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

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.

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:

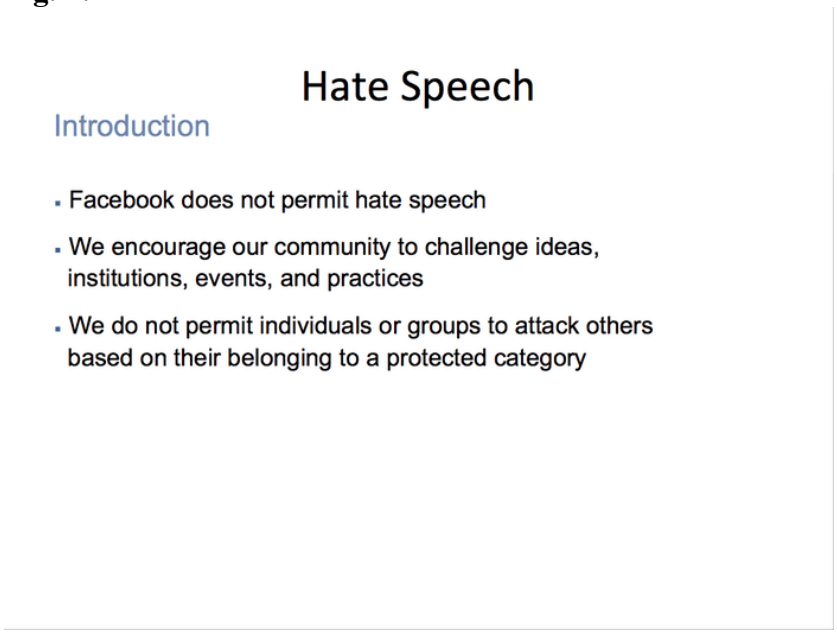


Fig. 2:

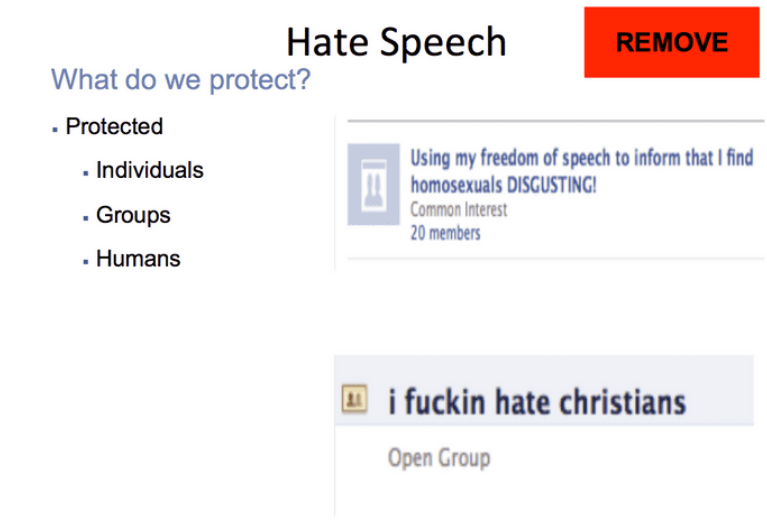


Fig. 3:

[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:

