

FACULDADE MERIDIONAL – IMED
ESCOLA DE DIREITO
PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU* EM DIREITO – PPGD
MESTRADO EM DIREITO

Álerton Emanuel Poletto

O DEVER JURÍDICO DE MODERAÇÃO DE CONTEÚDO PELO FACEBOOK

Passo Fundo, RS
2022

Álerton Emanuel Poletto

O DEVER JURÍDICO DE MODERAÇÃO DE CONTEÚDO PELO FACEBOOK

Dissertação apresentada ao Programa de Pós-graduação Stricto Sensu – Mestrado em Direito – da Faculdade Meridional – IMED, em sua área de concentração em Direito, Democracia e Tecnologia, Linha de Pesquisa Dimensões jurídico-políticas da Tecnologia e da Inovação, como requisito à obtenção do título de Mestre.

Orientador: Prof. Dr. Fausto Santos de Moraes

Passo Fundo, RS

2022

CIP – Catalogação na Publicação

- P765d POLETTTO, Álerton Emanuel
O dever jurídico de moderação de conteúdo pelo facebook / Álerton Emanuel Poletto. – 2022.
122 f., il.; 30 cm.
- Dissertação (Mestrado em Direito) – Faculdade IMED, Passo Fundo, 2022.
Orientador: Prof. Dr. Fausto Santos de Moraes.
1. Direitos fundamentais – Liberdade de expressão. 2. Democracia – Facebook. 3. Democracia – Moderação de conteúdo. I. MORAIS, Fausto Santos de, orientador. II. Título.

CDU: **316**

Catalogação: Bibliotecária Angela Saadi Machado - CRB 10/1857

Autor: Álerton Emanuel Poletto

Título: O DEVER JURÍDICO DE MODERAÇÃO DE CONTEÚDO PELO FACEBOOK

Dissertação apresentada ao Programa de Pós-graduação *Stricto Sensu* – Mestrado em Direito – da Faculdade Meridional – IMED, em sua área de concentração em Direito, Democracia e Tecnologia, Linha de Pesquisa Dimensões jurídico-políticas da Tecnologia e da Inovação, e aprovada pela banca examinadora.

Passo Fundo, RS, 16 de março de 2022.

Coordenadora Prof.^a Dr.^a Salete Oro Boff - IMED

Prof. Dr. Fausto Santos de Moraes - IMED

Prof. Dr. Vinicius Borges Fortes - IMED

Prof. Dr. Lucas Gonçalves da Silva - UFS

AGRADECIMENTOS

O agradecimento representa a valoração da trajetória percorrida. Por isso, agradeço demasiadamente minha família por todo suporte e compreensão, principalmente nos momentos difíceis durante o mestrado.

Estendo meus agradecimentos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio no desenvolvimento desta pesquisa prestado pela concessão de bolsa PROSUP.

Agradeço aos colegas do Grupo de Pesquisa IAJUS TEAM pelo espaço de discussão acadêmica e pelo compartilhamento de vivências na pesquisa jurídica.

Igualmente, estendo os agradecimentos aos professores do Programa de Pós-Graduação em Direito da Faculdade Meridional – IMED, os quais contribuíram na construção desta pesquisa, e aos colegas que dividiram o desafio de realizar o mestrado.

Por fim, meu sincero agradecimento e gratidão ao Professor Fausto Santos de Moraes, orientador deste trabalho, responsável por me instigar a realizar uma pesquisa científica no Direito com seriedade. És um grande exemplo acadêmico de pesquisador e professor.

RESUMO

Considerando a extensão da rede social do Facebook, surgiu a necessidade da plataforma de gerir seu ambiente digital e controlar o que circula dentro de sua plataforma. Desse modo, o Facebook utiliza-se da atividade de moderação de conteúdo com o objetivo de filtrar e controlar tudo o que é publicado na sua rede social, podendo implicar na remoção, bloqueio ou restrição do conteúdo considerado violador do ambiente digital. Pode-se observar que tal atividade se relaciona com o exercício da liberdade de expressão online. Considerando que direitos fundamentais são amplamente exercidos no ambiente do Facebook, as limitações, restrições e a proteção dos direitos fundamentais devem ser constitucionalmente adequadas com a ordem jurídica e fundada na teoria dos direitos fundamentais. Assim, parte-se do problema de se existe um dever jurídico do Facebook em moderar o conteúdo publicado pelos usuários, sob pena de incidir em uma proteção deficiente dos direitos fundamentais dos usuários-cidadãos. Trabalha-se com a hipótese de existir um dever jurídico de moderação do conteúdo, a qual é evidenciada através do objetivo central do trabalho de verificar se de fato há o dever jurídico. Para isso, o trabalho constitui-se de dois capítulos, a fim de responder aos objetivos específicos de apresentar como o Facebook modera os conteúdos, abordar a relação entre a moderação de conteúdo e os direitos fundamentais e, por fim, criticar o tratamento jurídico sobre a moderação de conteúdo frente ao dever jurídico imposto ao Facebook de moderar o conteúdo na sua rede social. O trabalho justifica-se institucionalmente à área de concentração do PPGD IMED em Direito, Democracia e Tecnologia e a respectiva linha de pesquisa Dimensões jurídico-políticas da Tecnologia e da Inovação ao eleger como objeto de pesquisa a proteção na ordem jurídica brasileira nas relações desenvolvidas na plataforma do Facebook e o dever/obrigação de moderação de conteúdo ao sustentar o dever de proteção aos direitos fundamentais integrando o objeto de pesquisa à dimensão jurídico-política da tecnologia, frente às repercussões da tecnologia nos direitos fundamentais. O trabalho foi elaborado com base no método fenomenológico-hermenêutico, a fim de verificar a relação entre a moderação de conteúdo e os direitos fundamentais, através dos meios técnicos de pesquisa bibliográfica na literatura da temática, bem como exploração das propostas jurídicas para a moderação de conteúdo. Essa pesquisa foi financiada por meio da concessão de bolsa CAPES/PROSUP.

Palavras-chave: Moderação de conteúdo. Facebook. Liberdade de expressão. Direitos fundamentais. Constitucionalismo Digital. Democracia. Tecnologia.

ABSTRACT

Considering the extension of Facebook's social media, the need arose for the platform to manage its digital environment and control what circulates within its platform. In this way, Facebook uses the activity of content moderation in order to filter and control everything that is published on its social media, which may imply in the removal, blocking or restriction of content considered infringing the digital environment. It can be seen that such activity is related to the exercise of freedom of speech online. Considering that fundamental rights are widely exercised in the Facebook environment, the limitations, restrictions and protection of fundamental rights must be constitutionally adequate with the legal order and founded on the theory of fundamental rights. Thus, it starts with the problem of whether Facebook has a legal duty to moderate the content published by users, under penalty of incurring in a deficient protection of the fundamental rights of users-citizens. It works with the hypothesis that there is a legal duty to moderate the content, which is evidenced through the main objective of the work of verifying whether there is in fact a duty. For this, the work consists of two chapters, in order to respond to the specific objectives of presenting how Facebook moderates content, addressing the relationship between content moderation and fundamental rights and, finally, criticizing the legal treatment of content moderation in face of the legal duty imposed on Facebook to moderate content on its social media. The work is institutionally justified in the area of concentration of the PPGD IMED in Law, Democracy and Technology and the respective line of research Legal-Political Dimensions of Technology and Innovation by choosing as research object the protection in the Brazilian legal order in the relations developed in the Facebook platform and the duty/obligation of content moderation by supporting the duty to protect fundamental rights by integrating the research object to the legal-political dimension of technology, in view of the repercussions of technology on fundamental rights. The work was elaborated based on the phenomenological-hermeneutic method, in order to verify the relationship between content moderation and fundamental rights, through technical means of bibliographical research in the thematic literature, as well as exploration of legal proposals for the moderation of contents. This was funded through a CAPES/PROSUP research grant.

Keywords: Content moderation. Facebook. Freedom of speech. Fundamental rights. Digital Constitutionalism. Democracy. Technology.

LISTA DE ILUSTRAÇÕES

Figura 1 -Esquema moderação de conteúdo pelo Facebook	34
Figura 2 - Telas de denúncia de conteúdo.	59
Figura 3 - Telas de experiência de remoção: notificação.	60
Figura 4 - Telas de experiência de remoção: revisão da decisão.	61
Figura 5 - Telas de alerta e indicação de informações adicionais.....	61
Figura 6 - Tela de cadastro no Facebook.....	65
Figura 7 - Padrões da Comunidade do Facebook.	69
Figura 8 - Violência e comportamento criminoso.....	70
Figura 9 - Segurança.	70
Figura 10 - Conteúdo questionável.	71
Figura 11 - Integridade e autenticidade.....	71

LISTA DE TABELAS

Tabela 1 - Sistema de strikes	37
Tabela 2 - Indicativos de moderação de conteúdo - nudez e atividade sexual de adultos	41
Tabela 3 - Indicativos de moderação de conteúdo - bullying e assédio	41
Tabela 4 - Indicativos de moderação de conteúdo – perigo infantil	42
Tabela 5 - Indicativos de moderação de conteúdo - organizações perigosas: terrorismo	42
Tabela 6 - Indicativos de moderação de conteúdo - organizações perigosas: ódio organizado	43
Tabela 7 - Indicativos de moderação de conteúdo – contas falsas	43
Tabela 8 - Indicativos de moderação de conteúdo – discurso de ódio	43
Tabela 9 - Indicativos de moderação de conteúdo – bens regulamentados: drogas	44
Tabela 10 - Indicativos de moderação de conteúdo – bens regulamentados: armas de fogo.....	44
Tabela 11 - Indicativos de moderação de conteúdo – spam	45
Tabela 12 - Indicativos de moderação de conteúdo – suicídio e automutilação	45
Tabela 13 - Indicativos de moderação de conteúdo – violência e incitação.....	46
Tabela 14 - Indicativos de moderação de conteúdo – conteúdo violento e gráfico ...	46

SUMÁRIO

1 INTRODUÇÃO	12
2 A MODERAÇÃO DE CONTEÚDO DE CONTEÚDO PELO FACEBOOK	17
2.1 Contexto informacional das plataformas de aplicação	17
2.1.1 Ascensão da tecnologia: a revolução da internet	17
2.1.2 Tecnologia e sociedade: a internet como elemento social	19
2.1.3 Novas relações digitais de poder pelas plataformas	21
2.1.4 The Facebook's Society	24
2.2 Delineando a moderação de conteúdo	27
2.2.1 Conceito e tipos de moderação de conteúdo	28
2.2.2 Processo de moderação pelo Facebook	31
2.2.3 Das violações na plataforma	36
2.2.4 Indicativos do Facebook de moderação de conteúdo	40
2.3 Implicações da moderação de conteúdo	47
2.3.1 Questões tecnológicas da moderação	52
2.3.2 Governança na moderação de conteúdo pelo Facebook	56
2.3.3 Termos de Serviço e Padrões da Comunidade: regras da moderação de conteúdo pelo Facebook	65
3 DESAFIOS JURÍDICOS À MODERAÇÃO DE CONTEÚDO	Erro! Indicador não definido.
3.1 Desafios Internacionais na regulação do Facebook	Erro! Indicador não definido.
3.1.1 Estados Unidos	Erro! Indicador não definido.
3.1.2 Reino Unido	Erro! Indicador não definido.
3.1.3 Alemanha	Erro! Indicador não definido.
3.1.4 Portugal	Erro! Indicador não definido.
3.1.5 França	Erro! Indicador não definido.

