prerrogativa exclusiva do judiciário para decisões desse tipo*[324].

Um posicionamento nessa linha de certo modo briga com a realidade: ignora, no caso das grandes redes sociais, *o que elas são* e *o que elas fazem*. Parece partir de uma premissa de que se tratam de plataformas neutras que não exercem julgamentos valorativos em suas atividades-fim, uma premissa incorreta. Costuma ignorar também o modelo do marco jurídico definidor da Web 2.0: a seção 230 do "Communications Decency Act*"* – que separa completamente a imunidade de responsabilização civil de intermediários digitais de um suposto dever de neutralidade. Uma certa confusão que é até compreensível, seja porque as próprias empresas durante muito tempo não difundiam publicamente informações sobre a extensão da moderação realizada por elas, seja porque esse *novo tipo de função editoral* foi *se desenvolvendo e* se *aprofundando* ao *longo dos últimos anos*, à medida em que novos problemas e desafios se apresentavam (Capítulo 3).

Em um nível mais fundamental, um ponto de vista que atribui ao judiciário uma capacidade institucional exclusiva para lidar com problemas e casos difíceis envolvendo discursos e expressões no mundo digital ignora também *o mundo tornado Cosmópolis*, no qual a regulação de discurso por estados convive necessariamente com essa emergente regulação privada e transnacional (Capítulo 1). Por isso, fecha seus olhos para

[324] Este argumento da tese corrobora um posicionamento praticamente consolidado entre especialistas de direito digital que analisam o Marco Civil da Internet e o tema da responsabilidade civil de provedores de aplicações. Ainda assim, o restante desta seção pretende reforça-lo com novas e singulares perspectivas condizentes com o desenvolvimento do trabalho. Nesse sentido, reconhecendo a prerrogativa de plataformas para derrubarem conteúdos com base em seus termos de uso: Renato Opice Blum, Paulo Sá Elias e Renato Leite Monteiro, "Marco regulatório da internet brasileira: 'Marco Civil'", artigo publicado pelo portal *Migalhas*, em 20/06/2012; Carlos Affonso Souza e Ronaldo Lemos, *Marco Civil da Internet: construção e aplicação*, Editar Editora, 2016, especialmente pp. 100 e seguintes; Carlos Affonso Souza e Chiara Spadaccini de Teffé, "Responsabilidade dos provedores por conteúdos de terceiros na internet", artigo publicado por *Consultor Jurídico*, em 23/01/2017.

possibilidades de governanças *mais inteligentes*, que conjugam a convivência dessas esferas decisórias distintas.

Afinal, há *potencial positivo* em uma competição entre sistemas decisórios concorrentes sobre o que pode ou não pode ser permitido nos gigantescos ambientes das grandes redes sociais. O judiciário e as plataformas possuem *capacidades institucionais muito distintas para lidar com postagens problemáticas* – que podem ser ora inapropriadas, ora danosas, ora ilícitas, por vezes dotadas de todas essas características ao mesmo tempo.

As redes sociais podem tomar medidas *menos drásticas* (ou seja, com menor grau de intervenção ao direito de liberdade de expressão) do que apenas a derrubada de postagens, pois há uma gradação de respostas possíveis para conteúdos considerados fronteiriços (Capítulo 2-F): podem diminuir sua visibilidade e distribuição, embora mantendo-a publicada no perfil original. Também podem manter uma postagem no ar, mas atrelá-la a algum aviso sobre a falsidade ou incorreção de seu conteúdo, o que dá ao público um contexto informativo mais correto e acurado, levando em conta o ambiente informativo de todo o ecossistema da internet[325]. Redes sociais, afinal, implementam suas políticas entre as lógicas do permitido/proibido e *também* do visível/invisível, que marcam essa *nova espécie de função editorial*. O judiciário, por sua vez, aplica a lógica permitido/proibido a partir da juridicidade do conteúdo: tratando-se de conteúdo ilícito, pode determinar sua derrubada nos termos do Marco Civil da Internet.

Além disso, redes sociais possuem uma capacidade própria de monitoramento e filtragem que lhes permite identificar mais rapidamente conteúdos problemáticos, impedindo sua distribuição – tal como no caso da transmissão do ataque do atirador no massacre de Christchurch, na Nova Zelândia (Capítulo 2-A). O mesmo argumento vale para postagens de "deepfakes", por exemplo. Aguardar um pronunciamento judicial ou mesmo por uma notificação, em casos assim, seria aguardar por uma decisão totalmente ineficaz, depois de uma viralização já consolidada. Essas capacidades técnicas são singulares a cada plataforma, pois elas definem o "código" que determina as condições

[325] Ao comentar a disposição do Facebook de combater a publicação de vídeos digitalmente manipulados, os chamados "deepfakes", Monica Bickert, "head of global policy" da empresa, exemplifica: "se simplesmente removêssemos todos os vídeos manipulados sinalizados pelos verificadores de fatos (´fact-checkers´) como falsos, os vídeos ainda estariam disponíveis em outros lugares na Internet ou no ecossistema de mídia social. Ao deixá-los e rotulá-los como falsos, estamos fornecendo às pessoas informações e contextos importantes" – "Facebook bans deepfakes but permits some altered context", reportagem publicada pelo *The Wall Street Journal*, em 07/01/20.

de publicações de discursos em seus ambientes. Por isso, estão *melhor posicionadas* para apresentar *soluções mais ágeis e menos custosas* para lidar com esses conteúdos problemáticos.

Ainda por esse lado, uma pessoa que se sente prejudicada por uma postagem (casos de "bullying", injúria, difamação, etc) também pode potencialmente ter acesso a mecanismos mais ágeis e baratos de reclamação levam à derrubada da publicação, *sem a necessidade de um acesso mais trabalhoso e custoso ao judiciário* – ainda que essa última opção fique preservada para medidas posteriores, especialmente para demandas indenizatórias contra seu autor.

Além disso, o judiciário também costuma agir a partir de uma *lógica de litigância* pressupõe em regra a existência de uma parte prejudicada. Decisões proferidas nos termos do Marco Civil da Internet são normalmente tomadas nos casos em que há provocação por uma parte interessada na derrubada de um conteúdo – casos que envolvem a oposição entre a liberdade de expressão e direitos de personalidade[326]. Mas essa lógica é insuficiente para lidar com os desafios próprios a esses ambientes digitais. Problemas colocados por conteúdos de desinformação com graves consequências – como informações falsas sobre vacinas obrigatórias, relacionando-as a doenças ou a transtornos de saúde – ou ataques generalizados contra minorias não envolvem a existência de uma parte prejudicada identificável. Daí que *a curadoria exercida pelas redes sociais não deve pressupor a existência de uma reclamação por parte de terceiro, condição presente na arquitetura comum de acesso ao judiciário*[327].

Por fim, resta a questão sobre a possibilidade de as redes sociais retirarem do ar, por decisão autônoma, conteúdos considerados *perfeitamente lícitos*. É compreensível que esse problema levante um mal-estar em face do direito à liberdade de expressão. A retirada do ar de conteúdos lícitos parece contrariar uma noção intuitiva de que isso fere os direitos que as pessoas possuem de publicar qualquer conteúdo considerado legal no

[326] Carlos Affonso Souza e Ronaldo Lemos destacam que o Marco Civil da Internet "certamente direciona o equacionamento de uma eventual divergência entre vítima e provedor para o Poder Judiciário. Aqui, o Marco Civil reconhece que é justamente o Judiciário a instância legítima para o deslinde da questão" – *Marco Civil da Internet: construção e aplicação*, Editar Editora, 2016, pp 102 e seguintes.

[327] Não se ignora que pode haver uma demanda por derrubada de postagens apontadas como ilegais feita no âmbito de uma ação coletiva, como pelo Ministério Público, Defensoria Pública ou associações privadas, por exemplo. Além de serem casos excepcionais, não alteram essa diferença qualitativa, em termos de capacidade institucional, de um processo que leva a decisões judiciais decorrentes de uma litigância trabalhosa e custosa promovida por atores externos em comparação a decisões autônomas de plataformas.

país em que vivem. Logo, deveriam essas plataformas observar os mesmos parâmetros de legalidade (determinados pelas leis do estado) e de constitucionalidade (oponíveis ao estado) em casos envolvendo a liberdade de expressão?

Essa solução seria problemática e inadequada, por uma série de razões. Pornografia, nudez e conteúdos violentos, por exemplo, costumam ser perfeitamente lícitos – ainda assim, uma proibição faz sentido para que uma rede social não se torne uma central de distribuição de pornografia ou de materiais chocantes, o que comprometeria sua própria atividade econômica, iniciativas e razões de ser[328] – ou, em outras palavras, sua *autonomia privada.*

Além disso, plataformas são poderosas – mas não são "a" internet. Embora possuam centenas de milhões ou bilhões de usuários, operam dentro de um ecossistema da esfera digital com diversos outros atores. Suas regras valem, respectivamente, apenas para cada plataforma. Mesmo entre elas, uma coexistência de diferentes conjuntos de regras e a constante disputa por mercado garantem algum grau de dinamismo competitivo em suas formulações, evoluções e adaptações normativas. Isso também as distancia de regras do estado, que possuem a pretensão de eficácia total para toda sua comunidade política. Cidadãos estão naturalmente subordinados às leis de onde vivem – que potencialmente implicam uso da força monopolizada pelo estado, com risco de prisão, por exemplo. Leis nacionais e regras de plataformas gigantes e globais, assim, são normas provenientes de *contextos sociais diferentes*, que implicam *circunstâncias fáticas e jurídicas diversas*.

Os exemplos já mencionados de conteúdos de desinformação ou de ataques generalizados contra minorias também ajudam a compreende porque *redes sociais podem trabalhar com categorias de conteúdos problemáticos que não sejam coincidentes com os critérios de legalidade*. Um cidadão comum pode ter uma opinião totalmente equivocada sobre riscos de vacinas obrigatórias ou sobre a não existência do Holocausto – e viver em uma sociedade onde o direito lhe garante a possibilidade de se manifestar

[328] Para Kate Klonick, isso iria "provavelmente criar uma internet que ninguém quer (...) – fazendo com que os atuais problemas de discurso de ódio online, bullying e terrorismo, com os quais muitos ativistas e acadêmicos se preocupam, fossem muito piores". Klonick justifica porque seria incabível, à luz da jurisprudência constitucional americana, opor direitos constitucionais da Primeira Emenda às plataformas por meio da "state action doctrine". Embora esta tese seja construída sobre uma premissa diversa (ou seja, de incidência dos direitos fundamentais nas relações privadas) a substância do argumento permanece – Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech", *Harvard Law Review Volume 131* (2018), p. 1658.

em ambos os sentidos. Mas se o funcionamento dos ambientes institucionais das redes sociais implica riscos de amplificação que elevam essas opiniões a problemas amplos de desinformação, *é razoável dizer que elas devem ter autonomia para combater os riscos inerentes às suas próprias atividades, dentro de seus próprios ambientes*. Trechos editados de um vídeo de um crime de um atirador em um programa jornalístico podem cumprir uma função informativa dentro de um contexto específico (situação que já costuma gerar debates de ética profissional); o mesmo vídeo em seu formato original pode ser proibido por redes sociais para evitar a glorificação do crime ou a exposição de vítimas.

Não é necessário aprofundar aqui um debate sobre como as *noções de risco e de dano* tradicionalmente cumprem um papel importante na delineação do que seriam os limites da liberdade de expressão[329]. O argumento a ser feito por ora é de que há *riscos específicos a serem considerados nos ambientes das redes sociais, que justificam a adoção de critérios eventualmente mais restritivos do que aqueles aplicados ao debate púbico em geral, previstos pelo direito público.*

Em termos de teoria constitucional, que remontam à ideia da incidência de direitos fundamentais nas relações entre particulares (no caso, a liberdade de uma empresa definir as regras de discursos em seus ambientes *versus* a liberdade de expressão de usuários), *se as redes sociais não tiverem um grau de liberdade de criação de regras para seus ambientes que em alguns aspectos sejam mais restritivos do que as regras estatais, isso implica negar a ideia de que tenham qualquer tipo de autonomia para definir o que é dito em seus ambientes*[330]. E a negação da ideia de algum grau de autonomia a essas

[329] Ronaldo Porto Macedo Júnior, por exemplo, chama a atenção para o que considera "o desenvolvimento de um rico aparato conceitual" pela jurisprudência constitucional americana, ao longo de décadas, para lidar com a distinção entre discursos protegidos pela Primeira Emenda e aqueles passíveis de proibição em função de um dano social real ou possível. Em Schenk v. United States (1919), a Suprema Corte aplicou o critério de "perigo claro e real" ("clear and presente danger"), que posteriormente foi alterado para uma versão de teste de "má tendência" ("bad tendency test") em *Whitn*ey v. California (1927*)*: "se um discurso tivesse uma 'má tendência' para causar atos ilegais, poderia ser constitucionalmente proibido". Desde Brandenburg v. Ohio (1969) vige um critério mais protetivo à liberdade de expressão, que permite a defesa pública do uso de força ou de atos ilegais, exceto "quando essa postura ('advocacy') é dirigida para incitar ou produzir iminentes ações ilegais e é provável que incite ou produza de fato tais ações" (critérios de intenção, iminência e probabilidade) – "Freedom of expression: what lessons should we learn from US experience?", *Revista Direito GV Volume 13, n. 1, São Paulo (2017),* pp. 291-292.