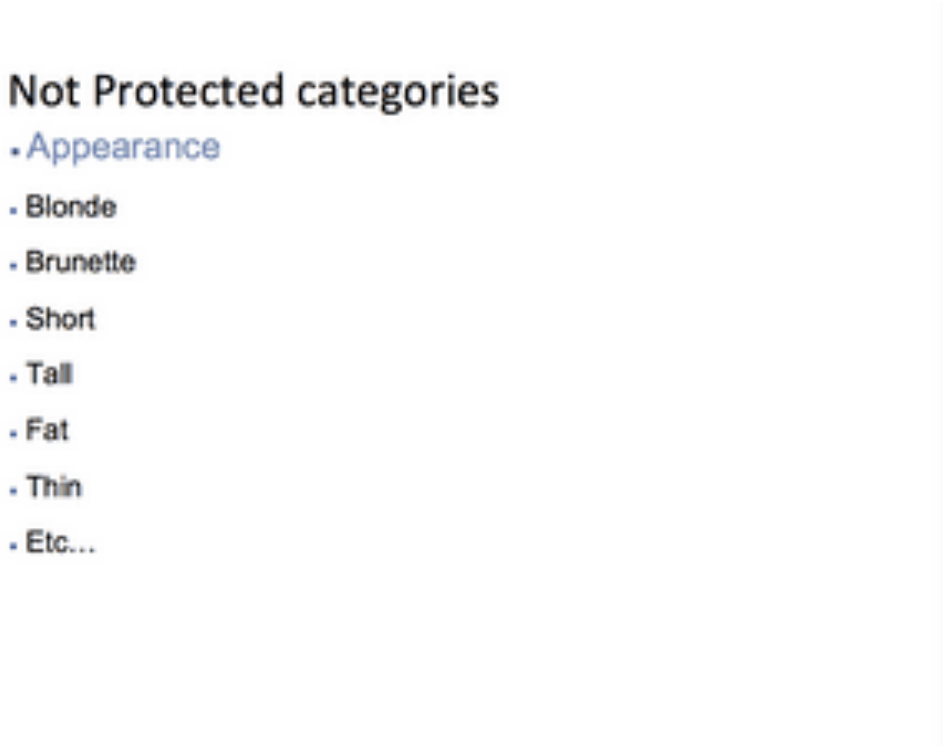


Fig. 8:

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:

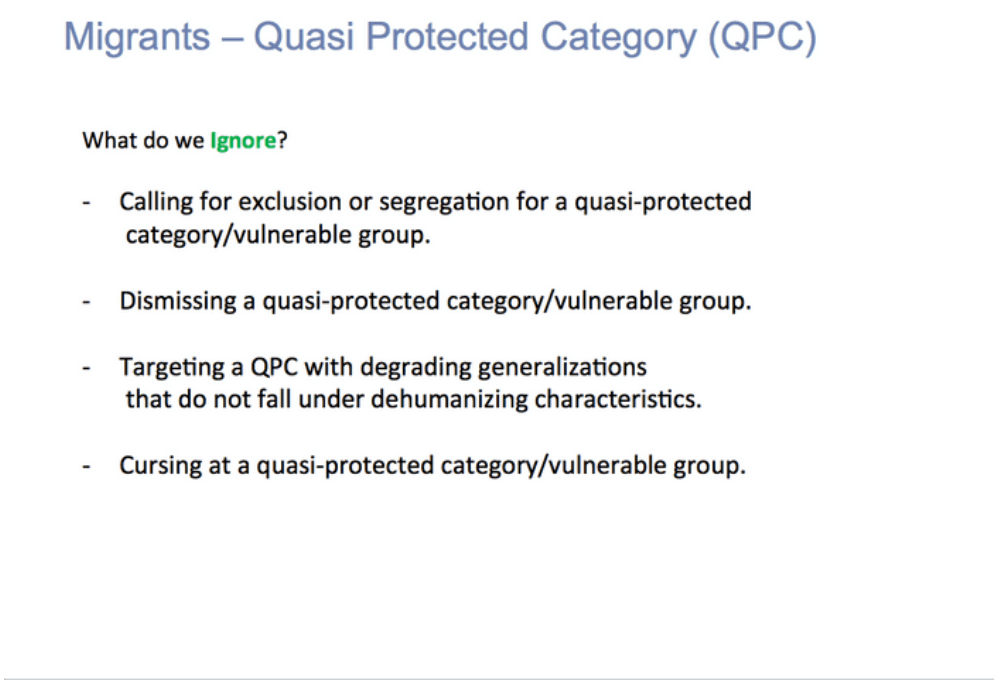

Migrants – Quasi Protected Category (QPC)
What do we Ignore?
- Calling for exclusion or segregation for a quasi-protected category/vulnerable group.
- Dismissing a quasi-protected category/vulnerable group.
- Targeting a QPC with degrading generalizations that do not fall under dehumanizing characteristics.
- Cursing at a quasi-protected category/vulnerable group.