3.1.6 União Europeia	Erro! Indicador não definido.
3.2 O Facebook como gestor da liberdade de expressão	Erro! Indicador não definido.
3.2.1 A constitucionalização digital	Erro! Indicador não definido.
3.2.2 Dos direitos fundamentais aos termos de uso do Facebook: implicações teóricas	Erro! Indicador não definido.
3.2.3 Liberdade de expressão online	Erro! Indicador não definido.
3.2.4 Legitimação das plataformas para a liberdade de expressão	Erro! Indicador não definido.
3.3 Propostas jurídicas sobre a moderação de conteúdo	Erro! Indicador não definido.
3.3.1 Medidas Legislativas	Erro! Indicador não definido.
3.3.1.1 Marco Civil da Internet	Erro! Indicador não definido.
3.3.1.2 Lei Geral Proteção de Dados	Erro! Indicador não definido.
3.3.1.3 Projeto de Lei nº 2.630/2020	Erro! Indicador não definido.
3.3.1.4 Projeto de Lei nº 3.227/2021	Erro! Indicador não definido.
3.3.2 Discussões Judiciais	Erro! Indicador não definido.
3.3.2.1 A responsabilidade sobre a moderação de conteúdo	Erro! Indicador não definido.
3.3.2.2 A fórmula de ponderação da internet	Erro! Indicador não definido.
4 CONCLUSÃO	Erro! Indicador não definido.
REFERÊNCIAS	

1 INTRODUÇÃO

A internet mudou radicalmente as relações pessoais e as formas de comunicação social, manifestação e debate público, principalmente com o surgimento e ascensão das redes sociais. As plataformas digitais, criadas e administradas por empresas privadas, tornaram-se um ambiente no qual direitos fundamentais são frequentemente exercidos e limitados por esses agentes privados de acordo com suas próprias regras.

A rede social do Facebook se tornou a plataforma mais utilizada globalmente, contendo mais de 3 bilhões de usuários, os quais publicam diariamente milhões de conteúdo na rede social. Ante a necessidade de gerir seu ambiente digital, ampliou-se a tarefa do Facebook em controlar o conteúdo publicado pelos usuários por meio da moderação de conteúdo. Tal atividade é exercida massivamente por sistemas tecnológicos de Inteligência Artificial e também por suporte humano.

O Facebook faz uso da atividade de moderação de conteúdo com o fim de filtrar e controlar tudo o que é publicado na sua rede social, podendo incorrer na remoção, bloqueio ou restrição do conteúdo considerado violador do ambiente digital. Contudo, tal atividade se relaciona diretamente com o exercício da liberdade de expressão nesse ambiente digital, implicando ao Facebook estabelecer os limites do que pode ou não circular na sua rede social.

Sobreleva evidenciar alguns exemplos marcantes do conflito entre a moderação e a liberdade de expressão de usuários/personalidades políticas, o presidente Jair Bolsonaro e o presidente americano Donald Trump. Em março de 2020, ambos presidentes tiveram publicações removidas no Twitter, Facebook e Instagram, em razão dos posts e tweets estarem em desacordo com as orientações de combate à pandemia do COVID-19, destacando o potencial lesivo da desinformação sobre a doença à saúde pública. Outro exemplo, em janeiro de 2021, Donald Trump teve sua conta suspensa no Twitter, Instagram, Youtube e Facebook, após os protestos violentos e invasão ao Capitólio por parte dos seus apoiadores, como uma medida de emergência.

Direitos fundamentais, especialmente a liberdade de expressão, são exercidos nas relações dentro do ambiente da internet e das redes sociais, implicando

em um dever de proteção eficiente desses direitos frente aos eminentes abusos e conflitos. Tradicionalmente, incumbe ao Estado o dever de proteção dos direitos fundamentais, entretanto, o Estado brasileiro não possui controle e condições de fiscalização das novas relações advindas do mundo digital, podendo incorrer em uma proteção deficiente dos direitos fundamentais dos usuários-cidadãos.

Considerando o fenômeno da moderação de conteúdo, implicam alguns questionamentos jurídicos relativos ao possível dever jurídico do Facebook em moderar conteúdos publicados pelos usuários sob o prisma dos direitos fundamentais. Neste aspecto, impende reforçar que direitos fundamentais são exercidos no ambiente digital, urgindo um dever de proteção.

A presente pesquisa tem como objeto de estudo um marco teórico que envolve o Constitucionalismo digital e a eficácia horizontal da proteção dos direitos fundamentais exercidos no ambiente digital, a fim apoiar as compreensões acerca das ações do Facebook, como bloqueio, remoção e filtro de conteúdo publicado em sua plataforma social, estão de acordo com os limites constitucionalmente adequados para a limitação de direitos fundamentais, de acordo com a teoria dos direitos fundamentais e com vistas à ordem jurídica brasileira.

Posto isso, o problema em discussão fundar-se-á em se existe e qual a extensão de dever jurídico do Facebook em moderar o conteúdo publicado pelos usuários, sob pena de implicar na proteção deficiente dos direitos fundamentais online, com vistas à ordem jurídica brasileira.

O dever de proteção dos direitos fundamentais importa em uma proibição de proteção deficiente, do qual o Estado não pode abrir mão, por mais que surjam novos desafios para o Estado controlar e fiscalizar a proteção dos direitos exercidos nos ambientes digitais contra abusos. Assim, considera-se a hipótese da existência de um dever jurídico de moderação de conteúdo pelo Facebook, sob pena de incidir em uma proteção deficiente dos direitos fundamentais dos usuários-cidadãos.

Está implícito nesse dever de proteção o poder que o Facebook tem de moderar o conteúdo publicado pelos usuários em sua plataforma e a justificação do uso de tal poder decorre tanto dos termos e padrões do Facebook quanto da proteção dos direitos fundamentais, o que não implica que esse poder não possa ser questionado juridicamente.

Constitui-se como objetivo geral da pesquisa verificar se o Facebook possui o dever jurídico de moderar os conteúdos publicados na rede social, especificando-se em apresentar como o Facebook modera os conteúdos na plataforma, abordar a relação entre a moderação de conteúdo com os Direitos Fundamentais, e criticar o tratamento jurídico sobre a moderação de conteúdo.

O presente trabalho se justifica por fatores de diferentes ordens. São elas de característica social, jurídico-constitucional, tecnológica, institucionais do programa do mestrado e objeto de pesquisa do grupo de pesquisa IAJUS TEAM, bem como a atualidade da discussão ao considerar os movimentos nacionais e internacionais de discussões de regulação da moderação de conteúdo.

Sob a perspectiva social, a presente pesquisa se mostra relevante em virtude da centralidade que as plataformas sociais no convívio social, uma vez que a tecnologia se torna, cada vez mais, fundamental para a existência humana. Quanto ao fator jurídico-constitucional, as novas relações provenientes do uso da internet fomentam novos desafios para o Estado, especialmente, no que tange à proteção dos cidadãos em um espaço digital, transnacional e gerido por agentes privados.

Outrossim, pelos mecanismos tradicionais o Estado não possui condições de exercer controle do ambiente digital, devido a esse ambiente ter como principal característica a falta de fronteiras territoriais e jurídicas. Assim sendo, implica ao Direito impor o dever jurídico para que as plataformas digitais promovam ambientes seguros ao exercício dos direitos fundamentais.

A pressão jurídica para que os direitos fundamentais sejam protegidos eficientemente e estabelecer limites ao controle privado no exercício de direitos fundamentais pode ser traduzida pelo constitucionalismo digital, o qual intenta garantir eficácia horizontal dos direitos fundamentais exercidos na internet.

Finalmente, a pesquisa impende grande valor ao programa de pós-graduação *stricto sensu*, uma vez que busca investigar as regulações jurídicas da tecnologia, a fim de garantir a proteção dos direitos fundamentais nos ambientes digitais, tendo por fundamento as bases teóricas do programa, relativas à área de concentração do PPGD/IMED em Direito, Democracia e Tecnologia e a respectiva linha de pesquisa Dimensões jurídico-políticas da Tecnologia e da Inovação ao eleger como objeto de pesquisa a proteção na ordem jurídica brasileira nas relações

desenvolvidas na plataforma do Facebook e o dever/obrigação de moderação de conteúdo.

Outrossim, a pesquisa consolida os estudos desenvolvidos no grupo de pesquisa IAJUS, em especial, acerca das consequências jurídicas, éticas e democráticas decorrentes dos impactos jurídicos da moderação de conteúdos pelas redes sociais e da manipulação comportamental através das plataformas sociais, exigindo, portanto, uma resposta eficiente do Direito ao sustentar o dever de proteção aos direitos fundamentais integrando o objeto de pesquisa à dimensão jurídico-política da tecnologia, frente às repercussões da tecnologia nos direitos fundamentais.

A metodologia adotada na pesquisa, em relação à linha de abordagem e elaboração da temática refere-se ao método fenomenológico-hermenêutico, como um meio para compreender e expressar a percepção sobre a relação da moderação de conteúdo com os direitos fundamentais, partindo de um possível dever jurídico do Facebook em moderar o conteúdo na sua rede social e buscando identificar as principais implicações jurídicas do tema.

Nessa senda, a partir da hipótese formulada de que o Facebook deve moderar sob pena de uma proteção deficiente dos direitos fundamentais, implicam questionamentos jurídicos logicamente organizados e respondidos nos dois capítulos propostos no *corpus* da pesquisa: compreender a atividade de moderação de conteúdo pelo Facebook e verificar os desafios jurídicos da moderação.

Com efeito, considerando o referencial assimilado, foi possível estabelecer os meios técnicos da pesquisa pelas etapas jurídicas para verificação do dever jurídico do Facebook em moderar conteúdos, iniciando pela contextualização de como é realizada a moderação de conteúdo pelo Facebook, para, na sequência, relacionar a moderação com a teoria dos direitos fundamentais.

Buscando exprimir a os meios técnicos para cumprir objetivos específicos, primeiramente, importa em delinear como a moderação de conteúdo ocorre por meio de pesquisa monográfica do modo técnico dessa operação. Posteriormente, a expressão do fenômeno se dará pela pesquisa das regras que regem o Facebook no sítio informacional da própria plataforma. Por fim, para alcançar o objetivo de propor uma crítica ao tratamento jurídico da moderação, importa explorar as regulações

internacionais sobre moderação e as propostas jurídicas para a moderação de conteúdo com fundamento na teoria dos direitos fundamentais.

Pelo ponto de vista procedimental, é do tipo exploratória e utiliza o auxílio de fontes bibliográficas. Houve a exploração de obras de vários autores, os quais debatem a temática abordada, configurando-se o corpus da pesquisa realizada, de forma que, a partir delas, pode-se verificar se o Facebook possui o dever jurídico de moderar os conteúdos considerados violadores publicados em sua plataforma, a fim de garantir uma proteção eficiente dos direitos fundamentais.