[330] Virgílio Afonso da Silva explica porque não é possível demandar que entes privados promovam apenas as restrições consideradas necessárias (de acordo com a *regra da proporcionalidade*) a direitos fundamentais nas relações contratuais: "exigir que particulares adotem, nos casos de restrição a direitos fundamentais, apenas as medidas estritamente necessárias – ou seja, as menos gravosas – para o atingimento dos fins perseguidos nada mais é do que retirar-lhes a autonomia de livremente dispor sobre os termos de seus contratos (...). Se aos particulares não resta outra solução que não a adoção das medidas estritamente

plataformas não faz sentido por razões substantivas – relativas a *o que são*, *o que fazem*, e a quais papéis estruturais esses atores possuem e devem possuir no ecossistema da liberdade de expressão e de livro fluxo de informações na sociedade. O argumento pela existência de um novo tipo de liberdade editorial desenvolvido no Capítulo 3-D justifica a existência desse grau de autonomia privada para a delineação de regras de discursos.

De um modo mais geral, essa *relação dinâmica entre o direito e regras formuladas com algum grau de autonomia por atores institucionais da infraestrutura da liberdade de expressão sempre esteve presente* – como no caso da imprensa ou de universidades, por exemplo. Autores adeptos da "abordagem institucional à liberdade de expressão" enfatizam que a capacidade normativa de estados sempre conviveu com outras provenientes de demais esferas, pois a sociedade produz instituições que desenvolvem uma capacidade singular de estabelecer regras em seus próprios domínios e dentro de suas áreas de competência[331]. Sob essa ideia, em contextos institucionais específicos não faz sentido transpor a mesma lógica ou critérios que caracterizam direitos de um agente discursivo frente ao estado – esses contextos limitados podem justificar regras mais restritivas em nome de finalidades específicas, como no caso de universidades que validam o mérito dos discursos de seu corpo discente de acordo com critérios científicos e acadêmicos[332].

As considerações feitas até aqui buscam desvelar, a partir de perspectivas jurídicas e institucionais, alguns sentidos para essa convivência entre campos decisórios distintos sobre a derrubada de conteúdos em redes sociais. Mas a dicotomia por ora

---

necessárias, não se pode mais falar em autonomia" – *A constitucionalização do direito: os direitos fundamentais nas relações entre particulares*, Malheiros, 2005, pp. 163-164.

[331] Ver: Paul Horwitz, *First Amendment Institutions*. Harvard University Press (2012); Frederick Schauer, "Towards an Institutional First Amendment", *Minnesota Law Review Volume 89* (2005). Essa abordagem possui uma preocupação central: promover a interpretação da liberdade de expressão levando em consideração a existência e o desenvolvimento de instituições sociais que de algum modo sirvam a seus propósitos, viabilizando objetivos normalmente identificados com ela – como a expressão individual, a produção de conhecimento, a circulação de discursos e de ideias, a difusão de informações e a promoção do livre debate público. Sob essa formulação, contextos institucionais adquirem importância nos casos ou problemas que envolvam instituições sociais identificáveis e que operam funções conectadas à liberdade de expressão a partir de regras e culturas próprias, desenvolvidas com relevante grau de autonomia em relação às regras jurídicas estatais, por meio de padrões profissionais ou corporativos próprios, que surgem e se desenvolvem na sociedade de modo dinâmico ao longo do tempo. Trata-se de uma proposta teórica de autores americanos que reagem ao que percebem como incoerências e contradições decorrentes de uma abordagem exclusivamente individual à liberdade de expressão por parte da jurisprudência constitucional da Primeira Emenda.

[332] Se uma pessoa possui a liberdade de expressão (frente ao estado) de negar a existência do Holocausto, por exemplo, isso não significa que ela não possa ser reprovada em uma prova de história ao dizê-lo.

permanece – e, com ela, os desafios para a tutela da liberdade de expressão por parte dos mecanismos tradicionais do direito do estado.

O constitucionalismo digital demanda que o direito nacional não abra mão de uma visão cosmopolita, atenta ao ambiente digital da Cosmópolis – com a qual interage, ao lado de demais países, e sobre a qual possui relevante capacidade de influência e regulação, sem que isso implique pretensões irrealistas de imposições verticais externas ("top-down") a atores e cenários globais[333]. Levar isso em conta permite que as ferramentas tradicionais do direto do estado – que possuem forte poder vinculante dentro de seus territórios – operem soluções e políticas com impactos efetivos e realistas – mais realistas do que, por exemplo, as despropositadas afirmações de que caberia apenas ao judiciário tomar decisões pela derrubada de conteúdos em ambientes digitais.

Por isso, as duas linhas constitutivas do constitucionalismo digital apresentadas neste trabalho – *proteção a um devido processo digital* e *apresentação de argumentos substantivos em casos de restrições a direitos fundamentais* – informam também a formulação de propostas normativas para o judiciário e o legislativo brasileiro diante do tema da moderação de conteúdo pelas grandes redes sociais.

4.E – Constitucionalismo digital, moderação de conteúdo e perspectivas normativas para o judiciário brasileiro

O judiciário por certo mantém sua competência "tradicional" para avaliar pedidos de derrubada de conteúdos, nos termos do art. 19 do Marco Civil da Internet – um vasto campo decisório que traz consigo sua própria sorte de riscos e problemas para a liberdade de expressão. Mas convive também com a governança privada de conteúdos, até porque opera por meio de uma estrutura comparativamente muito cara, lenta e de acesso dispendioso para os usuários. Para além de suas próprias decisões sobre manutenção ou derrubada de postagens, a partir de critérios de legalidade/ilegalidade, *cabe ao judiciário também, diante da redação atualmente em vigor do Marco Civil da*

[333] "Porque as empresas de tecnologia governam com um substantivo grau de autonomia, proteger direitos fundamentais na era digital requer que o constitucionalismo lide com um poder de governança descentralizado. Imposições externas, verticalizadas, por parte de governos, não serão suficientes para mudar a maneira com que essas empresas concebem seus papeis de governantes" – Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 119-120.

Internet, avaliar as restrições de publicações impostas pelas plataformas por meio de suas decisões autônomas. E, eventualmente, revertê-las.

Se uma plataforma decide derrubar um conteúdo, dentro de seu campo de autonomia, ela assume o risco jurídico da avaliação da licitude ou constitucionalidade de seu comportamento por parte do judiciário.

Isso pode parecer trivial – mas em sentido contrário, no já mencionado julgamento do TJ-SC envolvendo o Youtube e a derrubada de uma sátira musical[334], sua empresa controladora Google utilizou-se do argumento segundo o qual o Marco Civil da Internet implicava a possibilidade de responsabilidade civil *apenas* nos casos de descumprimentos de ordens judiciais, *dando a entender que suas decisões autônomas de derrubada estariam cobertas por uma espécie de imunidade desse tipo de responsabilização*. Esse argumento foi corretamente afastado pela decisão unânime do TJ-SC – muito embora o julgamento tenha incorretamente se valido também da premissa equivocada de que qualquer ordem de retirada caberia exclusivamente ao judiciário[335].

A decisão deveria ter chegado ao mesmo resultado de mérito (imposição de indenização à empresa de tecnologia) baseada apenas no fato de que a sátira constituía manifestação lícita que não fere a legislação de direito autoral brasileira, pois é por ela permitida – avaliando, assim, a licitude da derrubada do conteúdo. É pertinente observar que casos que envolvem a suposta violação de direito autoral fazem referência a esse que é um conceito legal. A empresa não possui liberdade para definir o que seria direito autoral no Brasil, uma categoria específica e legalmente determinada, diferente de outras categorias que correspondem à sua autonomia editorial (como restrições a discursos de ódio, negação do Holocausto, e afins). *Ao invocar a violação de direito autoral como fundamento para a derrubada de conteúdo, a plataforma abre sua decisão a eventual revisão pelo judiciário sobre a aplicação dessa categoria legal* – ao contrário da hipótese de apenas aguardar pronunciamento judicial pela retirada e, com isso, permanecer acobertada pela regra do art. 19 do Marco Civil da Internet.

---

[334] Trata-se do caso já mencionado no Capítulo 4-C, relativo à sátira da música "10%", de Maiara e Maraisa: Tribunal de Justiça de Santa Catarina, apelação cível nº 0000447-46.2016.8.24.0175, 3ª Câmara de Direito Civil, votação unânime, julgamento em 06/02/18.

[335] Segundo a decisão, "não se trata, aqui, de exigir que a plataforma decida o direito no caso concreto que, como se sabe, é competência precípua do Poder Judiciário. O que se esperava, na verdade, é que a segunda ré, em observância à legislação de regência, optasse por manter o conteúdo no ar ininterruptamente até decisão judicial em sentido contrário".

Assim, como o caso acima também demonstra, *as razões fornecidas por uma plataforma para a restrição de uma publicação importam para que o judiciário possa avaliar a razoabilidade da medida*.

Para a maior parte dos casos, nos quais as derrubadas são feitas com base nas regras dos termos de uso da própria plataforma, cabe ao judiciário avaliar a razoabilidade dessa restrição – ainda que esses critérios não coincidam com os padrões de legalidade/ilegalidade. Durante a formulação desse juízo, deve ser levada em conta a autonomia editorial da plataforma para estabelecer regras eventualmente mais restritivas do que os padrões de direito público. Considerando essa autonomia (que pode impor, por exemplo, a derrubada de conteúdos de desinformação ou de discursos de ódio), o judiciário deve se pronunciar se a restrição é ou não razoável. *Esse é um papel tradicional que o judiciário exerce em casos envolvendo a eficácia horizontal de direitos fundamentais, que pressupõe uma dimensão de autonomia por parte de entes privados*[336].

Nesse contexto, não é surpreendente que as novas disputas em torno das regras de moderação de conteúdo – a definição das normas sociais prevalecentes em um ambiente digital – sejam levadas também ao judiciário[337]. Esse raciocínio de "judicial review" pressupõe, ademais, que uma plataforma apresente as razões que levaram à restrição da liberdade de expressão, proposição condizente com as linhas constitutivas do constitucionalismo digital. *A falta de apresentação de razões deve militar contra a plataforma em um processo judicial que discuta uma decisão de moderação de conteúdo*.

Nesse exato sentido, uma recente sentença de primeiro grau condenou o Facebook a reativar o perfil de um escritor, bem como a indenizá-lo em R$ 5 mil por

---

[336] Ver: Virgílio Afonso da Silva, *A constitucionalização do direito: os direitos fundamentais nas relações entre particulares*, Malheiros, 2005, especialmente pp. 148-170. Essa autonomia, de modo geral, deve ser compreendida também no contexto dos problemas e desafios próprios à moderação de conteúdo nos ambientes digitais – que, como visto, envolvem particularidades como *soluções de escala*, *possibilidades tecnológicas de monitoramento e filtragem diversas*, *implementações imperfeitas*, etc (Capítulo 2).

[337] Novamente, a questão da nudez (já tratada no Capítulo 4-B) pode ser exemplificativa. Uma ação atualmente em curso no judiciário paulista problematiza a derrubada de uma conta no Instagram voltada a imagens artísticas de nudez – "Instagram veta, fotográfa vai à Justiça e recupera perfil de nus artísticos", reportagem publicada pelo portal *Universo Online*, em 17/11/2019. O processo encontra-se sob segredo de justiça e, por isso, não foi diretamente acessado – Tribunal de Justiça de São Paulo, processo nº 1038392-59.2018.8.26.0002, Foro Regional de Santo Amaro. Conforme informações fornecidas pela advogada da autora, Camila Ramalho, seus principais argumentos decorrem da eficácia horizontal de direitos fundamentais e da necessidade de contraditório e ampla defesa para a derrubada de contas. A sentença de primeiro grau determinou o restabelecimento da conta, chegando a avaliar que o conteúdo das fotos não incidia na proibição da regra. A conta foi derrubada de novo posteriormente e o processo segue em andamento.

danos morais, pelo período em que a página ficou fora do ar. A juíza responsável baseou sua decisão *principalmente no fato de que o Facebook não informou quais teriam sido as regras violadas de seus termos de uso – nem ao usuário, durante o procedimento digital, nem ao judiciário, durante sua defesa processual*[338]. A falta de qualquer motivação foi (corretamente) determinante para a decisão contrária à empresa.

E, ainda que as regras estabelecidas sejam consideradas razoáveis e, portanto, lícitas, *o judiciário também pode avaliar se houve danos decorrentes de equívocos ou erros em sua aplicação.*

Em meio a seus esforços para combater desinformações às vésperas das eleições brasileiras de 2018, o Facebook derrubou a página de um usuário chamado Fernando Haddad, entendendo se tratar de um perfil falso o que era apenas um homônimo real do candidato presidencial. Ao julgar o consequente processo, o tribunal de justiça paulista manteve uma indenização no valor de R$ 2 mil reais ao usuário. E enfatizou que o problema não havia sido a derrubada inicial do perfil, compreensível em meio a iniciativas da empresa para zelar pela real identidade dos usuários, mas sim a falta de correção do equívoco, mesmo depois de serem fornecidas as informações devidas[339].

Prevalece, contudo, *a percepção de que a capacidade do judiciário de influenciar as regras de moderação de conteúdo das redes sociais é essencialmente limitada*. Ainda que as decisões judiciais procurem supervisionar as razões das decisões tomadas pelas empresas, fomentando uma devida transparência decisória, seus impactos serão restritos. A máquina lenta e cara do judiciário será acionada em um número ínfimo de casos – e os indícios dão conta de que os impactos financeiros dessa litigância serão

[338] Tribunal de Justiça de Minas Gerais, processo nº 5000133-18.2019.8.13.0045, 2º Juizado Especial Cível da Comarca de Caeté-MG, sentença proferida em 17/01/20. Ver também: "Juíza condena Facebook a indenizar usuário que teve conta apagada", reportagem publicada pelo site *Consultor Jurídico*, em 25/01/20. Segundo a decisão, "constata-se dos autos que o autor não recebeu nenhuma notificação prévia e fim de regularizar a situação de sua página no Facebook, tendo sua conta desativada repentinamente, sem qualquer justificativa. Ademais, efetuou diversos contatos com a requerida na tentativa de recuperar sua conta, porém não obteve sequer uma explicação do motivo pelo qual a conta foi desativada".