Fig. 12:

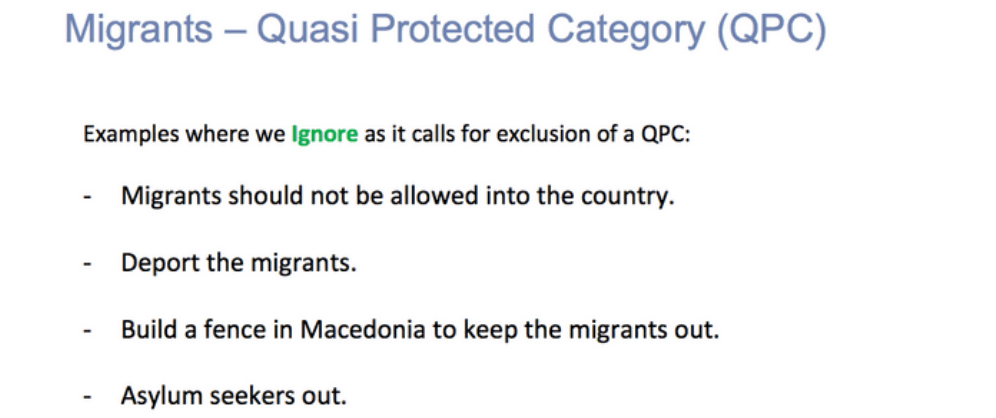

Migrants – Quasi Protected Category (QPC)
Examples where we Ignore as it calls for exclusion of a QPC:
- Migrants should not be allowed into the country.
- Deport the migrants.
- Build a fence in Macedonia to keep the migrants out.
- Asylum seekers out.

Fig. 13:

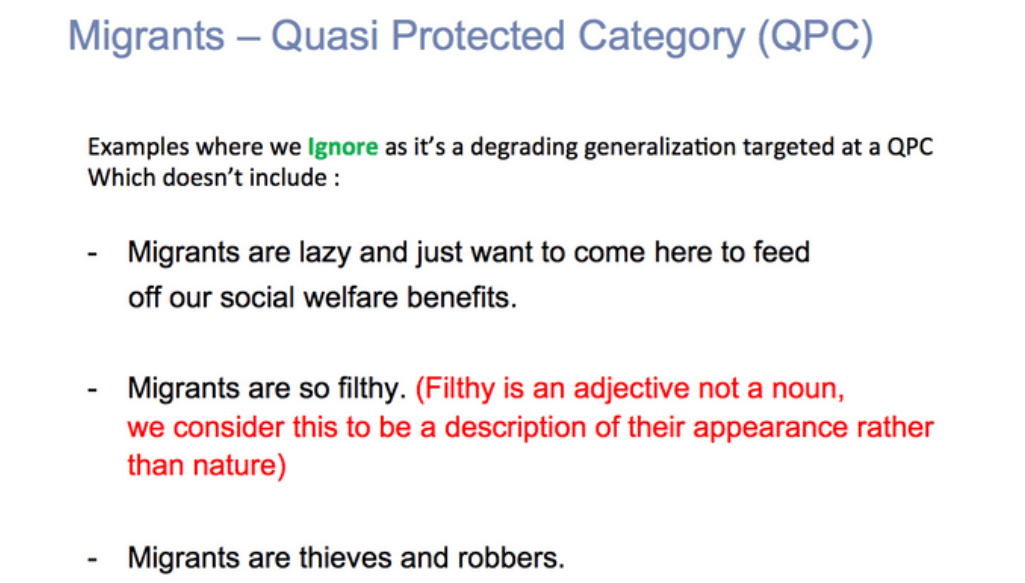

Migrants – Quasi Protected Category (QPC)
Examples where we Ignore as it's a degrading generalization targeted at a QPC
Which doesn't include :
- Migrants are lazy and just want to come here to feed off our social welfare benefits.
- Migrants are so filthy. (Filthy is an adjective not a noun, we consider this to be a description of their appearance rather than nature)
- Migrants are thieves and robbers.