2 A MODERAÇÃO DE CONTEÚDO DE CONTEÚDO PELO FACEBOOK

Com o propósito de melhor explorar a pesquisa foi dividido o capítulo em três títulos. A pesquisa inicia contextualizando a ascensão da internet e das redes sociais online e como a tecnologia se tornou um elemento social, até as mudanças e implicações na sociedade ao passo que o Facebook construiu o monopólio das redes sociais.

Posteriormente, impende explanar a atividade da moderação de conteúdo pelas plataformas de redes sociais online, bem como as implicações que decorrem dessa atividade pelo Facebook para, por fim, abordar os impactos jurídicos decorrentes da moderação de conteúdo sobre os usuários-cidadãos.

2.1 Contexto informacional das plataformas de aplicação

Em síntese, no presente título busca-se expor o contexto social das plataformas de aplicação iniciando pela ascensão da tecnologia e da internet na construção do as redes sociais representam, destacando os impactos nas relações sociais, especialmente sobre o impacto do Facebook na vida dos usuários.

2.1.1 Ascensão da tecnologia: a revolução da internet

Inicialmente, cumpre contextualizar a ascensão da tecnologia e a internet, a fim de formar uma breve base histórica-social que permita refletir acerca das mudanças com o advento da internet e, posteriormente, as redes sociais, e assim poder, ao longo do trabalho, identificar os impactos das atividades do Facebook, em especial da moderação de conteúdo.

Os avanços tecnológicos, ao longo da história da humanidade, propiciam mudanças sociais significativas desde o período neolítico até hoje, com o advento das redes sociais. Assim, no século XX, abriu-se a possibilidade para o compartilhamento de dados na forma de dígitos, resultando no “desenvolvimento de uma teia de conexões descentralizadas que veio a se tornar a internet” (MARTINO, 2014, p. 12).

A origem da Internet se deu como parte de uma rede de operações militares norte-americanas nos anos de 1950 e 1960. Contudo, o que conhecemos hoje por 'Internet', isto é, a rede de navegação e compartilhamento de dados, a chamada *World Wide Web*, foi desenvolvida em 1991, por Tim Berners-Lee e seus colegas no Centro Europeu de Pesquisas Nucleares.

A transformação do modelo militar de internet, iniciou pela criação das páginas e sites na Web 1.0. Com a mudança do mundo offline para o digital, as interações entre os usuários foi o ponto central para a dinamicidade das plataformas na Web 2.0 pela constante transformação do ambiente (MARTINO, 2014, p 12-13).

Não obstante, Castells afirma que o novo mundo de comunicação, com difusão a internet, pela primeira vez, foi possível integrar “muitos com muitos” em escala global, em uma “Galáxia da Internet” (2015, p. 7). O desenvolvimento tecnológico a partir da globalização acelerou os avanços científicos ao ponto de estimar-se que nos dez primeiros anos do século XXI equivalem aos avanços tecnológicos de todo o período do século XX (CALLEJÓN, 2020, p. 585).

Em 2003, era estimado que para 2010 havia aproximadamente 2 bilhões de usuários na internet (CASTELLS, 2015, p. 8). Entretanto, um relatório de 2020 promovido pelo *We Are Social* e *Hootsuite* aponta que mais de 4.54 bilhões de pessoas usam a internet no mundo todo (KEMP, 2020, p. 8), representando uma superação da estimativa inicial.

Outros dados que importam destaque do relatório *Digital 2020*, dispõem que mais de 5.19 bilhões de pessoas possuem aparelhos de celular e 3.80 bilhões são usuários ativos nas redes sociais, de uma estimativa populacional global de 7.75 bilhões de pessoas (KEMP, 2020, p. 8).

A globalização não apenas apressou o desenvolvimento tecnológico “como mudou a relação dos seres humanos com as dimensões de espaço e tempo previamente conhecidas”. O desenvolvimento tecnológico, também, vem transformando o jogo político e constitucional pela aceleração do tempo histórico e ampliação do espaço público (CALLEJÓN, 2020, p. 585).

De outra banda, a tecnologia da informação encontra-se, atualmente, equivalente ao que a eletricidade foi na Era Industrial, sendo a internet o tecido de nossas vidas (CASTELLS, 2015, p. 7). A título de comprovação, o relatório *Digital*

2020, expõem que a média de uso da internet no Brasil é de 9 horas e 17 minutos diários, entre pessoas de 16 a 64 anos, todo o dia, em qualquer dispositivo (KEMP, 2020, p. 43).

Conforme aponta o relatório, o Brasil é o terceiro país com maior média de uso da internet (KEMP, 2020, p. 43). Não obstante, a Internet passou a ser um elemento da sociedade, que a transforma pelo impacto da hiperconectividade, uma vez ser já impossível separar a tecnologia da sociedade.

2.1.2 Tecnologia e sociedade: a internet como elemento social

A ascensão da tecnologia importou na intensificação do uso da internet de forma a não ser mais possível separar a sociedade da internet. Assim, cumpre-se, neste título, apresentar a relação da tecnologia e da sociedade, bem como os impactos jurídicos-sociais da internet.

Sociedade em rede (CASTELLS, 2011), civilização da informação (ZUBOFF, 2019) sociedade filtrada (PARISER, 2012), comunidades virtuais (CASTELLS, 2015) são alguns termos que buscam definir as mudanças da sociedade com a chegada da internet e das redes sociais.

Afirmções conflitantes foram associadas à internet por ser um novo meio de comunicação e estabelecer novos padrões de interação social. De um lado, a comunicação online por meio de comunidades virtuais iria substituir as formas de interação humana limitadas territorialmente; outra tônica preceitua que a internet conduziria ao isolamento social, criando um colapso nas interações face a face em ambientes fora do virtual (CASTELLS, 2015, p. 110).

A interatividade que impulsiona o formato atual da web apresenta um panorama novo, uma vez que, simultaneamente, os produtores são o público e este também é criador do conteúdo que circula na web. “A internet deixou de ser apenas uma rede de computadores e se consolidou como uma rede de pessoas” (PINHEIRO, 2013, p. 196).

Nesta senda, “a Internet passou a ser a base tecnológica para a forma organizacional da Era da informação: a rede” (CASTELLS, 2015, p. 7). Todavia, tem-se que a tecnologia não determina a sociedade, nem a sociedade escreve o curso da

transformação tecnológica, dado que diversos fatores impactam nos processos de inovação, descoberta e aplicação social, como a própria iniciativa empreendedora e a criatividade (CASTELLS, 2011, p. 43). A nova estrutura social baseada em redes foi estruturada através de três processos: demandas econômicas por flexibilidade administrativa e globalização do capital, da produção e do comércio; exigências sociais por direitos à liberdade individual; e avanços tecnológicos nas áreas da computação e telecomunicação. Com a internet, esses processos formaram uma nova sociedade, a sociedade em rede (CASTELLS, 2015, p. 8).

Os desafios da sociedade digital são grandes e repercutem tanto na esfera pública quanto na esfera privada. De fato, a esfera pública e a privada convergiram de modo que não há mais espaço entre o público e privado, “pois nessa atual sociedade, o espaço público é inundado com questões eminentemente privadas, numa invasão à intimidade da pessoa, por vezes contra a vontade dela” (GONÇALVES DA SILVA; CARVALHO, 2017, p. 69).

A comunicação social também sofreu transformações na sociedade da informação, de modo que a comunicação difundida dentro do meio ambiente digital pelos meios de comunicação deve garantir os o princípio democrático (FIORILLO; SOUZA, 2016, p. 202). Sobreleva destacar que, para Fiorillo, o meio ambiente digital deriva do Direito Ambiental, uma vez que tem como objeto de estudo os bens relativos aos grupos formadores da sociedade no meio digital e na internet (2014). Nessa esteira, a internet virou um elemento social, a ponto de compreender um mercado global de *smart home* [casa inteligente] dentro de um sistema de *Aware Home* [Lar Consciente] de conexão das informações dos indivíduos dentro do seu lar e realizando a coleta de dados desse ambiente para a destinação comercial (ZUBOFF, 2019, p. 19-20). Além disso, convém explicar a denúncia feita a esses processos que expõe a mutação do poder capitalista para as atividades no mundo digital pela busca da modulação comportamental através de um capitalismo de vigilância (ZUBOFF, 2019).

Não obstante, a sociedade também é afetada pela estrutura das mídias e da internet na era da personalização (PERISER, 2012, p. 15). A promessa da personalização do conteúdo na internet implica viver uma vida dentro de um filtro-

bolha criado por algoritmos¹, no qual esse filtro molda nossas experiências individuais de forma a nos confinarmos em um universo pessoal (PERISER, 2012, p. 13).

Igualmente relevante é destacar que as plataformas privadas que criaram e controlam essa infraestrutura na internet são os Novos Governadores na era digital, controlando o discurso digital e criando uma cultura democrática global em uma estrutura de governança dirigida pelas plataformas (KLONICK, 2018, p.1663-1664). Os Novos Governadores são entidades privadas de autorregulação que têm motivação econômica e normativa para refletir a cultura democrática e as expectativas de liberdade de expressão de seus usuários (KLONICK, 2018, p. 1669-1670).

No que tange às transformações sociais, a migração da sociedade para o espaço da internet, isto é, para o ambiente digital, cria uma estrutura técnica-operacional de um conjunto de práticas pessoais conectadas a uma rede de computadores de forma a institucionalizar uma cibercultura para uma comunidade virtual desterritorializada, transversal e universal (PIERRE, 1999, p. 126-130).

Outro aspecto a ser sublinhado compreende a relação jurídica da internet na sociedade firmando como um direito humano. Alguns países da Europa, como a Finlândia, reconhecem o acesso a Internet como um direito fundamental, bem como a Organização das Nações Unidas (ONU) pronunciou-se que o acesso à internet é um direito humano. No Brasil, existe um Projeto de Emenda Constitucional nº 6 de 2011 que buscava alterar o rol de direitos sociais, a fim de introduzir o acesso à internet.

Feitas tais constatações, cabe assentar que as mudanças sociais com a eclosão da internet foram significativas para a identificação da sociedade como atualmente se figura. Não obstante, as plataformas digitais também possuem um papel nessas transformações, principalmente nas relações que passam a ser digitais.

2.1.3 Novas relações digitais de poder pelas plataformas

As mudanças a partir da internet podem ser percebidas não só na sociedade, mas também na transformação nas diversas relações para o mundo digital.

¹ Por algoritmos, define-se: Um algoritmo é uma sequência de instruções minuciosas que informa ao computador o que ele deve fazer (DOMINGOS, 2017, p. 20).

Assim, neste título, será abordada a ideia fundacional das redes sociais e a transformação das relações pelas redes sociais.

Inicialmente, cumpre apresentar uma conceituação do que seja redes sociais, bem como delinear as principais características. Nessa perspectiva, redes sociais são um tipo de “relações entre seres humanos pautadas pela flexibilidade de sua estrutura e pela dinâmica entre seus participantes [...] definidas por seu caráter horizontal, desprovido de uma hierarquia rígida” (MARTINO, 2014, p. 55).