[339] Tribunal de Justiça de São Paulo, apelação cível nº 1122918-53.2018.8.26.0100, 3ª Câmara de Direito Privado, votação unânime, julgamento em 24 de setembro de 2019. Ver também: "Facebook é condenado por bloquear perfil de homônimo de Haddad", reportagem publicada pelo *Consultor Jurídico*, em 08/01/20. Segundo a decisão, "o dano, na verdade, não decorreu do bloqueio da página pessoal do apelante, que ocorreu no exercício regular do direito, mas sim, da manutenção da suspensão, mesmo após a comprovação de que ele não havia descumprido as normas da empresa, fato que deu ensejo à propositura da presente ação. Se o apelado tivesse agido com a cautela devida, analisando a documentação enviada pelo apelante e, após a constatação de que ele não infringiu as normas de serviço da plataforma, tivesse reativado a sua página em tempo razoável, o dano não teria sido configurado".

negligenciáveis para as plataformas. Uma perspectiva talvez mais promissora seja a judicialização, em tutela coletiva, de direitos decorrentes do devido processo digital – recursos contra decisões, apresentação de regras claras e definidas e transparência decisória, por exemplo. Mas seu principal papel deve ficar reservado às decisões para derrubada de conteúdo a pedido das partes, a partir de juízos de legalidade/ilegalidade, nos termos do Marco Civil da Internet – uma seara por si só complexa, mas que não se confunde com o campo das regras de discursos autônomas das grandes redes sociais.

4.F – Constitucionalismo digital, moderação de conteúdo e perspectivas normativas para uma atualização do Marco Civil da Internet

O legislativo, por sua vez, possui uma capacidade naturalmente mais ampla para criar regras e incentivos que moldem a governança de discursos pelas redes sociais, potencialmente condizentes com os predicados do constitucionalismo digital. O extenso debate sobre como governos de todo o mundo devem reagir ao chamado "techlash" (espécie de "ressaca tecnológica") tem ganhado força e envolve temas diversos como proteção à privacidade e concentração de mercado por poucas empresas, entre outros. O próprio Mark Zuckerberg tem defendido publicamente desde 2019 um maior nível de regulamentação sobre as plataformas de internet, em especial em quatro áreas: eleições (com "padrões comuns para verificar quem é um ator político", além de abordar "importantes questões sobre como campanhas políticas usam dados e microdirecionamento"), privacidade, proteção de dados e, também, a moderação de conteúdos considerados problemáticos pelas redes sociais[340]. Este tópico pretende contribuir especificamente com perspectivas para a legislação brasileira a partir dos dispositivos vigentes do Marco Civil da Internet, apresentando avaliações e algumas propostas incrementais de atualização legislativa.

Como regra geral, a isenção de responsabilidade civil de intermediários digitais prevista pelo Marco Civil da Internet deve ser mantida. Sem desconsiderar que há diferentes relações de custos e benefícios entre os regimes possíveis de responsabilidade civil de intermediários digitais[341], *um dos principais aspectos relevados por esta pesquisa é o fato de que, mesmo sob regimes legais que não implicam responsabilização civil*

[340] Mark Zuckerberg, "The internet needs new rules. Let´s start in these four areas", artigo publicado pelo *The Washington Post*, em 30/03/19.

[341] Ver nota de rodapé nº 303, neste Capítulo.

direta, plataformas digitais cedem a outros tipos de incentivos, que não a pura e simples obrigação legal, e criam amplos sistemas de moderação de conteúdo, essenciais para enfrentarem conteúdos problemáticos (ilícitos ou não). Essa moderação não decorre de direta imposição legal, mas ainda assim é feita.

Um maior grau responsabilização civil de plataformas por postagens de usuários criaria um incentivo excessivo para a derrubada de conteúdos, o que seria problemático do ponto de vista da liberdade de expressão. A concorrência de um sistema judicial e outro autônomo por parte de plataformas evita esses incentivos excessivos pela supressão de conteúdos, pois as plataformas, por seu próprio modelo de negócios, possuem um interesse inato em viabilizar as publicações de seus usuários.

Não se ignora que isso aumenta o custo de derrubada de conteúdos considerados ilícitos – pois do ponto de vista das regras da atual legislação, isso seria feito somente mediante o custoso e lento acesso ao judiciário. Mas a história do desenvolvimento das políticas de moderação de conteúdo pelas redes sociais indica que uma forte obrigação legal, por meio de regras de responsabilização civil, não é condição necessária para que essa moderação ocorra: as plataformas respondem a outros incentivos (como mercado e normas sociais) ao fazê-lo. Nesse sentido, se a governança global da internet se basear cada vez mais em legislações como a iniciativa alemã da NetzDG, aumenta a probabilidade de que a sobreposição de legislações nacionais fomente prioritariamente a derrubada de conteúdos em alta escala, criando práticas de censura excessivamente abrangentes.

A NetzDG[342] foi promulgada em 2017 e entrou em vigor no ano seguinte. A lei prevê uma série de obrigações a empresas responsáveis por "redes sociais"[343] na internet, desde que tenham pelo menos 2 milhões de usuários registrados na Alemanha. Esses provedores devem manter um canal específico, determinado pela lei, para que usuários ou entidades da sociedade civil notifiquem a plataforma a respeito de *conteúdos que são ilegais*, tomando por parâmetros cerca de vinte seções do código penal alemão, que

[342] No original alemão, "Netzdurchsetzunggesetz" – com o sentido aproximado de "Aplicação da lei para as redes".

[343] O termo é definido de forma abrangente: "provedores de serviços que, com intuito de lucro, operam plataformas de internet desenhadas para possibilitar que usuários compartilhem qualquer conteúdo com outros usuários ou que tornem esse conteúdo disponível ao público" (Seção 1), definição que pode dar margem a dúvidas. A lei exclui expressamente plataformas que ofereçam conteúdo jornalístico ou editorial, bem como plataformas voltadas a possibilitar "comunicações individuais ou disseminação de conteúdo específico" – Versão em inglês da lei disponível em https://germanlawarchive.iuscomp.org/?p=1245

incluem um amplo leque de delitos – desde incitação a crimes, violação de intimidade por fotografias, formação de organizações terroristas, difamações ordinárias e disseminação de imagens violentas. Para conteúdos que são "manifestamente ilegais", a plataforma deve retirá-los do ar em até 24 horas. Para postagens apenas "ilegais", esse prazo é de 7 dias[344]. As decisões e suas razões devem ser informadas às pessoas envolvidas – tanto autores da notificação, como da postagem em questão. *Na prática, a lei abandona a ideia de imunidade legal ("safe harbour") aos intermediários digitais*[345].

Plataformas que recebam mais de 100 notificações por ano devem publicar relatórios semestrais que detalhem a implementação dos mecanismos previstos pela lei, os critérios usados para avaliar os conteúdos reportados, bem como o número total de reclamações e a quantidade de conteúdos bloqueados, entre outros dados afins. A lei prevê penas de até 50 milhões de euros para diversas infrações – entre elas, deixar de apresentar os relatórios semestrais, desenvolver mecanismos para receber e avaliar as notificações ou "deixar de eliminar inconsistências nesse processo".

É prevista uma forma mitigada de revisão judicial apenas na hipótese de o governo impor multas para falhas diretas na remoção ou bloqueio de conteúdos ilegais – nesse caso, a corte com competência para revisar multas administrativas deve avaliar se o conteúdo de fato é ilegal. No entanto, a decisão é tomada apenas com argumentos prévios e escritos por parte do órgão governamental e da empresa de mídia social. A lei exclui expressamente a necessidade de audiência judicial para debates, bem como veda a possibilidade de recurso contra a decisão tomada. Além disso, não prevê a participação de quaisquer outros envolvidos – como a pessoa que fez a reclamação ou o próprio autor do conteúdo[346].

Ainda permanecem dúvidas relevantes sobre como a implementação da lei irá ocorrer na prática ao longo dos anos. Entre elas, por exemplo, se uma pequena porcentagem de erros de avaliação sobre ilegalidades pode gerar uma responsabilização

[344] A lei prevê, ainda, que as empresas possam delegar esses processos decisórios a um órgão externo, financiado por uma ou mais delas, desde que haja aprovação de sua atuação pelo ministério da justiça, observadas algumas condições – uma espécie de órgão auto regulador do mercado.

[345] "Germany: the act to improve enforcement of the law in social networks", *legal analysis* pela *Article 19*, publicada em agosto de 2017.

[346] "Germany: the act to improve enforcement of the law in social networks", *legal analysis* pela *Article 19*, publicada em agosto de 2017.

administrativa da empresa[347]. Até o momento, apenas o Facebook recebeu uma multa – no valor de 2 milhões de euros – por, de acordo com o governo, apresentar em seus relatórios de transparência um número subrepresentativo do total de reclamações que recebeu sobre conteúdos ilegais em suas páginas[348].

A NetzDG é talvez o símbolo mais importante da reação de países europeus contra a dominância de empresas americanas sobre a internet comercial[349]. Ainda está para ser avaliada a extensão de seus efeitos na governança privada de discursos pelas grandes redes sociais na Alemanha: uma avaliação inicial aponta que, na verdade, a lei serviu principalmente para que plataformas fossem mais diligentes nas aplicações de seus próprios termos de uso, motivadas pelos imperativos de transparência sobre os modos com que lidam sobre as reclamações recebidas[350]. Ainda assim, a iniciativa já influencia

[347] "Germany: the act to improve enforcement of the law in social networks", *legal analysis* pela *Article 19*, publicada em agosto de 2017.

[348] "Germany fines Facebook for underreporting hate speech complaints", reportagem publicada pelo portal da *Deutsche Welle*, em 02/07/2019. A controvérsia, na verdade, resulta do fato de que o Facebook *separou* os mecanismos normais de "flagging" de seus conteúdos (em razão de violações a seus termos de uso) e o formulário digital para notificações de ilegalidades que é mandatório por parte da NetzDG. Por isso, o relatório do Facebook declarava ter recebido no ano anterior 866 reclamações referentes a 1704 postagens, que levaram a 362 remoções. Esse número era significativamente menor que o total de outras plataformas – o Google disse ter recebido 215 mil reclamações no Youtube, com 58 mil remoções, e o Twitter apontou um total de 265 mil reclamações, que levaram à derrubada de 29 mil postagens. A diferença é que o Google/Youtube e o Twitter agregaram o canal para notificações de ilegalidades previsto pelo NetzDG a suas interfaces ordinárias de "flagging". Parte da crítica do governo ao Facebook foi no sentido de que o "formulário NetzDG" era mais escondido e por isso menos acessível ao público que seus canais usuais para conteúdos problemáticos, além da separação em si entre os critérios de "ilegalidade" e "violações a termos de uso". Ver: Heidi Tworek & Paddy Leerssen, "An analysis of Germany's NetzDG law", artigo publicado pelo *Transatlanic High Level Working Group on Content Moderation Online and Freedom of Expression – Institute for Information Law (Universiteit Van Amesterdam)*, em abril de 2019; Jenny Gesley, "Germany: Facebook found in violation of 'anti-fake news' law, artigo publicado pelo *Global Legal Monitor – Library of Congress of the United States*, em 20/08/2019.

[349] Um exemplo claro também de conflito sobre padrões de liberdade de expressão que marca a convivência na Cosmópolis. O governo alemão já tentava desde anos anteriores fazer com que as grandes plataformas cumprissem protocolos mutuamente acordados contra conteúdos de discursos de ódio no ambiente online. A lei foi proposta após as autoridades considerarem que os resultados eram insatisfatórios. Vide Capítulo 1-B e também: David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, pp. 65-74.

[350] Heidi Tworek & Paddy Leerssen, "An analysis of Germany's NetzDG law", artigo publicado pelo *Transatlanic High Level Working Group on Content Moderation Online and Freedom of Expression – Institute for Information Law (Universiteit Van Amesterdam)*, artigo publicado em abril de 2019.

diretamente outras propostas legislativa em andamento em demais países, como na França[351] e na Inglaterra[352].

No Brasil, a lei alemã claramente inspira um projeto de lei apresentado à Câmara dos Deputados em fevereiro de 2020, que prevê a criação do "*rito sumário para a retirada de conteúdos ilegais de redes sociais*": o PL altera busca alterar o Marco Civil da Internet para que seja prevista a remoção de conteúdos "obviamente ilegais" em 24 horas, enquanto a "conteúdos ilegais" seria aplicado um prazo de 7 dias: "conteúdos reclamados e não retirados ou bloqueados nos prazos estabelecidos (...) sujeitam o provedor de aplicações de internet à multa de até R$ 100.000,00 por reclamação não atendida"[353].

De um ponto de vista que busca avançar o constitucionalismo digital, a NetzDG implementa diversos dispositivos que podem ser avaliados como positivos, por fomentarem a transparência e prestação de contas ("accountability") por parte das grandes redes sociais. É nesse viés que se encontra uma forte virtude da lei: as obrigações impostas às empresas para criarem canais de reclamações por usuários, formularem análises que levem a conclusões transparentes e publicarem relatórios periódicos que forneçam dados sobre as práticas de moderação de conteúdo são medidas cruciais para uma governança de discursos que permita um nível apropriado de controle social, construção de um devido processo digital e fornecimento de razões para a restrição de direitos fundamentais.