Fig. 14:

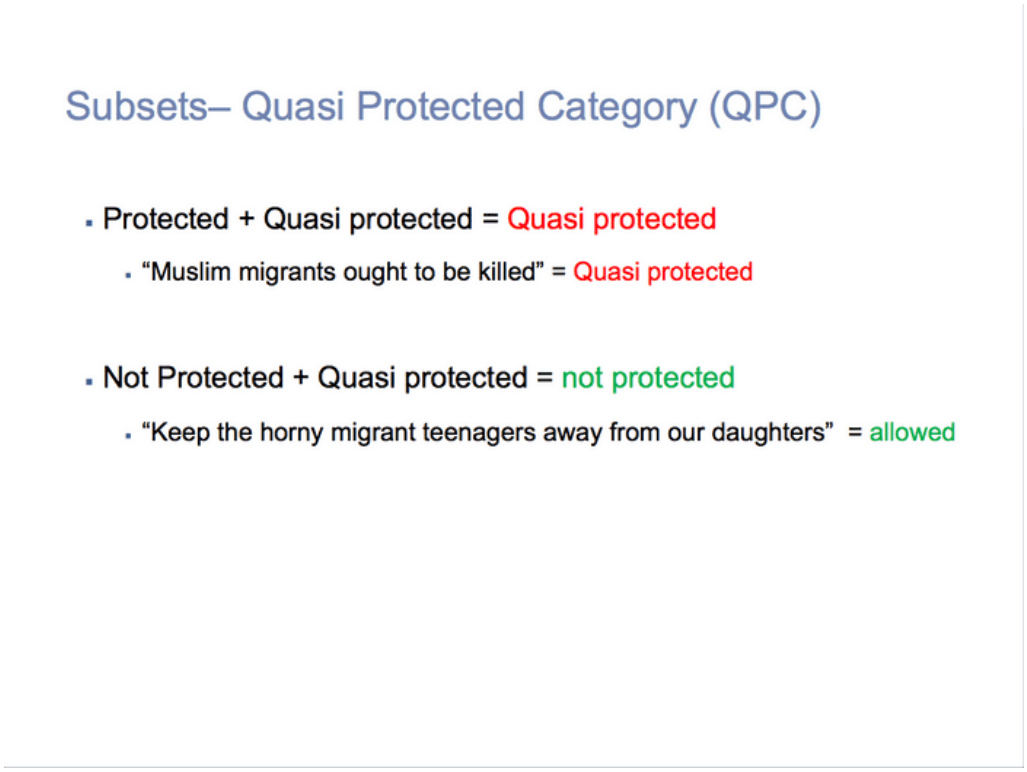

Subsets– Quasi Protected Category (QPC)
Protected + Quasi protected = Quasi protected
"Muslim migrants ought to be killed" = Quasi protected
Not Protected + Quasi protected = not protected
"Keep the horny migrant teenagers away from our daughters" = allowed

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:

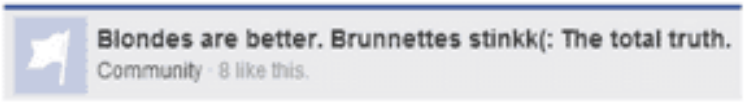


Fig. 19:
In separate notes, Facebook tells moderators to ignore this message and caption because it says "filthy" is not the same as "filth" and "calling migrants" thieves is not violating

Fig. 20:

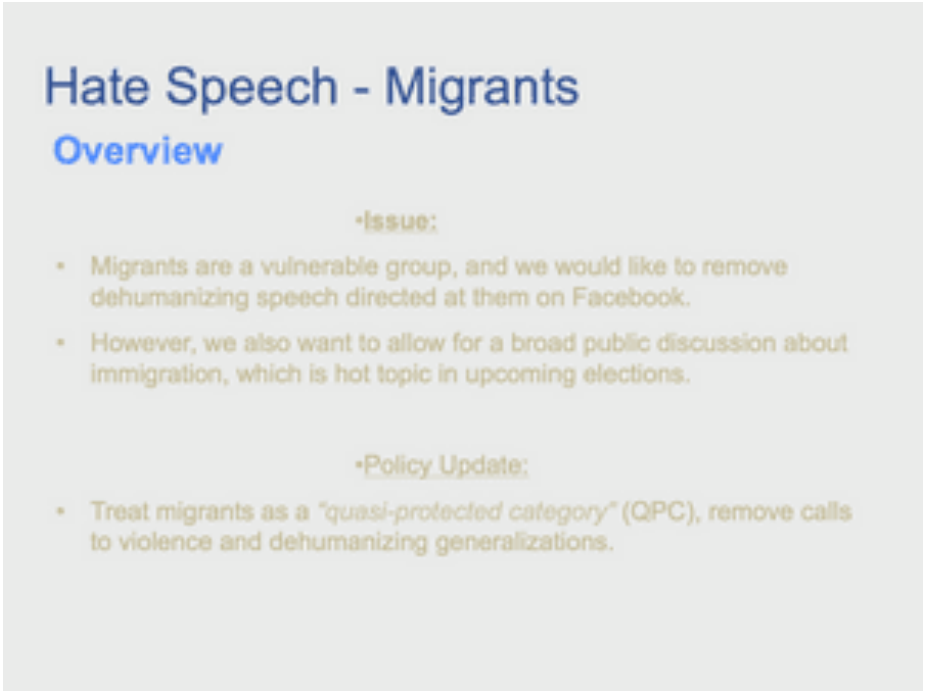


Fig. 21:

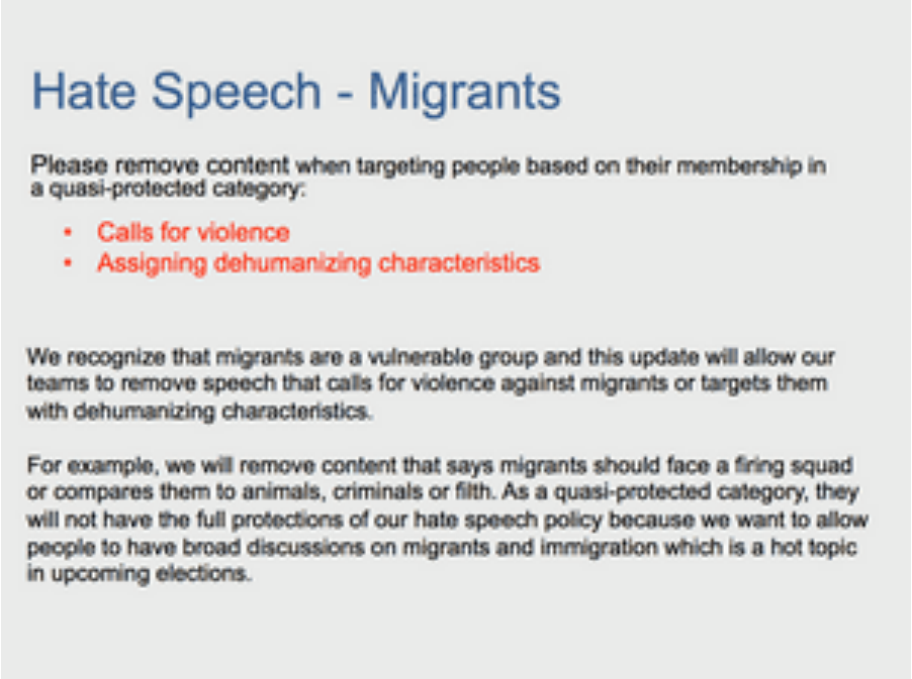

Hate Speech - Migrants
Please remove content when targeting people based on their membership in a quasi-protected category:
• Calls for violence
• Assigning dehumanizing characteristics
We recognize that migrants are a vulnerable group and this update will allow our teams to remove speech that calls 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:

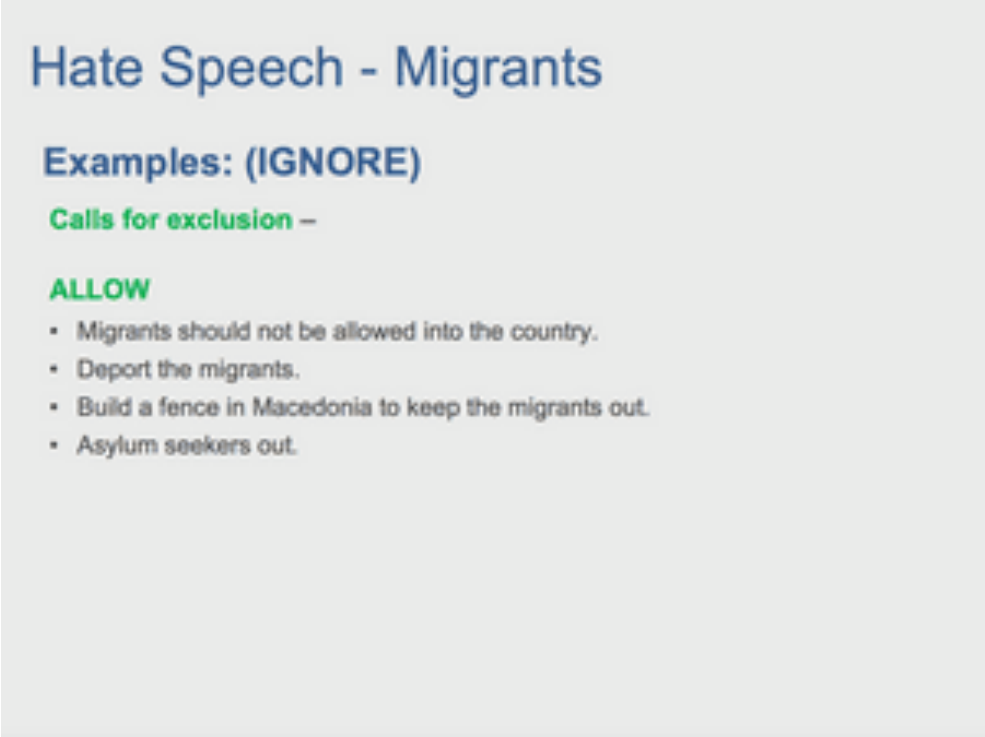

Hate Speech - Migrants
Examples: (IGNORE)
Calls for exclusion –
ALLOW
• Migrants should not be allowed into the country.
• Deport the migrants.
• Build a fence in Macedonia to keep the migrants out.
• Asylum seekers out.

Fig. 24:

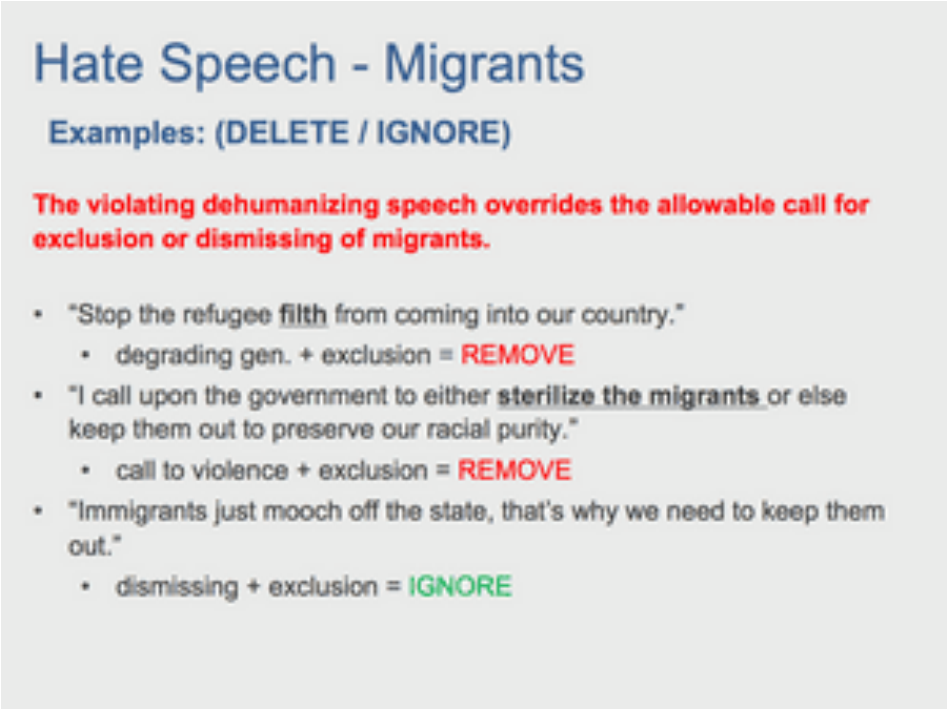
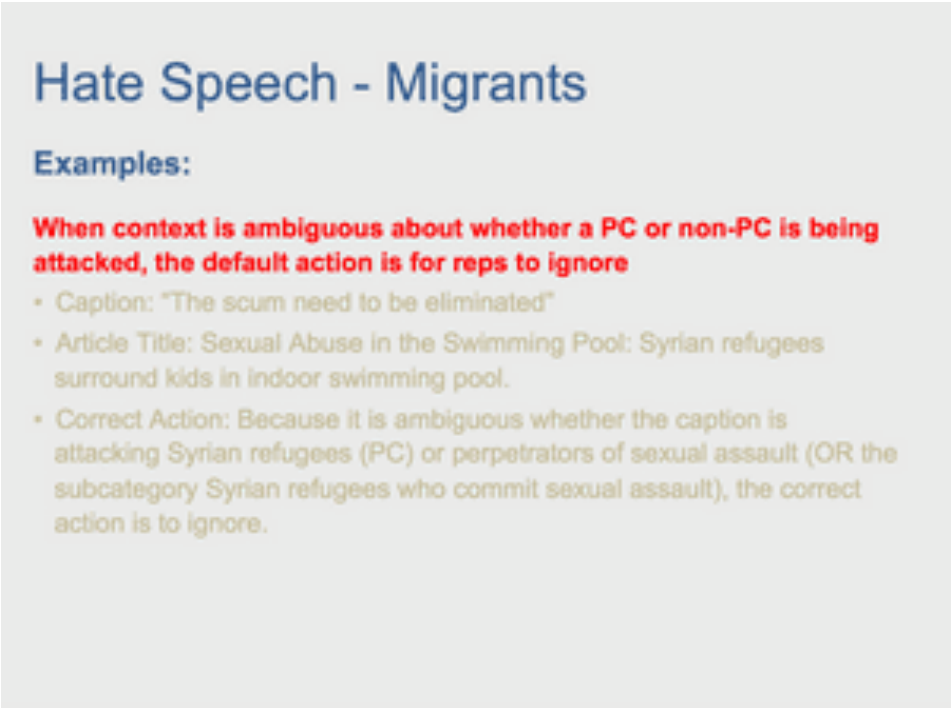

Hate Speech - Migrants
Examples: (DELETE / IGNORE)
The violating dehumanizing speech overrides the allowable call for exclusion or dismissing of migrants.
• "Stop the refugee filth from coming into our country."
• degrading gen. + exclusion = REMOVE
• "I call upon the government to either sterilize the 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.