Outra tônica de apreciação expõe que as redes sociais “[...] foram suplantadas como ferramentas de organizações capazes de congregar recursos em torno de metas centralmente definidas”, as quais são obtidas por meio da definição de “tarefas em cadeias e comando e controle verticais e racionalizadas” (CASTELLS, 2015, p. 7-8).

Não obstante, impende fazer distinção com o que venha ser uma plataforma. O termo plataforma refere-se a intermediários online operando em diversos setores, como o Facebook na mídia e o Uber no transporte (CAPLAN, 2018, p. 8). Isto é, as plataformas são sites e serviços online que hospedam, organizam e divulgam o conteúdo e as interações que os usuários compartilham, “sem ter produzido ou encomendado (a maior parte) esse conteúdo, construída sobre uma infraestrutura, sob essa circulação de informações, para processamento de dados para atendimento ao cliente, publicidade e lucro” (GILLESPIE, 2018, p. 18).

O uso da ideia de redes sociais no ambiente virtual importa na transposição das barreiras do mundo físico para o mundo online, de forma que as interferências entre os dois mundos são inevitáveis. Exercendo, inclusive, influência no poder político na troca de informações, uma vez que os usuários não encontram barreiras institucionais ou governamentais devido à arquitetura horizontal das redes na utilização para fins políticos e democráticos (MARTINO, 2014, p. 58-59).

O relatório *Digital 2021:Brazil* aponta uma crescente de 150.0 milhões de usuários de redes sociais em janeiro de 2021, equivalente à 70.3% da população total do país na mesma data (KEMP, 2021, p. 17). Ademais, o relatório aponta uma média diária de uso de redes sociais entre os brasileiros de 16 até 64 anos de 3h42min em janeiro de 2021 (KEMP, 2021, p. 22), a qual na data em 2020 possuía a terceira maior média global com 3h31min (KEMP, 2020, p. 92).

Esse relatório representa em números a relação entre a população brasileira e as redes sociais evidenciando a ascensão dos cidadãos para a noção de usuário-cidadão. Posto isso, a migração social para o mundo online expande sem fronteiras o poder dos agentes globais, em especial das empresas de tecnologia e comunicação como o Facebook.

As redes sociais digitais se tornaram um espaço relevante de debate público democrático, assim as plataformas privadas que criaram e controlam essa infraestrutura são *The New Governors* [os Novos Governadores] na era digital (KLONICK, p. 1663).

Nessa perspectiva, as redes sociais exercem três funções centrais no espaço público: primeiro, as redes sociais facilitam a participação pública na política, na cultura e na arte; segundo, as redes sociais conseguem organizar facilmente o debate público para que os usuários comuniquem-se entre si; terceiro, as redes sociais online curam a opinião pública ao regular a propagação e o alcance dos conteúdos, através da personalização dos resultados e feeds, ou seja, pela retirada ou reorganização do conteúdo ao aplicar os padrões da comunidade e os termos de serviço (BALKIN, 2020, p. 4).

Por moderar esse espaço público, as redes sociais se apoderaram de um poder advindo do empoderamento dos indivíduos nesse espaço, através do mapeamento da internet e dos comportamentos dos usuários, estruturados em uma economia de vigilância (ZUBOFF, 2019, p. 120), pela ampliação de um ambiente onde direitos são frequentemente exercidos, tais como direito à informação, à liberdade de expressão e privacidade.

Frente ao exposto, cumpre evidenciar o escândalo envolvendo o Facebook e a empresa *Cambridge Analytica*, denunciado pela coleta de informações pessoais dos usuários para influenciar na política, especialmente em processos eleitorais democráticos em diversos países.

A face oculta das redes sociais foi exposta no documentário produzido pela Netflix, *Privacidade Hackeada* (2019), que denuncia um esquema de coleta e manipulação de dados pessoais dos usuários do Facebook, sem consentimento, com o intuito de traçar o perfil dos usuários e, assim, direcionar propagandas políticas para influenciar pleitos eleitorais.

A interferência pela *Cambridge Analytica* nas campanhas foi decisiva em diversos processos democráticos pelo mundo, como a campanha do *Brexit*, que votou pela saída do Reino Unido da União Europeia, a eleição presidencial de Donald Trump nos Estados Unidos em 2016, além de outras eleições presidenciais na Índia, no México, no Quênia e em Malta (PRIVACIDADE HACKEADA, 2019).

A própria estrutura das plataformas de redes sociais reduz a capacidade de intervenção pública (CALLEJÓN, 2020, p. 584), de modo que “no contexto da globalização, o Estado é cada vez mais incapaz de controlar os grandes agentes globais” (CALLEJÓN, 2020, p. 587), principalmente as companhias tecnológicas que detém o monopólio das redes sociais.

Uma característica essencial das plataformas digitais consiste na produção e dependência dos efeitos em rede: quanto maior o número de usuário, mais valiosa a plataforma se torna. O Facebook, por exemplo, representa uma das principais plataformas de redes sociais em razão da quantidade de usuários inscritos, os quais são influenciados a entrarem na rede pela política da socialização, na qual se você pretende usar uma plataforma para esse fim, você vai entrar na plataforma uma vez que a maioria de seus amigos e família já fazem parte da comunidade.

2.1.4 The Facebook's Society

Neste último título pretende-se explicar a idealização do Facebook, a sua construção, seus objetivos como rede social e os serviços que oferta divulgados nas informações oficiais da plataforma, para, por fim, apresentar sinteticamente os pontos problemáticos da moderação de conteúdo e das redes sociais.

Em 2003 iniciou-se o Facebook como uma rede de comunicação entre estudantes da Universidade de Harvard, exclusiva para o campus. No ano seguinte, o fundador, Mark Zuckerberg, abriu a rede social para a web e, desde então, houveram diversas transformações na plataforma, em particular a incorporação do FB Messenger, Instagram e WhatsApp pela Facebook Inc, em 2011, 2012 e 2014 respectivamente (FACEBOOK, 2021, online).

O ranking da plataforma social mais usada no ano de 2020 elenca as 4 plataformas do Facebook Inc. entre as 6 primeiras redes sociais mais utilizadas no

mundo, sendo o Facebook o detentor do primeiro lugar (KEMP, 2020, p. 95). Haja vista o exposto, o grupo Facebook domina a prestação do serviço global de aplicação social, quase em regime de monopólio.

Considerando a dimensão da rede social, impende destacar quais os objetivos e os serviços que o Facebook oferece, conforme informações da própria plataforma. Atualmente, o Facebook possui mais de 3 bilhões de pessoas conectadas à plataforma em todo o mundo e tem como princípios institucionais: dar voz às pessoas; criar conexão e comunidades; servir a todos; manter as pessoas seguras e proteger a privacidade; e promover oportunidade econômica (FACEBOOK, 2021, online).

A relevância social do Facebook pode também ser representada no relatório “Internet Health Report v0.1” de 2017, no qual mais de 55% dos brasileiros entrevistados consideram que não há nada na internet além do Facebook, isto é, para eles, “o Facebook é Internet”. Em contraponto, a mesma entrevista realizada nos Estados Unidos da América aponta um percentual de 5% de concordância com a frase “o Facebook é Internet” (MOZILLA, 2017, p. 26), representando uma ignorância informacional dos brasileiros.

Tendo em conta tal cenário, os produtos do Facebook compreendem “o Facebook (incluindo o navegador no aplicativo e o aplicativo para celular do Facebook), o Messenger, o Instagram (incluindo aplicativos como o Boomerang)”. Outrossim, os serviços fornecidos são: fornecer uma experiência personalizada; conectar pessoas e organizações; permitir a manifestação dos usuários; ajudar a descobrir conteúdo, produtos e serviços do interesse do usuário; combater condutas prejudiciais, proteger e dar suporte para a comunidade online; usar e desenvolver tecnologias avançadas para fornecer serviços seguros; fornecer experiências consistentes e contínuas entre os produtos; e possibilitar acesso global ao serviços da plataforma (FACEBOOK, 2020, online).

O financiamento desses serviços pela plataforma é provido pela venda de anúncios e propagandas de empresas e organizações que contratam a plataforma. Assim, a rede social direciona aos usuários os serviços de terceiros de forma personalizada aos interesses de cada usuário, por meio da coleta de dados pessoais e informações sobre a atividade na rede social (FACEBOOK, 2020, online).

Contudo, o sucesso financeiro da rede social não está no direcionamento de propagandas políticas, mas sim na economia de vigilância da análise comportamental dos usuários que busca a predição comportamental para influenciar e moldar comportamentos de indivíduos e até grupos, em uma lógica denunciada como capitalismo de vigilância (ZUBOFF, 2019, p. 120-124).

Entretanto, essa análise não será objeto do presente trabalho, mas sim o serviço do Facebook em controlar o conteúdo que é publicado e circula na sua plataforma de rede social. Essa função de remover e filtrar as publicações dos usuários é denominada de moderação de conteúdo.

A moderação é um serviço indispensável para as plataformas, no que tange à proteção dos usuários, ocultação de conteúdo ilegal, bem como para a organização de todo o conteúdo que circula na rede. Nesta senda, a moderação torna-se um processo necessário no controle de conteúdo pornográfico, obsceno, violento, ilegal, abusivo e de ódio.

Para exemplificar apenas um dos pontos desse processo, no primeiro trimestre de 2021, o Facebook moderou em sua plataforma 31.8 milhões de publicações removidas por indicação de conter nudez adulta e atividade sexual, o que violaria as regras da plataforma. Todavia, após revisão, do total, apenas 633 mil publicações foram restauradas por não se enquadrarem no motivo da moderação (FACEBOOK, 2021d, online).

A vida online é vulnerável a alguns comportamentos da interface entre a estrutura e a informação, como o spam e assédio. Portanto, são contra esses abusos que a moderação deve proteger (GRIMMELMANN, 2015, p. 12). Entretanto, resta evidente que as plataformas são desenhadas para moderar e assim curar todo o conteúdo que circula na rede, determinando o que pode ou não estar disponível para o acesso dos usuários.

Não obstante, a remoção de conteúdos, por mais que sejam abusivos, põe em risco a violação de outros direitos, tais qual a liberdade de expressão, bem como alerta acerca da “concentração do poder decisório nessas plataformas que, ao atuarem como agentes de censura privada, acabam por definir as pautas que

permanecem ou são excluídas do debate público” (HARTMANN; MONTEIRO, 2020, p. 390)².

Malgrado a moderação desvirtuada implicar abusos a direitos, “as plataformas não apenas não sobrevivem sem moderação, como também não são plataformas sem moderação”. Isto é, a moderação de conteúdo “não é um aspecto auxiliar do que as plataformas fazem. É essencial, elementar, definidor” (GILLESPIE, 2018, p. 21).

Assim, uma plataforma totalmente “aberta” sem qualquer regulação ou moderação de conteúdo é uma utopia da ideia democrática, uma vez que todas as plataformas moderam conteúdos e impõem regras aos usuários, até porque, seria insustentável a utilização do serviço (GILLESPIE, 2018, p. 5).

Por fim, a moderação de conteúdo, especialmente promovida pelo Facebook, opera como um filtro do que pode ou não estar disponível na plataforma digital. Nesta senda, resta oportuno evidenciar essa atividade de forma a compreender seus processos, bem como as implicações aos usuários.