Nesse aspecto, a lei alemã não apenas inova, mas demonstra como demais legislações podem seguir essa linha e formularem com criatividade novos mecanismos que persigam aqueles objetivos. Entre eles, por exemplo: requisitos legais para que plataformas forneçam uma notificação a autores de postagens que foram reportadas, para que eles possam se manifestar antes de eventual decisão, bem como direito de recurso a esses autores, posterior à decisão pela efetiva derrubada; ou a manutenção de um repositório de todas as postagens derrubadas que possa ser acessado por autoridades

---

[351] "France online hate speech law to force social media to act quickly", reportagem publicada pelo *The Guardian*, em 09/07/19; "France: analysis of draft hate speech bill", artigo publicado por *Article 19*, em 03/07/19.

[352] "Britain's government announces plans to regulate Facebook, Twitter and TikTok", reportagem publicada por *Forbes*, em 12/02/20.

[353] Projeto de lei n. 283/2020 – que foi apensado ao projeto de lei n. 2.712/15, que também busca alterar o Marco Civil da Internet para prever a possibilidade de retirada de conteúdo, "por solicitação do interessado, de referências a registros sobre sua pessoa em sítios de busca, redes sociais ou outras fontes de informação da internet, desde que não haja interesse público atual na divulgação da informação não se refira a fatos genuinamente históricos", dispositivo este, por sua vez, declaradamente inspirado no chamado "direito ao esquecimento" europeu.

públicas, entidades da sociedade civil ou pesquisadores, garantindo um nível amplo de supervisão social sobre o que foi de fato tirado do ar[354].

O aspecto negativo e problemático da legislação alemã fica por conta da imposição de juízos sobre legalidade/ilegalidade às plataformas (com consequente responsabilização jurídica pelas decisões tomadas), por meio do afastamento da possibilidade de elas formularem juízos próprios de moderação de conteúdo que sejam desprovidos de uma vinculação direta com esses juízos de legalidade, a partir de uma liberdade editorial que responde a variados incentivos sociais. A recente consolidação dos gigantes sistemas de moderação de conteúdos demonstra que essa imposição não é necessária – e que a governança privada de discursos pode responder de modo mais produtivo a incentivos legais de natureza não punitiva.

Fomentar uma governança privada de discursos que seja deferente à lógica de direitos e responsiva a demandas sociais parece ser não apenas uma política regulatória mais inteligente para abordar os conteúdos problemáticos do ambiente online, mas também aquela melhor posicionada para resguardar os direitos da esfera digital. Essa perspectiva impede mecanismos de censura colateral por parte de estados, evitando também o cenário que Garton Ash denomina de "poder ao quadrado" entre governos e plataformas digitais[355]. E *permite abordar também problemas graves que não guardam correlação com parâmetros de ilegalidades, como é o caso de conteúdos com informações falsas que impliquem riscos sociais relevantes, incluindo campanhas de desinformação.* Essas razões se somam para apontar como benéfica a existência de regimes jurídicos distintos, a partir de capacidades institucionais distintas. Nada disso impede que mecanismos tradicionais do direito do estado demandem a derrubada de um conteúdo tido por ilegal – uma decisão que, em democracias, costuma ser de competências do judiciário ou de órgãos governamentais dotados de algum grau de imparcialidade.

---

[354] Exemplos fornecidos por: Heidi Tworek & Paddy Leerssen, "An analysis of Germany's NetzDG law", artigo publicado pelo *Transatlanic High Level Working Group on Content Moderation Online and Freedom of Expression – Institute for Information Law (Universiteit Van Amesterdam)*, artigo publicado em abril de 2019. Para os autores, "melhorar direitos de notificação pode promover uma equidade procedimental e nivelar o nível do jogo em favor da liberdade de expressão".

[355] Ver referência da nota de rodapé n. 46 no Capítulo 1.

No caso brasileiro, em um país no qual figuras políticas costumam ter um papel de proeminência nos pedidos de remoção de conteúdos feitos ao judiciário[356], não é difícil antever como plataformas digitais teriam um forte incentivo para derrubada de postagens que fossem reportadas por autoridades como alegados abusos do direito de crítica ou da liberdade de expressão, caso o modelo jurídico de responsabilidade civil do Marco Civil da Internet fosse revertido[357].

Por isso, analisando a atual redação do Marco Civil da Internet (de certo modo obsoleta, porque silencia por completo sobre a moderação de conteúdo feita por plataformas digitais), parece adequado que uma atualização legislativa: a) trabalhe prioritariamente com obrigações e incentivos para que a governança de conteúdos por redes sociais seja pautada por uma lógica de transparência, prestação de contas e implementação de um devido processo digital; b) mantenha a regra geral de isenção de responsabilidade civil a intermediários digitais, com exceções limitadas a situações específicas e justificadas.

Quanto ao primeiro ponto A, uma reforma legislativa brasileira deveria incorporar mecanismos que implementem os eixos dos "Princípios de Santa Clara para transparência e prestação de contas em moderação de conteúdo" para as grandes redes sociais (Capítulo 4-B). Nesse sentido, os aspectos já mencionados e tidos como positivos da legislação alemã fornecem bons parâmetros. Como medidas desse gênero implicam custos de operação, faz sentido também que a legislação module essas exigências para as grandes plataformas, que tenham a partir de um determinado número de usuários registrados no Brasil, tal como faz a legislação alemã, ao mirar exatamente as plataformas dotadas de maior impacto social. Seria possível a legislação determinar a produção de relatórios públicos periódicos, bem como a organização de um repositório que contenha as decisões tomadas a partir das reclamações recebidas, para que possam ser submetidas ao escrutínio de entidades da sociedade civil e pesquisadores.

---

[356] Como retrata o Projeto Ctrl+X da Associação Brasileira de Jornalismo Investigativo (Abraji), que monitora pedidos de supressão de informações da internet por parte de políticos – http://ctrlx.org.br.

[357] Atualmente, após promulgação da lei n. 13.488/2017, a legislação eleitoral replica em sua seara a isenção de responsabilidade civil a intermediários digitais prevista pelo Marco Civil da Internet a provedores de internet que possibilitam o impulsionamento pago de conteúdos, até ordem de remoção da Justiça Eleitoral. Uma reversão ou enfraquecimento desse modelo, a depender de futura decisão do STF, implica um claro incentivo para que plataformas derrubem postagens de teor político a partir da mera reclamação de autoridades ou candidatos. Ver: "Julgamento do Marco Civil no STF pode criar limbo jurídico durante eleições", reportagem publicada pelo *portal Jota*, em 06/03/20.

Mais importante ainda, uma atualização legislativa deveria incorporar a obrigatoriedade de mecanismos do devido processo digital, tais como: publicação prévia e clara das regras dos termos de uso, direito de notificação sobre reclamações feitas ou decisões de derrubada de um determinado conteúdo, direito de recurso contra uma decisão de derrubada em um determinado prazo razoável e necessidade de transparência com relação às razões das decisões de moderação de conteúdo[358].

Todos esses mecanismos – que seriam direcionados à *constituição de obrigações de meio, desvinculados de juízos de mérito sobre conteúdo*s – deveriam se tornar obrigações legais de grandes redes sociais, mediante imposição de penalidades administrativas ou judiciais. Uma construção nesse exato sentido integra o texto original do projeto de lei n. 2.630/20, em trâmite no Senado Federal, oficialmente chamado de "Lei das Fake News"[359].

Tais mecanismos poderiam também se tornar requisitos legais para que grandes plataformas (a partir de um determinado número de usuários) façam jus à regra de isenção de responsabilidade civil prevista pelo art. 19 do Marco Civil da Internet. Em sentido

[358] A lei brasileira já demonstra uma atenção voltada ao contraditório e à ampla defesa, *mas apenas como meio para ação ou defesa perante o judiciário, condizente com a regra de seu artigo 19*. Conforme dispõe seu artigo 20: "Sempre que tiver informações de contato do usuário diretamente responsável pelo conteúdo a que se refere o art. 19, caberá ao provedor de aplicações de internet comunicar-lhe os motivos e informações relativos à indisponibilização de conteúdo, com informações que permitam o contraditório e a ampla defesa em juízo, salvo expressa previsão legal ou expressa determinação judicial fundamentada em contrário; Parágrafo único: Quando solicitado pelo usuário que disponibilizou o conteúdo tornado indisponível, o provedor de aplicações de internet que exerce essa atividade de forma organizada, profissionalmente e com fins econômicos substituirá o conteúdo tornado indisponível pela motivação ou pela ordem judicial que deu fundamento à indisponibilização".

[359] Embora sua redação se autoproclame como a "Lei brasileira de liberdade, responsabilidade e transparência na internet". O projeto mira exclusivamente o combate a campanhas de desinformação em grandes redes sociais (que tenham a partir de dois milhões de usuários no país, conforme seu art. 1º) e aplicativos de mensagens privadas (como o Whatsapp). Em seu texto original, o art. 6º prevê a obrigatoriedade de relatórios de transparência em grande parte nos moldes aqui defendidos; seu art. 12, por sua vez, demanda mecanismo de recurso em face de decisões das plataformas que rotulem uma postagem como desinformação, pelo prazo de três meses, tanto para o autor do conteúdo quanto para o usuário autor de uma denúncia. Embora esses dispositivos em particular estejam alinhados com o posicionamento deste trabalho, parece ser negativo limitar tais iniciativas exclusivamente à moderação de conteúdo feita contra campanhas de desinformação, excluindo as decisões das plataformas nas demais searas de postagens problemáticas ou controversas. Mais importante ainda, a proposta busca impor uma obrigação de utilização pelas plataformas de "verificadores de fatos independentes" sob termos, condições e responsabilidades não muito esclarecidas. Trata-se, de qualquer modo, de um texto inicial de projeto de lei, que até o depósito desta tese não havia sido votado pelo Senado Federal.

semelhante, foram formuladas algumas propostas de atualização da legislação americana[360]. Como aponta David Kaye:

> "*Governos devem encorajar as empresas a adotarem regras transparentes e estratégias para suas aplicações, talvez até mesmo com requisitos regulatórios vinculantes – mas não através de uma regulação de mão pesada de conteúdo ou pelo medo de penalidades que poderiam comprometer a competição e inovação*"[361].

Quanto ao segundo ponto B, a legislação brasileira já prevê exceções pontuais e justificadas à regra geral do artigo 19 do Marco Civil da Internet (Capítulo 4-C), notadamente a aplicação do procedimento de notificação e retirada para imagens de nudez divulgadas sem consentimento de pessoas retratadas. Trata-se de uma estratégia regulatória acertada, que modula o regime jurídico legal para algumas situações e problemas específicos – especialmente nos casos nos quais a ilegalidade de um conteúdo seja de fácil constatação visual e menos dependente de análises contextuais, como imagens de pornografia infantil, por exemplo. Parte significativa da virtude de propostas que orbitam em torno de um eixo de governança transparente e da implementação legal de um devido processo digital está exatamente na desnecessidade de abandono ou enfraquecimento de um regime de isenção de responsabilidade civil a intermediários digitais, que cumpre importante papel na viabilização de um ambiente discursivo livre na internet – caminho oposto ao seguido pela lei alemã NetzDG.

Condizente com os objetivos desta pesquisa, esta análise manteve até aqui seu foco na moderação de conteúdo realizada pelas grandes redes sociais. Mas é importante

---

[360] Esses autores defendem que o regime de quase absoluta imunidade legal dada a intermediários digitais pela Seção 230 do "Communications Decency Act*"* deveria ter como contrapartidas algumas formas de obrigações legais, primordialmente mecanismos que contemplem um devido processo digital: Rebecca Tushnet, "Power without responsability: intermediaries and the First Amendment", *George Washington Law Review Volume 76* (2008), pp. 1015 e seguintes; Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 158-159. Danielle Citron e Benjamin Wittes vão além e fornecem uma proposta que parece ser marcada por uma insegurança jurídica inata: eles defendem que a imunidade legal ("safe harbour") seja condicionada à comprovação fática de que a empresa responsável implementa de fato um "conjunto de medidas razoáveis" para evitar ou retirar do ar as postagens com conteúdos ilegais, uma proposta que deixa em aberto a aplicação da lei a partir de uma categoria bastante indeterminada e sujeita a produção probatória casuística – Danielle Citron & Benjamin Wittes, "The Internet will not break: denying bad samaritans Section 230 immunity", *Fordham Law Review Volume 86* (2017), pp. 401-423.

[361] David Kaye, *Speech Police: the global struggle to govern the Internet*, Columbia Global Reports, 2019, p. 20.

registrar que o marco teórico do constitucionalismo digital permite pensar também em possibilidades legislativas protetivas de direitos fundamentais (e da liberdade de expressão, em particular) para outros grupos que fornecem serviços ligados à internet – principalmente aqueles que atuam no nível "backbone" (espinha dorsal) da rede.

Grandes redes sociais são importantes, mas não controlam níveis básicos de acesso à internet. Se é natural que elas exerçam um nível de liberdade editorial ao formularem suas regras de discursos, por conta de suas funções e modelos de negócios, esse aspecto parece não se aplicar a serviços de hospedagem, registradores de domínio, de defesa de servidores ou outros afins. Como aponta Jack Balkin, "diferentes partes da infraestrutura da internet deveriam ter diferentes responsabilidades para proteger a liberdade de expressão online"[362].