2.2 Delineando a moderação de conteúdo

A fim de compreender a prática de moderação de conteúdo pelo Facebook é preciso abordar o que venha a ser essa atividade, os tipos e as técnicas de moderação, as suas características. Na sequência, será explicitado o que acontece com os usuários e sua conta quando um conteúdo publicado acaba por violar alguma regra do Facebook.

Finalmente, busca-se apresentar os números dessa atividade divulgados pelo Facebook no portal de transparência da empresa nos relatórios de aplicação das regras da rede social de acordo com os tipos de conteúdo. Essas questões serão apresentadas neste título.

² Neste trabalho, a censura é um conceito que está relacionado com as discussões sobre a moderação de conteúdo e a liberdade de expressão. A expressão censura utilizada no texto não tem um sentido positivo ou negativo, significa o ato de impedir a circulação de determinada informação. O conceito adotado de censura quando implicar em um abuso será utilizada a expressão de censura negativa.

2.2.1 Conceito e tipos de moderação de conteúdo

Inicialmente, é mister explicitar e reforçar o já exposto sobre o que venha ser a moderação de conteúdo, apresentando uma conceituação e as técnicas operacionais de moderação utilizadas pelas plataformas. Dessa forma, nesse título busca-se apontar as técnicas de moderação que as plataformas exercem na organização do fluxo de conteúdo dos autores para os leitores.

Retomando algumas ideias já apresentadas no corpo da presente pesquisa, as redes sociais possuem um serviço indispensável e intrínseco a sua atividade que importa na moderação do conteúdo publicado nas plataformas. Essa atividade convém tanto para a proteção dos usuários de conteúdos problemáticos, quanto para a organização do conteúdo que o usuário irá consumir.

No que tange à proteção dos usuários, a moderação importa um processo indispensável em um ambiente público como uma rede social, a fim de controlar conteúdos que possuem cunho questionável, isto é, publicações que contenham conteúdo pornográfico, violento, abusivo e até ilícito.

Em relação à organização do conteúdo apresentado ao usuário, a moderação é central na determinação do que vai ser mostrado ou não para o usuário da rede. Vejamos o exemplo do Feed de Notícias do Facebook. O Feed de Notícias é uma lista de histórias em constante atualização de status, fotos, vídeos, links, atividade do aplicativo e curtidas de pessoas, Páginas e grupos (FACEBOOK, 2021c, online).

Portanto, a moderação de conteúdo compreende-se como uma atividade das plataformas de aplicação de seleção de todo o conteúdo publicado pelos usuários e marcação desse conteúdo como o que pode ou não estar no seu ambiente digital, consoante às suas regras, com o intuito de proteger os usuários e prevenir abusos.

Nessa esteira, a fim de compreender como o Facebook modera o conteúdo publicado é oportuno destacar 6 técnicas de operacionalização da moderação, especificamente na organização do conteúdo dos autores para os leitores, das quais: *deletion, editing, annotation, synthesis, filtering e formatting* (GRIMMELMANN, 2015, p. 58-59).

A primeira técnica de organização do conteúdo dá-se pela exclusão [*deletion*] que funciona removendo a publicação contrária às regras impostas pela plataforma. A técnica também pode ser aplicada nos serviços da plataforma, a exemplo do conteúdo efêmero, o qual a exclusão do conteúdo faz parte do sistema, como nas fotos enviadas no Snapchat e Instagram que desaparecem em segundos (GRIMMELMANN, 2015, p. 59).

Na sequência, tem-se a técnica de edição [*editing*], de modo que possibilita a alteração do conteúdo, objetivando corrigir erros de digitação alterando a essência de um post, isto é, a possibilidade de o moderador rejeitar a verdade do autor e substituí-las pela dele próprio (GRIMMELMANN, 2015, p. 59).

Outra operação de moderação é anotação [*annotation*], a qual se traduz pela adição de informações ou por comentários no conteúdo. Exemplifica-se essa operação pelo sistema de *feedback* e de avaliação dos compradores e vendedores nas plataformas de e-commerce, bem como os Likes e Comentários no Facebook (GRIMMELMANN, 2015, p. 59).

Por seguinte, a síntese [*synthesis*] é a combinação de transformação de conteúdo. O Wikipédia é o principal exemplo dessa organização de fluxo, uma vez que os usuários podem promover pequenas e heterogêneas mudanças sintetizadas em uma enciclopédia digital (GRIMMELMANN, 2015, p. 59).

A filtragem de conteúdo [*filtering*] é uma ferramenta não destrutiva como a exclusão, mas que pode resultar no mesmo efeito prático com a inclusão de algoritmos classificador do conteúdo que vai ser apresentado ao usuário. A operacionalização desse fluxo, portanto, sustenta-se como um motor de busca que filtra o conteúdo e apresenta apenas os mais relevantes aos leitores (GRIMMELMANN, 2015, p. 59).

Finalmente, a formatação [*formatting*] é o modelo de organização de conteúdo que apresenta uma tipografia melhorada ao exibir informações aos leitores, tornando a legibilidade ordenada e rápida (GRIMMELMANN, 2015, p. 59).

Em suma, a filtragem oculta conteúdo indesejado, a exclusão o remove imediatamente, a anotação permite que os leitores avaliem a relevância do conteúdo para eles e a síntese transforma várias contribuições de valor moderado em um de maior valor (GRIMMELMANN, 2015, p. 60).

A partir desses fatores, a moderação do fluxo de conteúdo auxilia os leitores a ver apenas o conteúdo que possuem preferência e ocultar os menos interessantes ou ilegais, sob o pálio de melhorar a experiência dos usuários ao utilizarem a rede social.

Antes de iniciar o próximo tópico, impende referenciar que o processo de moderação de conteúdo pelas plataformas ocorre em diversas escalas. Em uma escala menor, a moderação é feita manualmente por funcionários com uso limitado de tecnologias ao seu auxílio, identificada como um processo artesanal. Todavia, para empresas que necessitam grande escala de operações de moderação de conteúdo global utilizam tecnologias automatizadas, por um processo industrial (CAPLAN, 2018, p. 17 e 23).

Todos os dias, bilhões de postagens aparecem no Facebook, o que inviabiliza a moderação de conteúdo por processo manual, necessitando, assim, da implementação de sistemas algoritmos de inteligência artificial³ (IA) aliado à instantaneidade do mundo digital. Pontos que serão retomados ao longo da pesquisa.

Assim, a moderação de conteúdo pode acontecer antes de o conteúdo ser realmente publicado, moderação *ex ante*, ou depois que o conteúdo é publicado, moderação *ex post*. A moderação de conteúdo *ex ante* é o processo que acontece entre *upload* e publicação do conteúdo, essencialmente um processo automático executado por meio de triagem algorítmica sem o uso ativo humano tomando uma decisão (KLONICK, 2018, p. 1636).

Já a moderação *ex post* pode ser utilizada pelas plataformas como punição para maus comportamentos online e correção de erros (GRIMMELMANN, 2015, p. 67-68). A verificação posterior possui duas posturas de operação da moderação de conteúdo: proativa [*ex post proactive*] e reacionária [*ex post reactive*] (KLONICK, 2018, p. 1638).

³ Por Inteligência Artificial, filia-se ao entendimento: A inteligência artificial (IA) é um ramo da ciência da computação que busca, com interação multidisciplinar com outras áreas do conhecimento, a reprodução de ações cognitivas tipicamente humanas. Para tanto, a IA pode valer-se de diversas técnicas como estratégia de incremento de performance ou simplesmente de delegação de funções enfadonhas, repetitivas ou consideradas delegáveis e roboticamente praticáveis. Em outras palavras, a IA está associada à reprodução artificial da capacidade de organizar informações para uma solução aceitável de um problema (HARTMANN PEIXOTO, 2020, p. 17).

A moderação de conteúdo próativa implica nas plataformas buscando e removendo proativamente o conteúdo publicado, comumente desenvolvido com o auxílio de tecnologias de algoritmos. Por outro lado, a moderação posterior reacionária decorre da sinalização do conteúdo como potencialmente abusivo por outros usuários, resultando na revisão do conteúdo (KLONICK, 2018, p. 1638).

Todavia, alerta-se que essas ferramentas podem se tornar mecanismos de manipulação nas mãos de moderadores não engajados com as regras das plataformas e a ética ou pela operação de inteligências artificiais comprometidas. Assim, cumpre apresentar como se sucede a moderação de conteúdo pelo Facebook.

2.2.2 Processo de moderação pelo Facebook

Superado o tópico acerca da explicitação do que venha ser moderação de conteúdo e as técnicas utilizadas nesse processo pelas plataformas, cumpre explorar nesse título a operacionalização da moderação de conteúdo pelo Facebook, em específico, desde o Feed de Notícias à remoção das publicações, conforme diretrizes disponibilizadas pelo Facebook.

Inicialmente, ressalta-se que diariamente bilhões de conteúdos são publicados na plataforma incorrendo na tarefa do moderador em decidir o que fica ou não online. Frente a isso, sistemas de Inteligência Artificial (IA) são implementados para automatizar a triagem para identificar e remover conteúdo prejudicial (BARRET, 2020, p. 3).

A fim de diminuir o trabalho dos moderadores humanos, o Facebook tem investido em tecnologia de classificação com aprendizado de máquina [*machine learning*]⁴, isto é, em uma inteligência artificial que detecta automaticamente e proativamente conteúdo que vai contra as diretrizes da rede social (FACEBOOK, 2021d, online).

⁴ Define-se *machine learning*: Conjunto de métodos para identificar padrões em dados de forma autônoma e realizar projeções ou encaminhar uma ação. O Aprendizado de máquina associa-se a obtenção de volumes de dados elevados processados para a detectar padrões, a fim de realizar predição e recomendação de ações peculiares da atividade cognitiva humana (HARTMANN PEIXOTO, 2020, p.18).

Apenas um adendo, a aplicação de sistemas inteligente pelo Facebook abrange também outras atividades da plataforma, como o Feed de Notícias. A determinação do que vai ser apresentado no Feed de Notícias e a ordem de apresentação é disposta por um fluxo personalizado e diversificado de publicações classificadas por ordem de relevância dado pela atuação do algoritmo, denominado de *EdgeRank* (FACEBOOK, 2021c, online).

Alguns fatores influenciam a classificação do conteúdo pelo algoritmo, tais como: a frequência de interação com publicações de amigos, grupos ou Páginas (amigos e familiares têm prioridade); o tipo de publicação (foto, vídeo ou link); a quantidade de comentários, curtidas, reações e compartilhamentos que uma publicação; e quanto tempo faz que foi publicado (FACEBOOK, 2021c, online).

O processo de moderação de conteúdo pelo Facebook tem como estratégia gerencial “remover, reduzir e informar”. Isto é, a plataforma tem como preceito remover o conteúdo prejudicial contrários às políticas da rede social, reduzir a distribuição de conteúdo problemático que não viola as políticas do Facebook e informar as pessoas com contexto adicional para que possa decidir o que clicar, ler ou compartilhar (FACEBOOK, 2021d, online).

Em colaboração com especialistas globais em tecnologia, segurança pública e direitos humanos, o Facebook cria e atualiza suas políticas, com o intuito de construir recursos de segurança, para que os usuários possam relatar conteúdo e bloquear, ocultar ou deixar de seguir contas, por meio de ferramentas tecnológicas e recursos humanos (FACEBOOK, 2021d, online).

Nesta senda, a detecção de violações na rede social importa no esforço tecnológico de inteligência artificial do Facebook associado a equipes de revisão. Essa tecnologia do Facebook detecta e remove proativamente a grande maioria dos conteúdos violadores antes que os usuários os denunciem (FACEBOOK, 2021d, online).

O Facebook dispõe de diversas equipes de tecnologias que constroem e treinam inteligências artificiais para realizar a moderação de conteúdo na plataforma nas diversas áreas de conteúdo das diretrizes da rede social. Os sistemas de IA são dotados de aprendizado de máquinas, os quais conseguem aplicar as medidas de

moderação por meio de previsões e adaptações sobre os conteúdos, de acordo com a revisão da atividade (FACEBOOK, 2021d, online).

Nesse ponto convém destacar que a plataforma disponibiliza um sistema de revisão e marcação de conteúdo potencialmente lesivo ou abusivo para a utilização dos próprios usuários da rede social. A ferramenta de denúncia de conteúdo é disponibilizada ao lado do conteúdo publicado e implica a análise da denúncia e do conteúdo pelo Facebook contra possíveis violações das regras da rede social, servindo, portanto, como um sistema de *feedback* e alerta para a atividade de moderação do conteúdo (FACEBOOK, 2021c, online).

No que tange à estratégia "remover, reduzir, informar" cumpre imergir nas três abordagens do Facebook em relação a conteúdos moderados na plataforma atualizadas na plataforma de Transparência do Facebook. Em relação à primeira parte, remover, o Facebook possui regras que descrevem o que é e o que não é permitido na plataforma, assim, todo o conteúdo contrário a essas políticas são removidos (FACEBOOK, 2021d, online).

Apenas um adendo, a exemplificação do processo de moderação e representação dessa estratégia será posteriormente desenvolvido, especificamente no título 2.3 do trabalho, instituído para apresentação dos problemas decorrentes da moderação e os exemplos dessa atividade.

Após a identificação de um conteúdo violador na rede social, procede-se à notificação do usuário acerca do motivo da remoção a fim de evitar, no futuro, a postagem de outro conteúdo violador. Outrossim, o Facebook faz uso de um sistema de avisos para contar as violações de cada usuário para proceder sua responsabilização (FACEBOOK, 2021d, online).

Seguindo na estratégia "remover, reduzir, informar" do Facebook, a incidência de "reduzir" na moderação destina-se a conteúdos problemáticos que não se enquadram como caso de remoção. Portanto, se o conteúdo publicado "não violar os Padrões da comunidade do Facebook, mas ainda puder ser problemático ou de baixa qualidade, a empresa do Facebook reduzirá sua distribuição" aos usuários da rede social (FACEBOOK, 2021d, online).

O trabalho de redução foca-se em conteúdo problemático tanto no Feed de notícias quanto em comentários de publicações, com o intuito de promover segurança

e evitar fazer recomendações que possam ser de baixa qualidade, questionáveis, delicadas ou inadequadas para espectadores mais jovens (FACEBOOK, 2021d, online).

Contudo, alguns usuários podem querer ver esses conteúdos, mas outros podem não querer por achar problemático. Desse modo, conteúdos controversos, como desinformação e conteúdo de criadores que violam repetidamente as regras, ficam difíceis de encontrar na rede social devido a essa redução (FACEBOOK, 2021d, online).

De igual forma, a redução se aplica no sistema recomendação de conteúdo e contas, páginas, grupos e eventos que as pessoas ainda não seguem. A recomendação do Facebook está nos serviços de “Páginas que você pode gostar”, “Grupos em que você deve participar”, “Postagens sugeridas para você” e “Pessoas que você talvez conheça” (FACEBOOK, 2021d, online).

Por fim, a estratégia de “informar” representa em um mecanismo de aviso e sinalização de conteúdo potencialmente sensível ou enganoso, bem como adição de informações de verificadores de fatos independentes, a fim de fornecer um contexto sobre o conteúdo antes do usuário ler, confiar ou compartilhar a publicação (FACEBOOK, 2021d, online).

Dentre os mecanismos de avisos e sinalização de conteúdo sensível ou enganoso, destaca-se: telas de alerta sobre conteúdo sensível no Facebook; selos de verificação de páginas ou perfil; e telas de notificação de artigos desatualizados no Facebook (FACEBOOK, 2021d, online).

Com o intuito de representar graficamente o processo de moderação de conteúdo, desenvolveu-se a Figura 1 que sintetiza visualmente a atividade moderadora dentro da plataforma do Facebook. Veja-se:

Figura 1 -Esquema moderação de conteúdo pelo Facebook



Fonte: elaborado pelo autor, 2021.

Sintetizando o processo, quando uma publicação é postada na rede social o sistema moderador automaticamente realiza a classificação do conteúdo se viola ou não as regras e, assim, remover o conteúdo violador. Outrossim, se antes desse processo um usuário proceder a denúncia de uma determinada publicação, passa por uma revisão de um moderador, o qual decidirá pela permanência ou retirada do conteúdo.

Por fim, nos casos em que o conteúdo é removido da plataforma, tanto por IA ou revisão humana, o usuário que teve seu conteúdo oculto poderá solicitar revisão da decisão. Após a análise final, o conteúdo problemático poderá permanecer retirado da rede social ou, quando verificado equívoco, poderá retornar à disponibilidade.

Em suma, a moderação de conteúdo pelo Facebook ocorre sobre o conteúdo que circula dentro da plataforma, seja publicações ou comentários, tanto por perfis quanto por páginas, no Feed de notícias, em grupos ou mensagens privadas. Por conseguinte, resta explorar o sistema de penalidades da plataforma frente a violação de conteúdo.

2.2.3 Das violações na plataforma

Conforme abordado ao longo do trabalho, a moderação de conteúdo funciona como um filtro de verificação do conteúdo publicado com as regras de conteúdo permitido pela plataforma. Durante a atividade moderadora, seja por IA ou humana, ao detectar-se conteúdo violador das regras pode-se aplicar sanções aos usuários. Posto isso, neste título será abordado o que acontece com o usuário que violar as regras do Facebook.

A responsabilização do usuário poderá ser pela pena de restrição ou desativação da conta, de acordo com o histórico de violações e avisos e com qual política de conteúdo houve a violação. Não obstante, a plataforma oferece aos usuários a possibilidade de solicitar uma análise da decisão de remoção caso o usuário se oponha, através da opção “Discordar da decisão” (FACEBOOK, 2021d, online).

As violações, denominadas de *Strikes*, são computadas em sistema especificado pelo Facebook na Central de Transparência da plataforma, na aba de aplicação dos Padrões da Comunidade, o qual:

Para garantir que nosso sistema de avisos seja justo e proporcional, não contamos avisos sobre violações de conteúdo postado há mais de 90 dias para a maioria das violações ou há mais de 4 anos para violações mais graves. Também não contamos avisos para certas violações de política. Isso inclui quando alguém compartilha suas próprias informações financeiras, que removemos para evitar fraude, ou casos em que temos um contexto extra sobre a natureza da violação.

Se removermos vários trechos de conteúdo de uma vez, sem notificá-lo a cada remoção, também poderemos contá-los como um único aviso. Todas as advertências no Facebook ou Instagram expiram após um ano. (Tradução livre. FACEBOOK, 2021d, online).

O sistema de *strikes* não apenas se aplica a contas dos usuários, como também para páginas ou grupos no Facebook e implica diferentes sanções de acordo com o histórico de violações registradas anteriormente. Para a maioria das violações, no primeiro *strike* o usuário receberá um aviso, sem outras restrições, salvo quando a violação for gravosa a ponto de exigir uma medida por parte da rede social (FACEBOOK, 2021d, online).

Entretanto, havendo mais *strikes*, os avisos levarão às seguintes restrições padrões:

Tabela 1 - Sistema de strikes

Sistema padrão de <i>strikes</i> por violação de conteúdo no Facebook
Um <i>strike</i>: Aviso, sem outras restrições.
2 <i>strikes</i>: Restrição de um dia para criar conteúdo - como postar, comentar, usar o Facebook Live ou criar uma página.
3 <i>strikes</i>: restrição de 3 dias para a criação de conteúdo.
4 <i>strikes</i>: restrição de 7 dias para a criação de conteúdo.
5 ou mais <i>strikes</i>: restrição de 30 dias para a criação de conteúdo.

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Caso o conteúdo postado for contra às políticas mais severas do Facebook, a exemplo de exploração sexual infantil ou organização perigosa, o usuário/página/grupo infrator poderá receber restrições adicionais e mais longas de certos recursos, além das restrições padrão acima, como por exemplo, impedimento de criar anúncios por determinados períodos de tempo, bem como restrição de monetização e remoção das recomendações (FACEBOOK, 2021d, online).

Em relação às restrições adicionais, as contas dos usuários podem ser obstadas a criar conteúdo, como postar, comentar, usar o Facebook Live ou criar uma página, por um a 30 dias, bem como pode-se também desativar contas por violar repetidamente os Padrões da comunidade ou por certas violações muito graves (FACEBOOK, 2021d, online).

Retomando a ideia de que as redes sociais são um espaço público de interações sociais, as personalidades públicas também fazem parte desse ambiente e também estão sujeitas à moderação. Entretanto, contas de figuras públicas possuem regras de violações específicas as quais são consideradas o engajamento, alcance e influência da exposição, em razão de representarem um risco maior de danos ao violar as políticas.

A definição de quem se enquadra como figura pública no Facebook representa “funcionários do governo estadual e nacional, candidatos políticos para esses cargos, pessoas com mais de um milhão de fãs ou seguidores nas redes sociais e pessoas que recebem cobertura noticiosa substancial” (Tradução livre. FACEBOOK, 2021d, online).

Outro aspecto a ser sublinhado, importa na restrição de contas de figuras públicas durante distúrbios civis de modo a incitar ou promover eventos violentos ou agitação civil. De tal modo, a duração da restrição imposta pelo Facebook pode variar de um mês a 2 anos a critérios da empresa, conforme o grau de violação (FACEBOOK, 2021d, online).

Tendo em conta tal cenário, considera-se para determinar a restrição: gravidade da violação e o histórico de violações da personalidade pública no Facebook; a potencial influência da figura pública e o relacionamento com os envolvidos na violência; e a gravidade da violência e qualquer dano físico conexo, bem como o risco geral para a segurança pública (FACEBOOK, 2021d, online).

Por fim, a desativação da conta de um usuário poderá ocorrer em duas situações. A primeira situação se traduz no conhecimento pela plataforma de contas de indivíduos perigosos, criminosos sexuais condenados, contas de identidade falsa e contas criadas com o intuito de desviar os fins da rede social (FACEBOOK, 2021d, online).