Uma das poucas empresas existentes no mercado para distribuição de conteúdo e segurança de servidores de páginas da internet – a Cloudfare – baniu o site nazista "The Daily Stormer" de sua lista de clientes – inviabilizando, na prática, a existência de sua página na internet padrão e indexada, pois o conteúdo se mudou para a chamada "deep web". A prática rompeu com a tradição da empresa de não fazer qualquer juízo de valor sobre o conteúdo das páginas de seus clientes, exceto em caso de ilegalidades e de ordens judiciais válidas em suas jurisdições[363].

Isso aponta para a conveniência de uma atualização do Marco Civil da Internet também para garantir uma neutralidade editorial absoluta (ou, pelo menos, muito mais forte) para serviços que lidam com o direito de acesso à internet em geral. Embora hoje o art. 18 do Marco Civil da Internet preveja que provedores de conexão à internet não serão em nenhuma hipótese responsabilizados civilmente por danos decorrentes de conteúdos de terceiros, uma atualização legislativa poderia não apenas especificar melhor as características de serviços que identificam esses provedores, mas também criar *para esses*

---

[362] Jack Balkin, "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018), pp. 2036-2040.

[363] Segundo seu CEO, Matthew Prince, "isso foi minha decisão. Nossos termos de serviço reservam a nós o direito de romper com usuários com base apenas em nossa discricionariedade. Minha razão para tomar essa decisão foi simples: as pessoas por trás do Daily Stormer eram idiotas e eu estava cheio disso. Deixe eu ser claro: essa foi uma decisão arbitrária. Foi diversa do que eu havia discutido com minha equipe sênior no dia anterior. Eu acordei nesta manhã de mau humor e decidir expulsá-los da internet. Eu chamei nosso time jurídico e disse para eles que iríamos fazer isso. E chamei nosso time de segurança e falei para eles pararem com o serviço. Era uma decisão que eu podia tomar porque sou o CEO de uma grande empresa de infraestrutura da internet. Tendo tomado essa decisão, é importante que agora conversemos sobre porque isso é tão perigoso (...). Literalmente, eu acordei de mau humor e decidi que alguém não poderia estar na internet. Ninguém deveria ter esse poder". Para um relato do caso, ver: Nicolas Suzor, *Lawless: the secret rules that govern our digital lives*, Cambridge University Press, 2019, pp. 11-14.

outros grupos de atores uma *necessidade de ordem judicial, mediante juízo de ilegalidade, para a derrubada de qualquer conteúdo ou de negação de serviços*.

De modo geral, há margem significativa para que o legislativo atue dentro de parâmetros do constitucionalismo digital – e as soluções regulatórias de estados nacionais vão invariavelmente lidar com os problemas concretos que se apresentam, com as próprias soluções auto regulatórias apresentadas pelas plataformas e com demais movimentos regulatórios promovidos internacionalmente. *As propostas ora apresentadas certamente não possuem a pretensão de esgotar um complexo e multifacetado debate* (é possível pensar na conveniência de previsões legais que lidam especificamente com o período eleitoral ou com o microdirecionamento de anúncios políticos, por exemplo[364]), *mas pretendem contribuir com uma linha geral de atualização da política legislativa brasileira calcada em um diagnóstico sobre como as redes sociais funcionam de fato e quais são as potencialidades e limitações do direito do estado nesse cenário.*

4.G – Considerações finais do capítulo

A partir do marco teórico do constitucionalismo digital, este capítulo apresentou parâmetros normativos para lidar com os dilemas e problemas jurídicos que emergem das políticas de moderação de conteúdos pelas grandes redes sociais. Sua ênfase na noção de um devido processo digital e na necessidade de apresentação de argumentos para a restrição de direitos permite pensar em caminhos tanto para a seara do "direito das plataformas", quanto para mecanismos tradicionais do direito do estado, conjugando ambos esses planos a partir de uma lógica de proteção a direitos e de limitação de poderes.

Parte significativa dessa empreitada ocorre desatrelada do paradigma do estado nacional. Sem descuidar de um nível de autonomia editorial que é devido às plataformas, torna-se possível pensar em novos processos, desenhos institucionais e formas de governança que garantam um nível adequado proteção à liberdade de expressão em uma internet que se mantenha dominada por grandes intermediários digitais – que, por sua vez, também devem lidar com as novas formas de discursos problemáticos que surgem em seus ambientes, sob suas novas lógicas. O direito do estado deve reconhecer as complexidades que caracterizam a governança de discursos na internet para formular

[364] Seria possível elaborar critérios legais que forneçam parâmetros para a moderação de conteúdo específica de postagens ou propagandas políticas de partidos ou candidaturas habilitadas durante o período eleitoral, por exemplo. Eventualmente, com uma revisão automática por parte da Justiça Eleitoral.

regras e decisões jurídicas inteligentes, efetivas e que dialoguem de modo produtivo com essa dinâmica normativa internacional e transnacional.

Não há uma lista fechada de soluções prontas e integrais que possam ser apresentadas de antemão para os inúmeros problemas elencados ao longo deste trabalho, mas este capítulo procurou apresentar parâmetros normativos que, embora abrangentes, fossem concretos e consistentes em um nível suficiente para *articular uma visão constitucionalista, dotada de possibilidades práticas, para enfrentar o problema de pesquisa proposto.*

Considerações finais

Para além dos argumentos e conclusões já apresentados em cada capítulo anterior, esta parte final apresenta algumas últimas considerações gerais e complementares da tese.

Atualmente, não parece ser possível propor reflexões abrangentes a respeito das práticas da liberdade de expressão sem considerar as esferas digitais de debate público – e, por isso, também a governança privada de discursos a cargo dos grandes intermediários digitais. Não apenas porque essas plataformas conquistaram papéis proeminentes no novo ecossistema discursivo e informativo da sociedade, mas também porque, no caso das redes sociais, desenvolveram sofisticados sistemas institucionais de criação, interpretação e aplicação de regras e valores que de muitas maneiras emulam o funcionamento de sistemas jurídicos tradicionais, ainda que com novas características.

Por isso, uma reflexão constitucionalista sobre a liberdade de expressão que se mantenha restrita às normas do direito do estado hoje não basta para enfrentar muitas das questões relevantes que se impõem, inclusive em searas como discursos políticos e eleitorais, sempre conectados a problemas do autogoverno democrático. Assim como a leitura de decisões judiciais é parte essencial da pesquisa e do ensino do direito, é possível pensar que as decisões dos "novos governantes de discursos" tenham particular importância para uma análise e compreensão sobre o estado da arte da liberdade de expressão.

As grandes redes sociais agudizam alguns dilemas a respeito da permissão ou proibição de discursos, especialmente por conta da escala em que operam. Essas empresas quase sempre se encontram em meio ao fogo cruzado de pressões ora pela limitação a discursos, ora pela liberação de discursos. Não raro, essas demandas contraditórias agem ambas sobre um mesmo discurso ou problema. É possível ler os diversos casos apresentados neste trabalho sempre por meio dessas duas lentes: a pressão pela não restrição, motivada por um fundado receio de uma censura privada abrangente por parte dessas empresas poderosas, ou a pressão por uma maior limitação de conteúdos considerados problemáticos, motivada pelos novos tipos de riscos e danos que surgem a partir desses ambientes digitais.

Nenhuma dessas lentes é suficiente por si só para atacar o problema jurídico da remoção de conteúdos por decisões autônomas das grandes redes sociais. Se ao longo do

trabalho a leitura de seu texto explicitou essas ambiguidades, é porque se tem a convicção de que a pesquisa revela de que maneiras elas são inerentes ao objeto de estudo. Não parecer ser possível ser simplesmente contra ou a favor da atividade de moderação de conteúdo pelas redes sociais. Em meio a esses dois extremos, é necessário encontrar um caminho constitucionalista que preserve a incidência de direitos fundamentais, notadamente a liberdade de expressão, e ao mesmo tempo apresente alternativas realistas para os problemas de fato que surgem nesses mercados digitais de ideias, frequentados globalmente por multidões de milhões de pessoas. A tese buscou contribuir com esse objetivo, especialmente em seu Capítulo 4.

Por fim, esta pesquisa manteve desde seu início sua pretensão de apresentar uma análise mais abrangente sobre a moderação de conteúdo feita pelas grandes redes sociais, de modo a motivar a construção de conceitos e argumentos normativos na seara do direito constitucional. A despeito desse caráter mais generalista, este tema constitui campo que merece continuar sendo objeto de pesquisas, inclusive jurídicas. Enquanto a governança privada de discursos por empresas de tecnologia continuar operando espaços de debates públicos relevantes, essa será uma discussão em andamento, sem um ponto fixo de chegada.

Como já demonstra a bibliografia deste trabalho, há espaço para investigação, análise e crítica de como essa governança opera em diversas áreas: regulação de discursos eleitorais, curadoria algorítmica e formação de bolhas opinativas, efeitos discursivos das arquiteturas de publicações e desafios singulares que são postos em áreas específicas, como questões relativas a discursos de ódio ou campanhas de desinformação, entre outros.

Como se espera que tenha ficado claro, propostas de enfrentamento a todos esses problemas devem passar também pelas capacidades singulares das empresas de tecnologia que detêm as chaves de seus próprios ambientes – o que demanda, por parte das normas de direito público, regulações que sejam inteligentes e eficientes, cientes desse necessário diálogo com o "direito das plataformas" e zelando para que esse campo também opere com um grau satisfatório de respeito e deferência para com direitos fundamentais.

Bibliografia

Antonialli, Dennys. *A arquitetura da Internet e o desafio da tutela do direito à privacidade pelos Estados Nacionais*. Tese de doutorado apresentada à Faculdade de Direito da Universidade de São Paulo, 2017.

_____. "Drag Queen vs. David Duke: whose tweets are more 'toxic'?" *Wired*, artigo publicado em 25/07/2019.

Balkin, Jack. "Old-School/New-School Speech Regulation", *Harvard Law Review Volume 127* (2014): 2296-2342.

_____. "Free Speech is a triangle", *Columbia Law Review Volume 118* (2018): 2011-2055.

Bezanson, Randall. "The developing law of editorial judgment", *Nebraska Law Review Volume 78* (1999): 754-857.

Bowers, John; Zittrain, Jonathan. "Answering impossible questions: content governance in an age of disinformation", *The Harvard Kennedy School (HKS) Misinformation Review,* artigo publicado *em 14/01/20*, acesso em https://doi.org/10.37016/mr-2020-005

Blum, Renato Opice; Elias, Paulo Sá; Monteiro, Renato Leite. "Marco regulatório da internet brasileira: 'Marco Civil'", *Migalhas*, artigo publicado em 20/06/2012

Brito Cruz, Francisco Carvalho de. *Direito, democracia e cultura digital: a experiência de elaboração legislativa do Marco Civil da Internet*. Dissertação de mestrado apresentada à Faculdade de Direito da Universidade de São Paulo, 2015.

Brito Cruz, Francisco Carvalho de; Massaro, Heloísa; Oliva, Thiago; Borges, Ester. "Internet e eleições no Brasil: diagnósticos e recomendações", relatório publicado pelo centro de pesquisas *Internetlab*, em 26/09/2019, acesso em www.internetlab.org.br .

Celeste, Edoardo. "Digital Constitutionalism: mapping the constitutional responses to digital techonology´s challenges", *HIIG Discussion Paper Series No. 2018-02 (2018)*, acesso em https://ssrn.com/abstract=3219905

Citron, Danielle Keats; Norton, Helen. "Intermediaries and Hate Speech: fostering digital citizenship for our information age", *Boston University Law Review Volume 91* (2011): 1435-1484.

Citron, Danielle Keats; Wittes, Benjamin. "The Internet will not break: denying bad samaritans Section 230 immunity", *Fordham Law Review Volume 86* (2017): 401-423.

Cook, Timothy. "Introductory Essay", *in*: Timothy Cook (org.), *Freeing the presses: the First Amendment in action*. Louisiana State University Press, 2006.

Coutinho, Diogo R.; Kira, Beatriz. "Por que (e como) regular algoritmos?", *portal Jota*, artigo publicado em 02/05/2019.

Cram, Ian. "The Danish cartoons, offensive expression and democratic legitimacy", *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010.

Derakhshan, Hossein. "The Web we have to save", *Matter*, artigo publicado em 14/07/2015.

Docquir, Pierre François. "The Social Media Council: bringing human rights standards to content moderation on social media", *Centre of International Governance Innovation*, artigo publicado em 28/10/2019, acesso em http://cigionline.org

Douek, Evelyn. "Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation", *Hoover Working Group on National Security, Technology, and Law – Aegis Series Paper No. 1903* (2019), acesso disponível em http://www.hoover.org

_____. "Facebook´s 'Oversight Board': move fast with stable infrastructure and humility", *North Carolina Journal of Law and Technology Volume 21* (2019): 1-77.

_____. "Facebook´s Oversight Board Bylaws: for once, moving slowly", *Lawfare Blog*, artigo publicado em 28/01/20.

_____. "The rise of content cartels", *Knight First Amendment Institute – Columbia University*, artigo publicado em 11/02/20, acesso em http://knightcolumbia.org

Duan, Charles; Westling, Jeffrey. "Will Trump's executive order harm online speech? It already did.", *Lawfare Blog*, artigo publicado em 01/06/20.

Dworkin, Ronald. *O direito da liberdade: a leitura moral da constituição norte-americana*. Martins Fontes, 2006.

Fiss, Owen. *A ironia da Liberdade de Expressão*. Editora Renovar, 2005.

Garton Ash, Timothy. *Free Speech: ten principles for a connected world*. Yale University Press, 2016.

Gesley, Jenny. "Germany: Facebook found in violation of 'anti-fake news' law", *Global Legal Monitor – Library of Congress of the United States*, artigo publicado em 20/08/2019, acesso em http://loc.gov/law

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

Goldsmith, Jack. "The failure of Internet Freedom", *Knight First Amendment Institute – Columbia University,* artigo publicado em 13/06/2018, acesso em http://knightcolumbia.org

Gomes, Alessandra; Antonialli, Dennys; Oliva, Thiago Dias. "Drag queens e inteligência artificial: computadores devem decidir o que é 'tóxico' na internet?", *Internetlab*, artigo publicado em 28/06/2019, acesso em www.internetlab.org.br .

Gross, Clarissa Piterman. *Pode dizer ou não? Discurso do ódio, liberdade de expressão e a democracia liberal igualitária*. Tese de doutorado apresentada à Faculdade de Direito da Universidade de São Paulo, 2017.

Horwitz, Paul. *First Amendment Institutions*. Harvard University Press, 2012.

Jeong, Sarah. "Politicans want to change the internet´s most important law. They should read it first", *The New York Times*, artigo publicado em 26/07/2019.

Kadri, Thomas; Klonick, Kate. "Facebook v. Sullivan: public figures and newsworthiness in online speech", *Southern California Law Review Volume 93* (2019): 37-99.
_____. "How to make Facebook's Supreme Court work", *The New York Times*, artigo publicado em 17/11/2018.

Kaiser, Jonas; Rauchfleisch, Adrian. "Unite the right? How Youtube´s recommendation algorithm connects the U.S. far right", *Medium*, artigo publicado em 11/04/2018.
_____. "The implications of venturing down the rabbit hole", *Internet Policy Review: Journal of Internet Regulation,* artigo publicado em 27/06/2019.

Kaye, David. *Speech Police: the global struggle to govern the Internet*. Columbia Global Reports, 2019.

Kettemann, Mathias C; Schulz, Wolfgang. "Setting rules for 2.7 billion: a (first) look into Facebook's norm-making system – results of a pilot study", *Working Papers of the Hans-Bredow-Institut/ Leibniz Institute for Media Research (Hamburg) – Works in Progress #1* (2020), acesso disponível em http://www.hans-bredow-institut.de

Klonick, Kate. "The New Governors: the people, rules, and processes governing online speech", *Harvard Law Review Volume 131* (2018): 1598-1670.

_____. “Inside the team at Facebook that dealt with the Christchurch shooting”, *The New Yorker*, artigo publicado em 25/04/2019.

Kossef, Jeff. *The Twenty-Six Words That Created the Internet*. Cornell University Press, 2019, edição Kindle.

La Chapelle, Bertrand de; Fehlinger, Paul. “Jurisdiction on the Internet: from legal arms race to transnational cooperation”, *Global Comission on Internet Governance Paper Series nº 28* (2016), acesso em http://www.cigionline.org/

Lessig, Lawrence. *Code: version 2.0*. Basic Books, 2006.

Lewis, Anthony. *Make No Law: The Sullivan case and the First Amendment.* Vintage Books, 1991.

Macedo Júnior, Ronaldo Porto. “Freedom of Expression: what lessons should we learn from US experience?”, *Revista Direito GV Volume 13, n.1, São Paulo (2017)*: 274-302.
_____. “Fake News: a novidade de dizer mentiras”, *Revista de Jornalismo ESPM*, *edição julho-dezembro 2018*.

Mbongo, Pascal. “Hate Speech, Extreme Speech, and Collective Defamation in French Law”, *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010.

McNamee, Roger. *Zucked: waking up to the Facebook catastrophe*. Harper Collins, 2019, edição Kindle.

Meiklejohn, Alexander. *Political Freedom: the constitutional powers of the people*. Greenwood Press, 1979.

Moncau, Luiz Fernando Marrey. “Intermediários de Internet e Liberdade de Expressão: o mapa da busca de um delicado equilíbrio regulatório”, portal *Dissenso.org*, artigo publicado em 06/06/2018.

_____. *Direito ao esquecimento: entre a liberdade de expressão, a privacidade e a proteção de dados pessoais*. Editora RT, 2020.

Moncau, Luiz Fernando Marrey; Arguelles, Diego Werneck. “The Marco Civil da Internet and Digital Constitutionalism”: *in* Giacarlo Frosio (org.), *The Oxford Handbook of Online Intermediary Liability*. Oxford University Press, 2020.

Nitrini, Rodrigo Vidal. *Liberdade de informação e proteção ao sigilo de fonte: desafios constitucionais na era da informação digital*. Hucitec Editora, 2016.

Ortellado, Pablo; Ribeiro, Márcio Moretto. “O que são e como lidar com as notícias falsas”, *Sur – Revista Internacional de Direitos Humanos Volume 15, nº 27 (julho 2018).*

Post, Robert. "“Hate Speech”, *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010.

_____. "Participatory Democracy and Free Speech", *Virginia Law Review Volume 97, n. 3* (2011): 477-489.

_____. “Free Speech in the age of Youtube”, *Foreign Policy*, artigo publicado em 17/09/2012.

Redish, Martin. *The Adversary First Amendment: free expression and the foundations of American Democracy*. Stanford Law Books, 2013.

Roberts, Sarah T. *Behind the screen: content moderation in the shadows of social media*, Yale University Press, 2019.

Roose, Kevin. “A mass murderer of, and for, the Internet”, *The New York Times*, artigo publicado em 15/03/2019.

Rosen, Jeffrey. “The delete squad”, *The New Republic*, artigo publicado em 29/04/2013.

Sap, Marteen; Card, Dallas; Gabriel, Saadia; Choi, Yejin; Smith, Noah A. “The risk of racial bias in hate speech detection”, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019)*, pp.1668–1678; acesso direto em “Google’s algorithm for detecting hate speech is racially biased”, artigo publicado por *MIT Technology Review*, em 13/08/2019.

Schauer, Frederick. “The Exceptional First Amendment,”, *in:* Michael Ignatieff (org.), *American Exceptionalism and Human Rights*. Princeton University Press, 2005.
_____. “Towards and Institutional First Amendment, *Minnesota Law Review Volume 89* (2005): 1256-1279.

Silva, Virgílio Afonso da. *A constitucionalização do direito: os direitos fundamentais nas relações entre particulares*. Malheiros, 2005.
_____. “Colisões de direitos fundamentais entre ordem nacional e ordem transnacional”, *in:* Marcelo Neves (org.), *Transnacionalidade do direito: novas perspectivas dos conflitos entre lógicas jurídicas*. Quartier Latin, 2010.

Souza, Carlos Affonso; Lemos, Ronaldo. *Marco Civil da Internet: construção e aplicação*. Editar Editora, 2016.

Souza, Carlos Affonso; Teffé, Chiara Spadaccini de. “Responsabilidade dos provedores por conteúdos de terceiros na internet”, *Consultor Jurídico*, artigo publicado em 23/01/2017.

Sunstein, Cass. *Democracy and the problem of Free Speech*. The Free Press, 1995.
_____. *#Republic: divided democracy in the age of social media*. Princeton University Press, 2017.

_____. "Facebook can fight lies in political ads", *Bloomberg.com*, artigo publicado em 09/10/2019.

Summer, L. W. "Incitement and the Regulation of Hate Speech in Canada: a Philosophical Analysis", *in*: Ivan Hare e James Weinstein (org.), *Extreme Speech and Democracy*. Oxford University Press, 2010.

Suzor, Nicolas P. *Lawless: the secret rules that govern our digital lives*. Cambridge University Press, 2019.

Teuber, Gunther. *Constitutional Fragments: societal constitutionalism and globalization.* Oxford University Press, 2012.

Tushnet, Rebecca. "Power without responsability: intermediaries and the First Amendment", *George Washington Law Review Volume 76* (2008): 986-1016.

Tworek, Heidi; Leerssen, Paddy. "An analysis of Germany's NetzDG law", *Transatlanic High Level Working Group on Content Moderation Online and Freedom of Expression – Institute for Information Law (Universiteit Van Amesterdam),* abril de 2019, acesso em http://ivir.nl

Vaidhyanathan, Siva. "The Real Reason Facebook Won't Fact-Check Political Ads", *The New York Times*, artigo publicado em 02/02/2019.

Whitney, Heather. "Search engines, social media, and the editorial analogy", *Knight First Amendment Institute – Columbia University,* artigo publicado em 27/02/2018, acesso em http://knightcolumbia.org

Zittrain, Jonathan. *The future of the internet: and how to stop it*. Penguin, 2008, edição Kindle.

_____. “The hidden costs of automated thinking”, *The New Yorker*, artigo publicado em 23/07/19.
_____. “A jury of random people can do wonders for Facebook”, *The Atlantic*, artigo publicado em 14/11/2019.

Zuboff, Soshana. *The Age of Surveillance Capitalism: the fight for a human future at the new frontier of power*. Public Affairs, 2019.

Zuckerberg, Mark. “The internet needs new rules. Let´s start in these four areas”, *The Washington Post*, artigo publicado em 30/03/19.

Wu, Tim. “Is the First Amendment Obsolete?”, *Knight First Amendment Institute – Columbia University,* artigo publicado em 01/09/2017, acesso em http://knightcolumbia.org .

ANEXO 1 – Imagens de manual de treinamento interno distribuído pelo Facebook a seus moderadores, datado de 2016 e publicado pelo jornal *The Guardian* em 2017[365]:

Fig. 1:

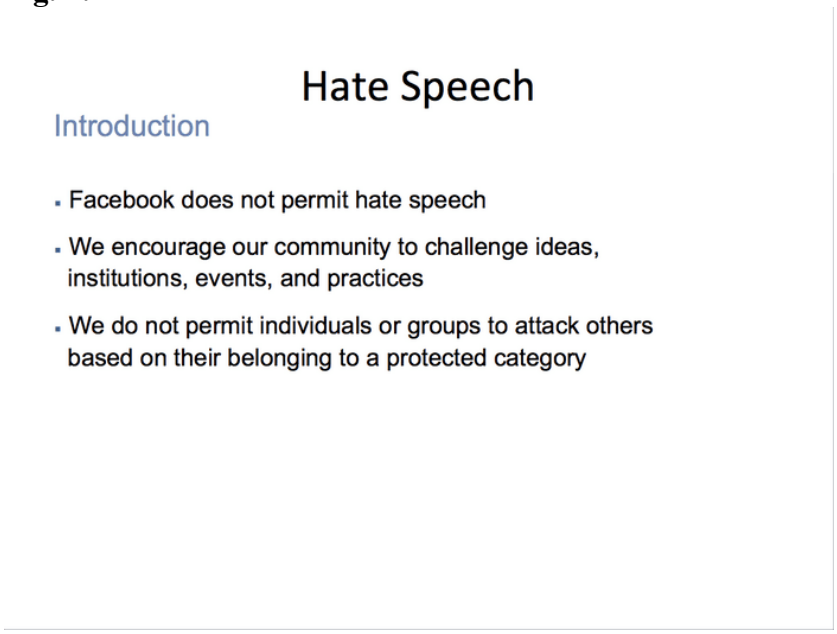


Fig. 2:

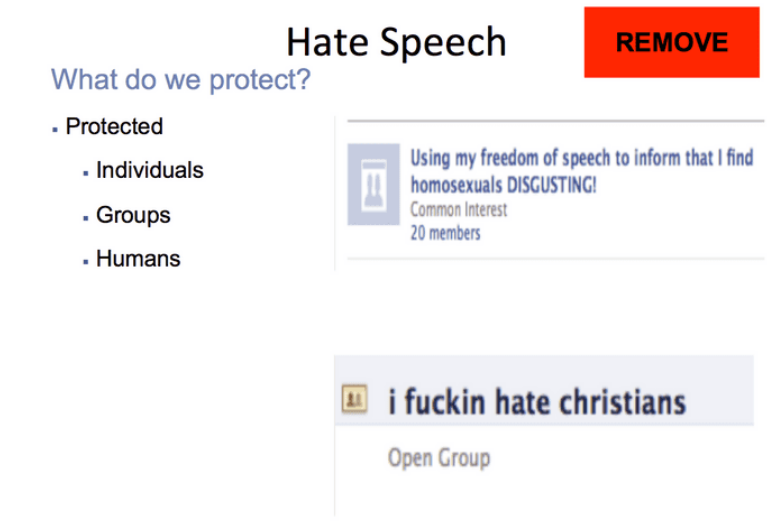


Fig. 3:

[365] Imagens conforme originais, em inglês. Acesso em 23/10/2019: https://www.theguardian.com/news/gallery/2017/may/24/hate-speech-and-anti-migrant-posts-facebooks-rules


Hate Speech
ALLOWED
What do we NOT protect?
Not protected
Concepts
Institutions
Beliefs
Anti-homosexuality
Common interest
15 members
I Hate Christianity
Request to Join
Info
Basic Info

Fig. 4:


Protected categories
Religious affilitation
We protect the followers of a religion. Not the religion itself
Christians ≠ Christianity
Bhuddists ≠ Bhuddism
Examples:
Catholics, Protestants, Muslims, Sunni, Shia
Scientologists
Mormons
Jehovah witnesses
Satanists
Atheists
etc…

Fig. 5:

Protected categories

Sexual orientation

- Heterosexual
- Homosexual
- Bisexual
- Asexual

Fig. 6:

Not Protected categories

Social class

- Rich
- Poor
- Middle class
- Working class
- Etc…

Fig. 7:

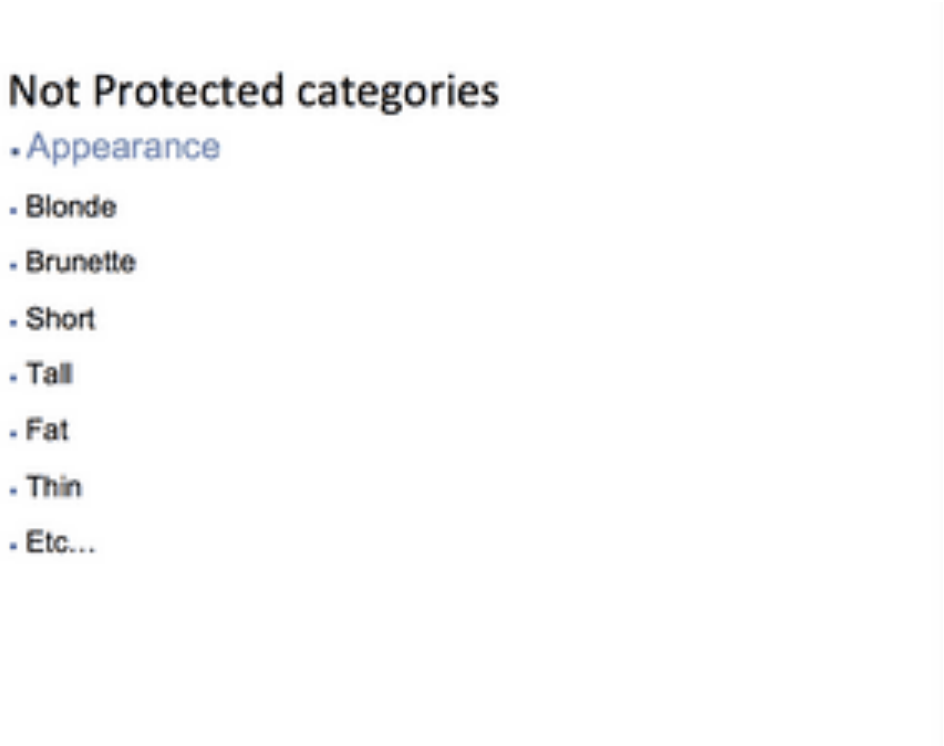


Fig. 8:

Fig. 9:

Not Protected categories

Countries

- Countries are not protected. People from a country are protected
- Ireland
- England
- France
- USA
- Brazil
- Spain
- Etc…

Fig. 10:

Quasi Protected Category (QPC)

People who cross an international border with intent to establish residency in a new country, regardless of whether their motivation is economic or political (defined as: migrants, refugees, immigrants, asylum seekers)

Fig. 11:

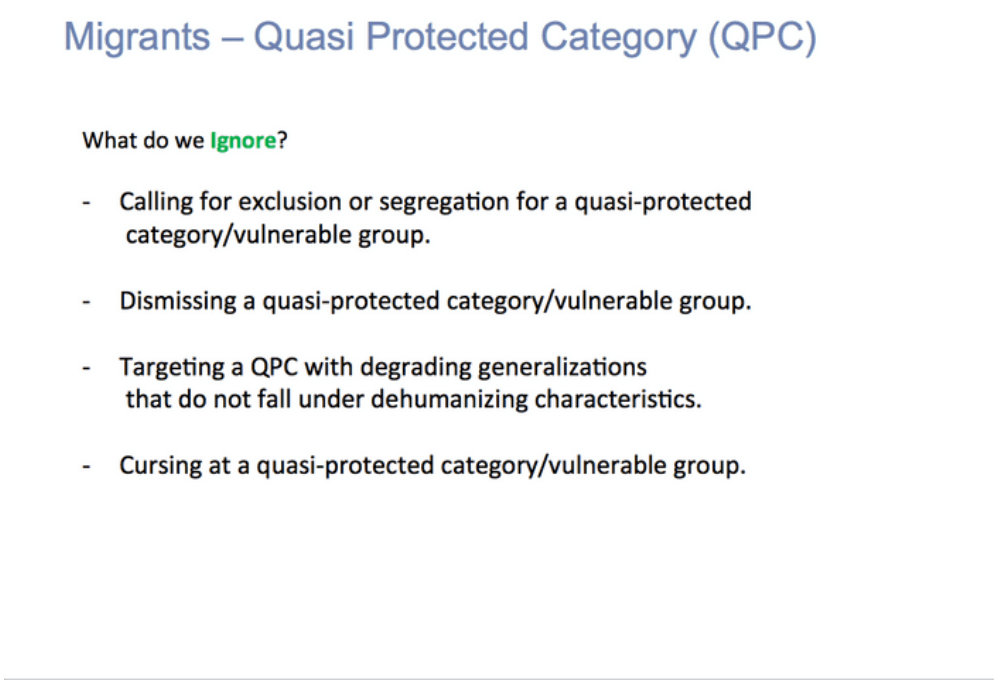

Migrants – Quasi Protected Category (QPC)
What do we Ignore?
- Calling for exclusion or segregation for a quasi-protected category/vulnerable group.
- Dismissing a quasi-protected category/vulnerable group.
- Targeting a QPC with degrading generalizations that do not fall under dehumanizing characteristics.
- Cursing at a quasi-protected category/vulnerable group.

Fig. 12:

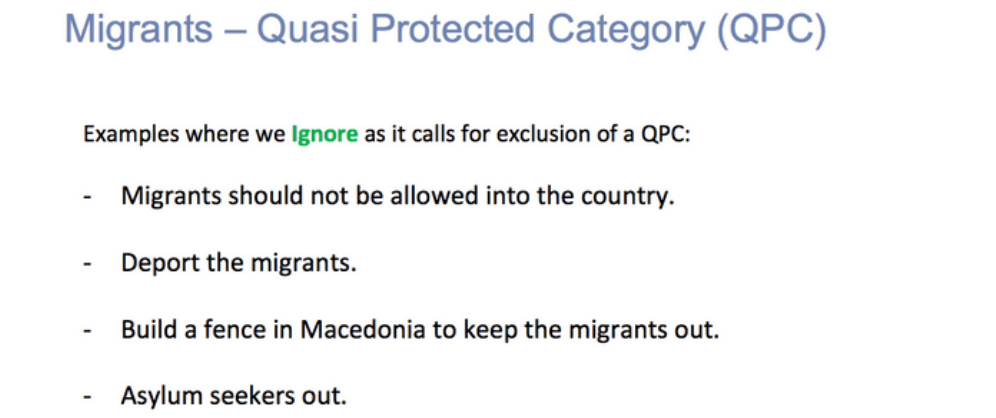

Migrants – Quasi Protected Category (QPC)
Examples where we Ignore as it calls for exclusion of a QPC:
- Migrants should not be allowed into the country.
- Deport the migrants.
- Build a fence in Macedonia to keep the migrants out.
- Asylum seekers out.

Fig. 13:

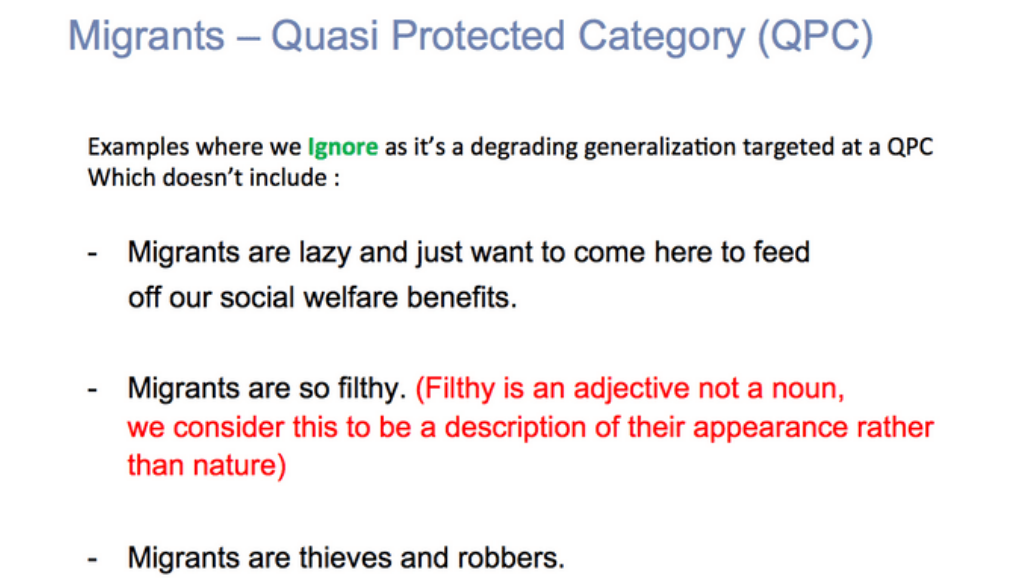

Migrants – Quasi Protected Category (QPC)
Examples where we Ignore as it's a degrading generalization targeted at a QPC
Which doesn't include :
- Migrants are lazy and just want to come here to feed off our social welfare benefits.
- Migrants are so filthy. (Filthy is an adjective not a noun, we consider this to be a description of their appearance rather than nature)
- Migrants are thieves and robbers.

Fig. 14:

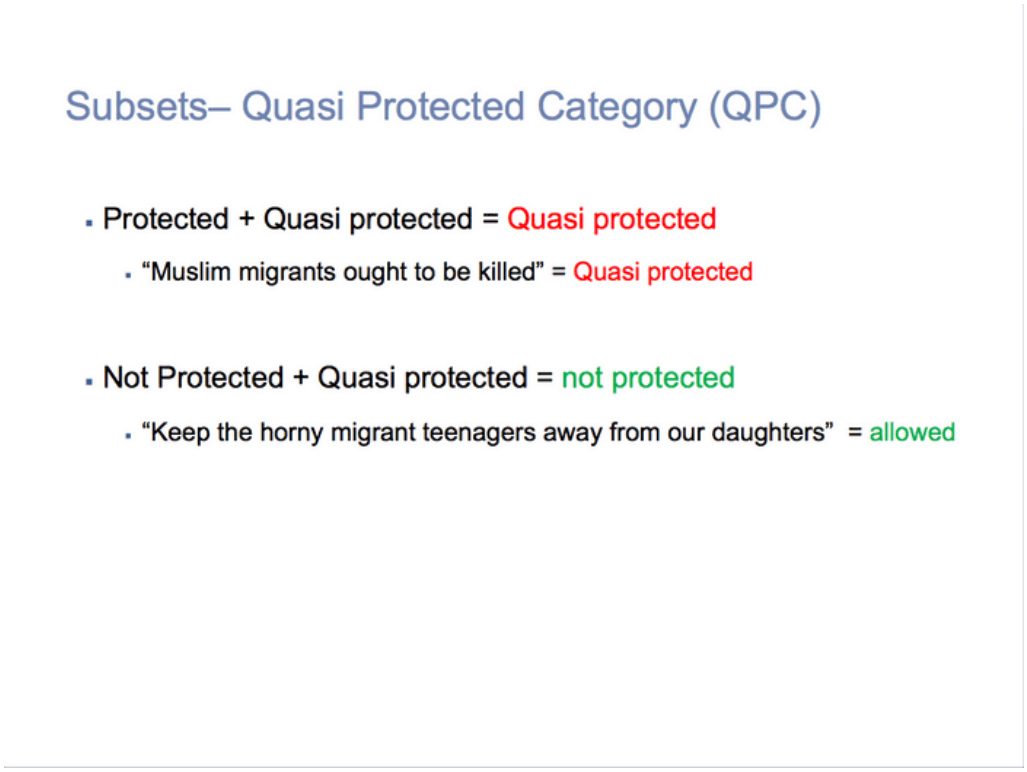

Subsets– Quasi Protected Category (QPC)
Protected + Quasi protected = Quasi protected
"Muslim migrants ought to be killed" = Quasi protected
Not Protected + Quasi protected = not protected
"Keep the horny migrant teenagers away from our daughters" = allowed

Fig. 15:

Hate Speech

Referencing protected categories

- Allowed
  - Calling someone as a PC ('You are such a Jew')
  - Identifying someone as a PC ('He's gay')
  - Claiming superiority ('French are the best')

- Not allowed
  - Claiming superiority if they mention another PC as inferiors
  - « Irish are the best, but really French sucks »

Fig. 16:

Hate Speech

Scenarios

- Dispute of historical events or hate crimes = allowed
  - 9/11 did not happen
  - Holocaust Denial: IP-Blocked
- Right-wing political parties = allowed
- Anti-immigration stances = allowed

Fig. 17:
These are examples of denigration speech Facebook alows

S)

Fig. 18:

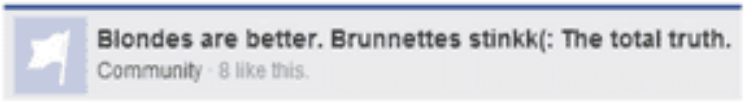


Fig. 19:
In separate notes, Facebook tells moderators to ignore this message and caption because it says "filthy" is not the same as "filth" and "calling migrants" thieves is not violating

Fig. 20:

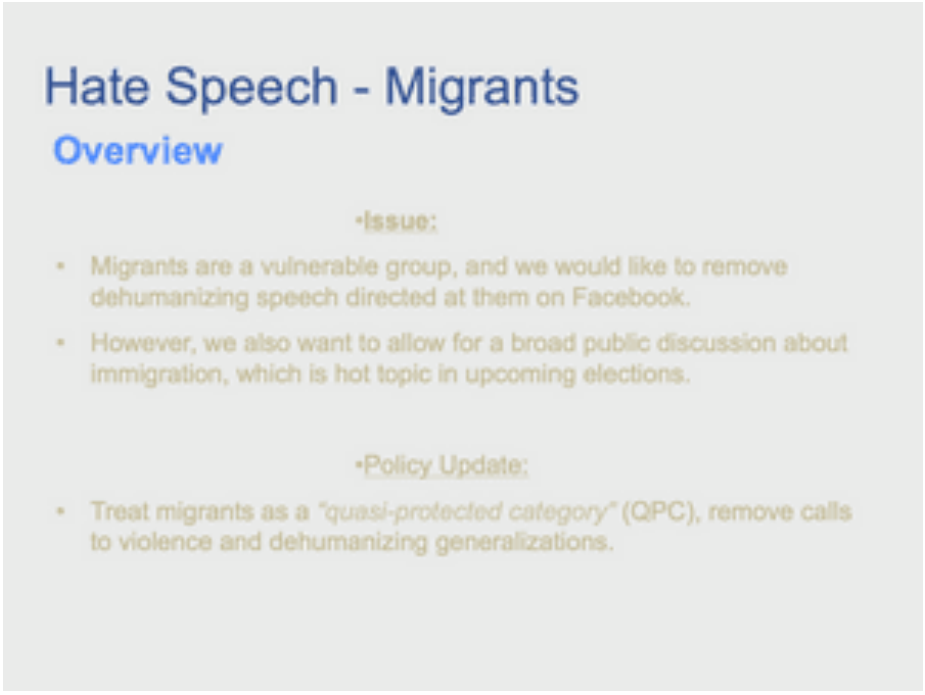


Fig. 21:

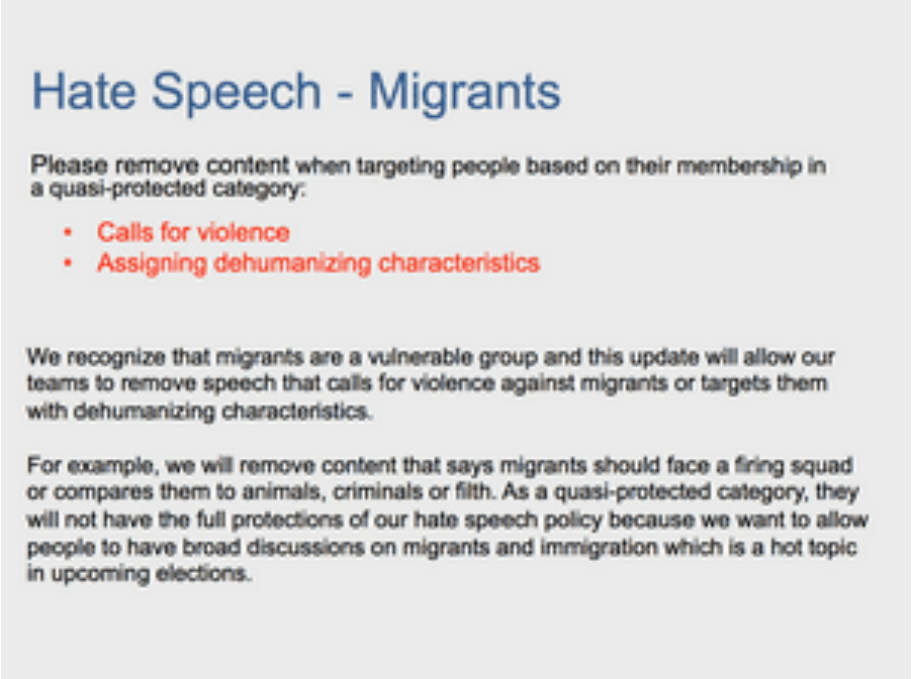

Hate Speech - Migrants
Please remove content when targeting people based on their membership in a quasi-protected category:
• Calls for violence
• Assigning dehumanizing characteristics
We recognize that migrants are a vulnerable group and this update will allow our teams to remove speech that calls for violence against migrants or targets them with dehumanizing characteristics.
For example, we will remove content that says migrants should face a firing squad or compares them to animals, criminals or filth. As a quasi-protected category, they will not have the full protections of our hate speech policy because we want to allow people to have broad discussions on migrants and immigration which is a hot topic in upcoming elections.

Fig. 22:


Hate Speech - Migrants
Examples: (DELETE)
Dehumanizing characteristics – REMOVE
• Migrants are scum.
• Migrants are filthy cockroaches that will infect our country.
• The migrant rats have arrived in Berlin.
• Refugees? They're all rape-fugees!
• Refugees are state-financed child molesters.
EDGE CASE – "Dismissing" an entire QPC should be an IGNORE
• Migrants are lazy and just want to come here to feed off our social welfare benefits.
• Migrants are so filthy.
• Migrants are thieves and robbers.

Fig. 23:

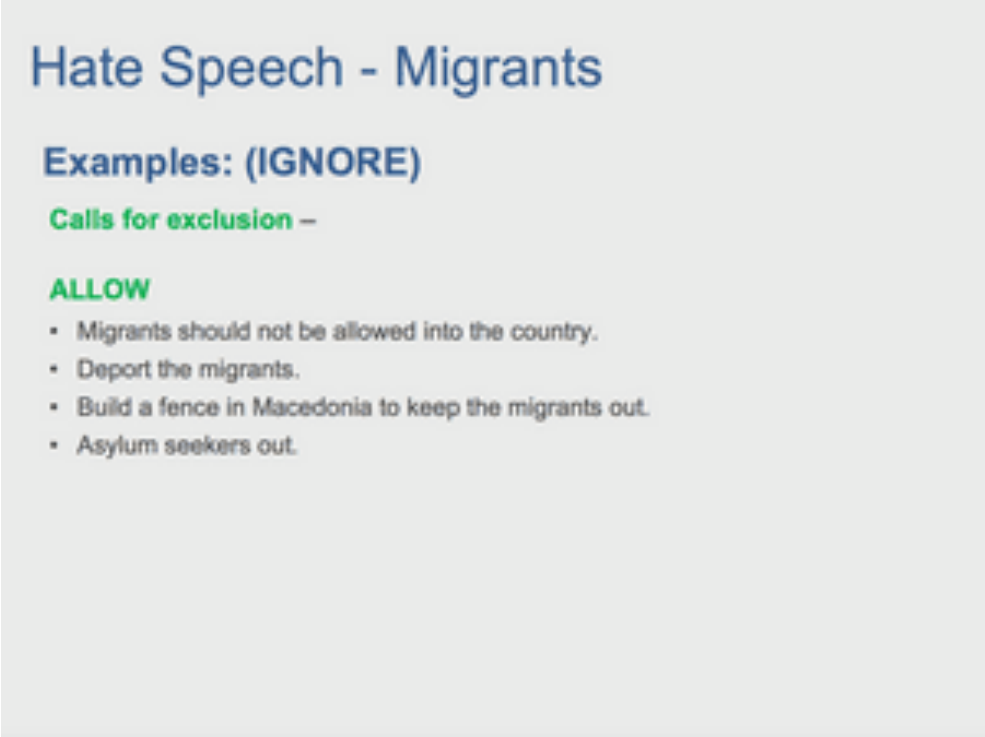

Hate Speech - Migrants
Examples: (IGNORE)
Calls for exclusion –
ALLOW
• Migrants should not be allowed into the country.
• Deport the migrants.
• Build a fence in Macedonia to keep the migrants out.
• Asylum seekers out.

Fig. 24:

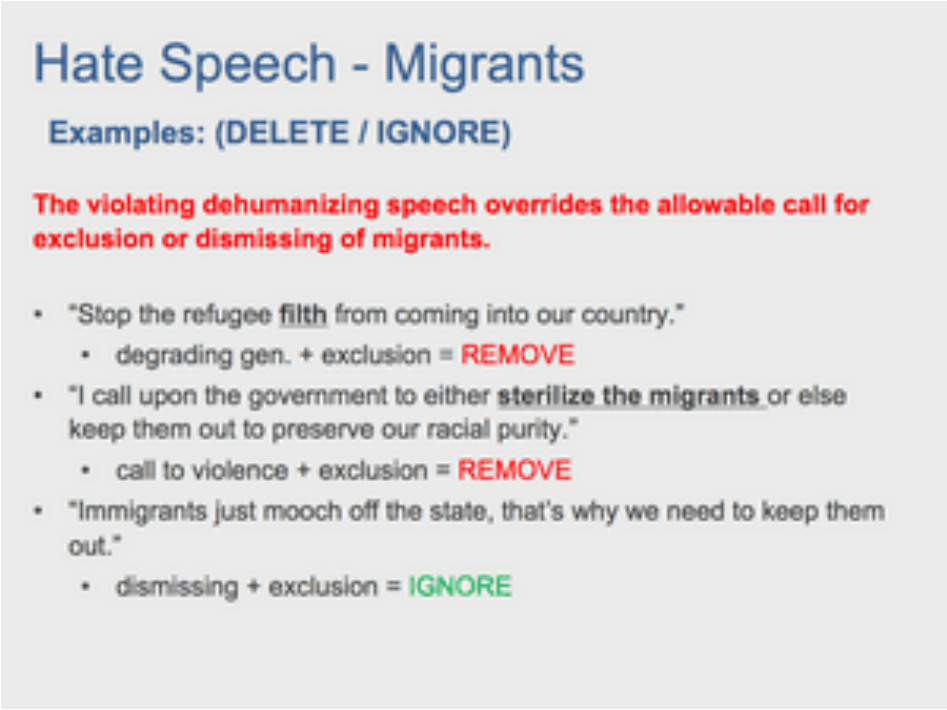
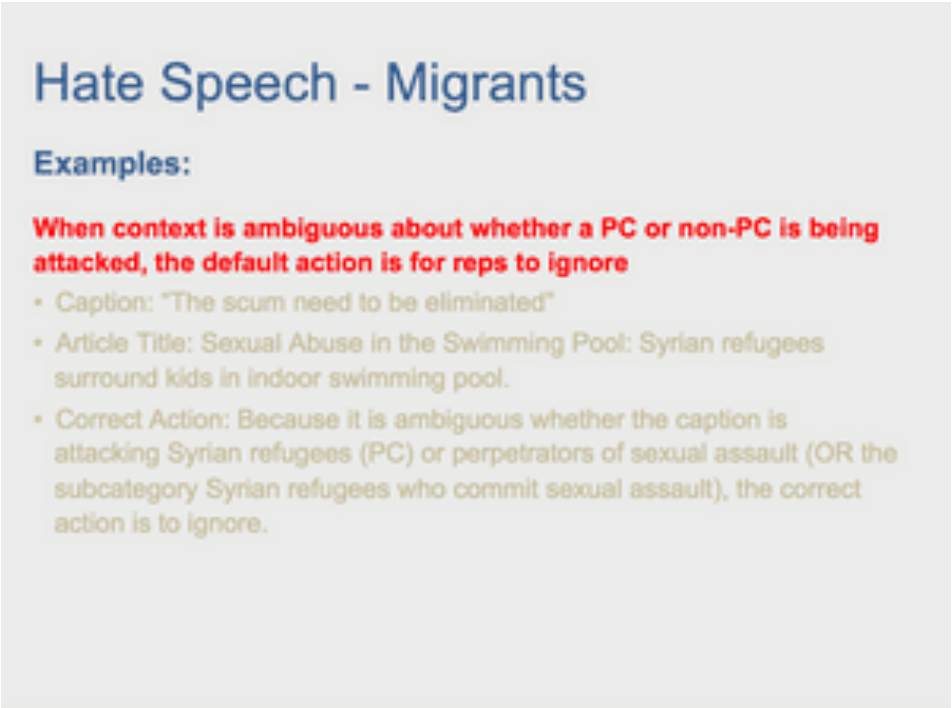

Hate Speech - Migrants
Examples: (DELETE / IGNORE)
The violating dehumanizing speech overrides the allowable call for exclusion or dismissing of migrants.
• "Stop the refugee filth from coming into our country."
• degrading gen. + exclusion = REMOVE
• "I call upon the government to either sterilize the migrants or else keep them out to preserve our racial purity."
• call to violence + exclusion = REMOVE
• "Immigrants just mooch off the state, that's why we need to keep them out."
• dismissing + exclusion = IGNORE

Fig. 25:


Hate Speech - Migrants
Examples: (DELETE/ IGNORE)
We will remove degrading generalizations or calls to violence against migrant subgroups of a PC.
• "I call upon the government to mandate all gay immigrants have a chip implanted in their brain so law enforcement can keep track of them." = REMOVE (call to violence)
• "The Sikhs who come to this country are filthy cows." = REMOVE (dehumanizing)
• "Islam is a religion of hate. Close the borders to immigrating Muslims until we figure out what the hell is going on." = IGNORE (exclusion)
• "Mexican immigrants are freeloaders mooching off of tax dollars we don't even have." = IGNORE (dismissing)

Fig. 26:


Hate Speech - Migrants
Examples:
When context is ambiguous about whether a PC or non-PC is being attacked, the default action is for reps to ignore
• Caption: "The scum need to be eliminated"
• Article Title: Sexual Abuse in the Swimming Pool: Syrian refugees surround kids in indoor swimming pool.
• Correct Action: Because it is ambiguous whether the caption is attacking Syrian refugees (PC) or perpetrators of sexual assault (OR the subcategory Syrian refugees who commit sexual assault), the correct action is to ignore.