Já a segunda situação decorre de reiteradas violações contra as regras da rede social. Ou seja, a empresa do Facebook poderá desativar a conta do usuário após 5 *strikes*, a depender ainda da gravidade e frequência das infrações. Em algumas situações, a violação pode ser grave o bastante para a conta seja desabilitada após uma ocorrência, tal como postar conteúdo de exploração sexual infantil (FACEBOOK, 2021d, online).

Em conclusão, da atividade moderadora que verificar violação das regras da rede social implica responsabilização ao usuário, conforme o sistema de avisos e violações exposto e de acordo com o histórico e gravidade do conteúdo. Outrossim, se ocorrido por figura pública os fatores influência e risco à segurança pública são envolvidos.

Não obstante, os dados coletados sobre as violações identificadas pela moderação de conteúdo são processados de acordo com métricas desenvolvidas pela plataforma. Posto isso, é relevante explicitar como os dados são medidos e processados pelo Facebook na elaboração dos relatórios que são disponibilizados e aplicação das medidas de aplicação das regras da plataforma.

A primeira métrica avaliada, prevalência, representa o impacto causado pelo conteúdo violador proporcional aos números de visualização do conteúdo, ou seja, “o número estimado de visualizações que mostraram conteúdo violador dividido pelo número estimado de visualizações de conteúdo total” (Tradução Livre. FACEBOOK, 2021d, online).

A segunda métrica envolvida é do conteúdo acionado, o que resulta em uma escala da atividade de fiscalização do Facebook. Nesse ponto são medidos os números de peças de conteúdo (postagens, fotos, vídeos ou comentários) e contas contrárias às regras (FACEBOOK, 2021d, online). Ademais, no momento em que um conteúdo é moderado recebe um rótulo com a descrição da política que o conteúdo violou.

Por conseguinte, a taxa proativa importa em uma métrica que apresenta a de porcentagem de todo o conteúdo ou contas que a moderação sinaliza como violadora antes que os usuários denunciem (FACEBOOK, 2021d, online). Essa métrica representa um indicador da eficácia da constatação de violações pela inteligência artificial do Facebook.

Posto isso, conforme informações disponibilizadas na Central de Transparência do Facebook, na aba de aplicação de tecnologias na plataforma, observa-se que:

A taxa em que podemos detectar proativamente conteúdo potencialmente violador é alta para algumas violações, o que significa que encontramos e sinalizamos a maior parte do conteúdo antes que os usuários o façam. Isso é especialmente verdadeiro quando conseguimos criar uma tecnologia de

aprendizado de máquina que identifica automaticamente o conteúdo que pode violar nossos padrões.

Essa tecnologia é muito promissora, mas ainda está a anos de ser eficaz para todos os tipos de violações. Por exemplo, ainda existem limitações na capacidade de entender o contexto e as nuances, especialmente para conteúdo baseado em texto. Isso cria desafios adicionais para a detecção proativa de certas violações. (Tradução livre. FACEBOOK, 2021d, online).

Outra métrica considerada refere-se ao conteúdo apelado. Conforme abordado, quando um conteúdo é removido o Facebook possibilita a revisão da decisão moderadora, assim, nessa métrica são computados os números de conteúdos que os usuários apelam das decisões de moderação (FACEBOOK, 2021d, online).

Por fim, a última métrica revela o indicativo de conteúdo restaurado, o qual anteriormente foi removido ou sinalizado como violador pelo Facebook. No conteúdo restaurado envolve não só o conteúdo restaurado após a apelação do usuário, como também a correção quando a própria plataforma identifica um erro na moderação (FACEBOOK, 2021d, online).

As sanções do Facebook poderão ser de aviso, restrição para criar conteúdo, suspensão da conta por prazo determinável e, a mais grave, desativação da conta. Nesta senda, convém apresentar os informativos, conforme as métricas de coleta de dados do Facebook, acerca da atividade de moderação de conteúdo.

2.2.4 Indicativos do Facebook de moderação de conteúdo

Este título tem como alvo apresentar os indicativos da moderação de conteúdo pelo Facebook divulgados no *site* da plataforma. Os dados foram obtidos na plataforma de transparência do Facebook dos relatórios de aplicação dos Padrões da Comunidade. Portanto, cumpre expor os dados da Central de Transparência do Facebook dos relatórios de aplicação dos Padrões da Comunidade⁵.

Trimestralmente o Facebook disponibiliza o relatório de aplicação dos Padrões da Comunidade, o qual divide-se por conteúdos, dentre os quais: nudez e atividade sexual de adultos; bullying e assédio; nudez infantil e exploração sexual de crianças; organizações perigosas; contas falsas; discurso de ódio; bens

⁵ Para acessar o relatório na íntegra: <https://transparency.fb.com/data/community-standards-enforcement/>

regulamentados - drogas e armas de fogo; spam; suicídio e automutilação; e conteúdo violento e gráfico.

A divisão dos trimestres representa: 1º trimestre entre janeiro e março; 2º trimestre entre abril e junho; 3º trimestre entre julho e setembro; e 4º trimestre entre outubro e dezembro. Desse modo, cumpriu-se observar os indicativos da atividade de moderação de conteúdo pelo Facebook em 2021.

De acordo com o relatório de moderação de conteúdo sobre nudez e atividade sexual de adulto, representado pela Tabela 2, os indicativos da atividade do Facebook importam no ano de 2021:

Tabela 2 - Indicativos de moderação de conteúdo - nudez e atividade sexual de adultos

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	31.9M	98.6	1.4	390k	73.2k	560k	633k
2ºtri	32.8M	98.9	1.1	536k	312k	55.9k	368k
3ºtri	34.7M	98.8	1.2	273k	234k	32.9k	267k
4ºtri	27.3M	97.7	2.3	311k	36.7	262k	299k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Quanto ao conteúdo marcado como bullying e assédio pelo sistema de moderação do Facebook, pode-se verificar os números da atividade na Tabela 3:

Tabela 3 - Indicativos de moderação de conteúdo - bullying e assédio

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	8.8M	54.1	45.9	936k	113k	179k	292k
2ºtri	7.9M	54.1	45.9	919k	121k	211k	332k
3ºtri	9.2M	59.4	40.6	1M	150k	283k	433k

4ºtri	8.2M	58.8	41.2	799k	121k	243k	364k
--------------	------	------	------	------	------	------	------

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Por sua vez, os dados de perigo infantil de nudez infantil e exploração sexual de crianças, no relatório de transparência disponibilizado pelo Facebook, representam na Tabela 4 os seguintes indicativos:

Tabela 4 - Indicativos de moderação de conteúdo – perigo infantil

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	5M	98.9	1.1	3.8k	300	46.7	47k
2ºtri	2.3M	97.9	2.1	4k	840	23.9k	24.7k
3ºtri	1.8M	97.1	2.9	3k	730	169.8k	170.8k
4ºtri	1.8M	97.5	2.5	3.7k	800	19.2k	20k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Em relação aos números de organizações perigosas, o relatório divide-se em terrorismo e ódio organizado. Primeiramente, em se tratando de conteúdo marcado como terrorismo, a Tabela 5 simboliza a moderação de conteúdo em 2021 pelo Facebook:

Tabela 5 - Indicativos de moderação de conteúdo - organizações perigosas: terrorismo

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	9M	99.6	0.4	38.8k	6.5k	62.6k	69.7k
2ºtri	7.1M	99.7	0.3	35.9k	3.5k	41.2k	44.7k
3ºtri	9.8M	99.7	0.3	52.3k	8.5k	69k	77.5k
4ºtri	7.7M	97.7	2.3	38.9k	5.2k	52.2k	57.4k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Por sua vez, o conteúdo moderado como ódio organizado, dentro da diretriz de organizações perigosas, evidencia a atividade do Facebook na Tabela 6:

Tabela 6 - Indicativos de moderação de conteúdo - organizações perigosas: ódio organizado

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	9.8M	98.6	1.4	149k	45k	455k	500k
2ºtri	6,2M	97.8	2.2	87.8k	33.7k	485k	518k
3ºtri	2M	96.4	3.6	53.2k	23.8k	433k	457k
4ºtri	1.6M	96.1	3.9	35.7	11.9k	115k	127k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Não obstante, a violação das regras do Facebook mede-se também em relação às contas falsas. A quantidade de contas falsas que a plataforma moderou importa apenas em três métricas representadas na Tabela 7:

Tabela 7 - Indicativos de moderação de conteúdo – contas falsas

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)
1ºtri	1.3B	99.8	0.2
2ºtri	1.7B	99.8	0.2
3ºtri	1.8B	99.8	0.2
4ºtri	1.7B	99.9	0.1

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Em se tratando de conteúdo marcado como discurso de ódio, a plataforma moderou em 2021, conforme apresentado na Tabela 8, o equivalente à:

Tabela 8 - Indicativos de moderação de conteúdo – discurso de ódio

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	25.2M	96.7	3.2	1.1M	84.8k	324k	409k
2ºtri	31.5M	97.6	2.4	1.4M	88.1k	328k	416k
3ºtri	22.3M	96.5	3.5	1.1M	90.7k	303k	394k
4ºtri	17.4M	95.9	4.1	769k	65.3k	228k	293k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

A comercialização de drogas e armas de fogo, por mais que seja regulamentada em diversos Estados, não é permitida no Facebook. Dessa forma, incide-se a moderação contra conteúdos para esse fim. Em relação aos indicativos de moderação de conteúdo de drogas desenhados pela Tabela 9, evidencia-se:

Tabela 9 - Indicativos de moderação de conteúdo – bens regulamentados: drogas

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	3.2M	94.1	5.9	36.4k	14.9k	97.1k	112k
2ºtri	2.3M	93.1	6.9	36.3k	6.1k	59.2k	65.3k
3ºtri	2.7M	96.7	3.3	43.9k	9.3k	83.6k	92.9k
4ºtri	4M	97.9	2.1	80k	27.7k	120k	148k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Ademais, quanto aos indicativos de moderação de conteúdo sobre armas de fogo, reflete-se na Tabela 10:

Tabela 10 - Indicativos de moderação de conteúdo – bens regulamentados: armas de fogo

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	1.9M	94.8	5.2	125k	39.6k	128k	168k
2ºtri	1.5M	94	6	82.7k	20.6k	120k	141k

3ºtri	1.1M	94.1	5.9	44.1k	9k	60.7k	69.7k
4ºtri	1.5M	92	8	63k	10.7k	59.8k	70.5k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Ainda nessa linha, no que tange ao conteúdo marcado como spam, cumpre destacar que os números dessa diretriz oscilam conforme os ataques cibernéticos dentro da plataforma. Assim, no ano de 2021, conforme representação da Tabela 11, foram moderados:

Tabela 11 - Indicativos de moderação de conteúdo – spam

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	919M	99.8	0.2	7.2k	1.3k	38.8M	38.8M
2ºtri	794M	99.7	0.3	14.6k	1.2k	30.6M	30.6M
3ºtri	777M	99.6	0.4	18.6k	1.2k	20.9M	20.9M
4ºtri	1.2B	99.6	0.4	24.6k	2k	54.1M	54.1M

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Conforme aponta o relatório de 2021 do Facebook, descrito na Tabela 12, os números de publicações moderadas por suicídio e automutilação importam:

Tabela 12 - Indicativos de moderação de conteúdo – suicídio e automutilação

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	5.2M	97.7	2.3	900	100	186k	186k
2ºtri	16.8M	99.3	0.7	300	60	280k	280k
3ºtri	8.5M	99	1	200	70	162k	162k
4ºtri	6.1M	98.8	1.2	200	50	95.2k	95.2k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Em atualização do relatório, a partir do terceiro trimestre de 2021, o Facebook acrescentou uma métrica de moderação de conteúdo marcado como violência e incitação. Tal diretriz, representada pela Tabela 13, possui como indicadores iniciais de avaliação de moderação:

Tabela 13 - Indicativos de moderação de conteúdo – violência e incitação

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
3ºtri	13.6M	96.7	3.3	435k	37.9k	210k	247k
4ºtri	12.4M	96.6	3.4	362k	33.9k	199k	233k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

Por fim, o último indicativo do relatório de 2021 diz respeito ao conteúdo violento e gráfico. A Tabela 14 expressa os indicadores da atividade de moderação pelo Facebook, nos termos:

Tabela 14 - Indicativos de moderação de conteúdo – conteúdo violento e gráfico

2021	Total	Moderação ativa (%)	Moderação após Denúncia (%)	Decisões apeladas	Conteúdo restaurado após apelação	Conteúdo restaurado sem apelação	Total de conteúdo restaurado
1ºtri	34M	99.6	0.4	3.7k	1k	10.5k	11.5k
2ºtri	30.1M	99.6	0.4	3.7k	700	12.1k	12.8k
3ºtri	26.6M	99.4	0.6	3.3k	700	7.9k	8.6k
4ºtri	25.2M	99.5	0.5	3.8k	600	5.5k	6.1k

Fonte: Elaborado pelo autor com base em Facebook, 2021d.

À vista do exposto, observa-se que a maioria dos temas, com exceção do *bullying* e assédio, a atividade de moderação de conteúdo possui altos níveis de moderação ativa pelas IA do Facebook e alguns tópicos, como nudez adulta e discurso de ódio, importam índices elevados de restauração de conteúdo e apelações das decisões moderadoras.

Outrossim, sobreleva destacar que os índices de publicações restauradas sem apelação podem ser compreendidos como a revisão da atividade de moderação de conteúdo e, portanto, revisão dos erros cometidos pela moderação pró-ativa, o que em alguns índices importam números elevados de erros.

Contudo, sobreleva notar que da moderação decorrem questões procedimentais e legais. De tal modo, importa explorar as implicações, tanto jurídicas quanto operacionais, que decorrem da atividade de moderação de conteúdo pelas plataformas digitais.

2.3 Implicações da moderação de conteúdo

O processo de moderação de conteúdo pelas redes sociais implica questões prejudiciais técnicas e práticas, bem como procedimentais e jurídicas. Posto isso, é relevante abordar os problemas que decorrem da fortificação social das redes sociais e conseqüentemente da moderação de conteúdo sobre os usuários frente ao fundamento que legitima a moderação de conteúdo pelo Facebook.

Conseqüentemente, impende apresentar alguns casos e exemplos de destaque dos problemas decorrentes da atividade de moderação de conteúdo expostos no título anterior. Os exemplos expostos buscarão demonstrar os equívocos da atividade de moderação na prática, bem como exemplos sucesso dessa atividade pelas plataformas.

Um dos primeiros casos contraditórios acerca da atividade de moderação de conteúdo decorreu do bloqueio de imagens de mães amamentando seus bebês por violarem as regras contra nudez e pornografia no Facebook. Em 2014, criou-se uma campanha online de manifestações por mães com a hashtag #FreeTheNipple [#libereomamilo], o que ensejou em uma pressão ao Facebook, que realizou a alteração nas suas políticas de publicações para que fotos de mães amamentando não sejam bloqueadas por moderadores de conteúdo (G1, 2014, online).

Outro exemplo de moderação de conteúdo que alertou o mundo para a nocividade dessa atividade foi a exclusão da fotografia histórica da Guerra do Vietnã, em 2016. A foto retrata crianças que correm chorando após um ataque aéreo feito pelos Estados Unidos, dentre elas, uma menina de 8 anos de idade correndo despida.

A imagem foi postada por um escritor norueguês no contexto de uma discussão sobre fotografias que mudaram a história de guerras, sendo, portanto, postada para exemplificar a questão. A conta do escritor foi suspensa pelo Facebook e a página do jornal *Aftenposten*, que noticiou o caso e repostou a imagem, também foi suspensa. As contas receberam do Facebook a seguinte mensagem: Qualquer fotografia de pessoas mostrando totalmente as genitálias, nádegas ou seios femininos será removida (PAYÃO, 2016, online).

Após diversas críticas, a rede social verificou o erro e republicou as postagens, justificando-se que a imagem de uma criança nua estaria violando os padrões da rede, podendo até ser qualificada como pornografia infantil. Entretanto, a plataforma reconheceu que a fotografia possuía uma importância histórica e global de estar documentando um momento na história da humanidade (PAYÃO, 2016, online).

Se por um lado houve diversas críticas ao Facebook por removerem conteúdo massivamente, em outra oportunidade criticou-se a falta de moderação em determinados episódios nos quais usuários transmitiam ao vivo pelo Facebook Live⁶ violência explícita.

Em 2019 na Nova Zelândia, um atirador atacou duas mesquitas e tirou a vida de 50 muçulmanos em uma ação terrorista, que ficou conhecido como o massacre de Christchurch. Durante o ataque, o atirador transmitiu em tempo real a chacina pelo Facebook Live, tendo ao vivo 200 visualizações e antes de ser removido o vídeo continha 4 mil visualizações (GRYGIEL, 2019, online).

A empresa alega que a IA de moderação que atua nas transmissões ao vivo não estava treinada para identificar violência explícita e, por isso, não impediu a transmissão do vídeo. Outra justificativa dada foi de que não houve denúncias durante a transmissão ao vivo e a primeira denúncia ocorreu após 12 minutos após de encerrada a transmissão (PAYÃO, 2019, online).

O Facebook divulgou que nas 24 horas seguintes ao massacre, 1,5 milhões de vídeos e imagens foram publicados na plataforma sendo que 1,2 milhões de publicações foram removidas durante o upload, isto é, foram identificadas antes da

⁶ Essa ferramenta do Facebook permite a transmissão de mídia ao vivo pelo usuário, permitindo o acesso de outros usuários em tempo real a transmissão.

carga completa na rede, sendo 300 mil removidas após o upload (GRYGIEL, 2019, online).

Não obstante, a moderação de conteúdo atinge também personalidades públicas em diversos contextos. Posto isso, cumpre apresentar alguns casos envolvendo figuras políticas e o Facebook. Retomando o exposto no título 2.1.3, a eleição presidencial dos Estados Unidos em 2016 foi marcada pelo esquema de manipulação eleitoral e uso político de dados envolvendo o Facebook e a empresa *Cambridge Analytica*, que ajudou Donald Trump a vencer o pleito.

Desde a denúncia e exposição, o Facebook busca “limpar sua imagem” e readquirir a confiança dos usuários, promovendo medidas de segurança dos dados dos usuários. De certo modo, isso influenciou a relação entre Trump e a plataforma, uma vez que decorreram diversas decisões de moderação contra publicações do norte-americano. Dentre as decisões que restringiram as publicações de Donald Trump, durante seu mandato de presidente, convém realizar uma breve linha do tempo e destacar as ações do Facebook por dois motivos: desinformação nas eleições e combate à pandemia da Covid-19.

Durante sua campanha eleitoral para a reeleição ao cargo de presidente dos Estados Unidos, Trump publicou em suas contas nas redes sociais, em especial no Facebook e também no Twitter, mensagens que espalharam desinformação, como teorias de conspiração contra o sistema americano de voto pelo correio, bem como anúncios políticos no Facebook que flagraram discursos de ódio e encorajamento de violência. Em relação à publicação de Trump que citava fraude da votação por correio, o Facebook rotulou a publicação como questionável emitindo um aviso de isenção de responsabilidade dizendo que a possibilidade de fraude eleitoral é remota (CULLIFORD, 2020, online).

Na época, o Facebook tolerou as postagens, implicando críticas e manifestações contra a tolerância da rede, tais como paralisação virtual sobre o problema pelos funcionários da plataforma. O CEO do Facebook, Mark Zuckerberg, manifestou-se, em junho de 2020, no sentido de que a plataforma permitiria o máximo de expressão dos usuários, salvo o iminente risco de dano e perigo explicitados. A decisão de Zuckerberg de permitir ativo certas postagens de Trump fez com que diversos funcionários solicitassem o desligamento da empresa Facebook Inc. (ORTUTAY, 2020, online).

Dias após a declaração de Zuckerberg, a campanha de Trump realizou publicações e anúncios políticos contra adversários políticos vinculando à figura de um triângulo de cabeça para baixo, representando a classificação nazista para os prisioneiros políticos nos campos de concentração. Nessa oportunidade, o Facebook removeu as publicações e anúncios com a justificativa de violarem a política contra o ódio organizado na rede social (O'SULLIVAN, 2020, online).

No que tange ao contexto da pandemia causada pela Covid-19, a plataforma, aliada com as recomendações da Organização Mundial de Saúde, impôs diretrizes de moderação mais inflexíveis devido à circulação de informações falsas sobre a doença, tratamentos e combate.

Em agosto de 2020, o Facebook removeu da plataforma, um vídeo de Trump postado na sua conta de uma entrevista à Fox News, a qual o presidente americano dizia que crianças eram “quase imunes” ao novo coronavírus. A justificativa da remoção da publicação pelo Facebook pautou-se na verificação de que o conteúdo do vídeo continha informações falsas, implicando desinformação nociva sobre a covid (BBC NEWS, 2020, online).

Outro caso no contexto pandêmico se resume na remoção de uma publicação de Trump, na qual ele afirmava que a Covid-19 seria menos letal que uma gripe sazonal, em outubro de 2020. O Facebook retirou o post da rede por violação às regras sobre informações falsas sobre a Covid-19 (CNN, 2020, online).

Na sequência temporal dos casos envolvendo Trump e Facebook, a eleição presidencial norte-americana declarou derrota a Donald Trump, o que ensejou na alegação pelo candidato de fraude eleitoral. Nas suas redes sociais, Trump se manifestou contra o reconhecimento do pleito e convocou seus apoiadores para protestarem contra a certificação do resultado das eleições realizadas pelo congresso americano.

Convocados por Trump, em 6 de janeiro de 2021, seus apoiadores organizaram manifestações na capital dos Estados Unidos e inflamados pelo discurso de Donald Trump de não aceitar o resultado e incitando a marcha até o congresso. Na oportunidade, ocorreu a sessão de certificação dos votos do Colégio Eleitoral. Por conseguinte, o Capitólio foi invadido por manifestantes pró-Trump e as forças de

segurança conseguiram controlar a situação e retomar o Capitólio após três horas de confronto (G1, 2021a, online).

Nesse período, Trump publicou um discurso nas suas redes sociais pedindo para que os manifestantes se retirassem do lugar pacificamente, mas, novamente, alegou fraude na eleição. Após publicar, a plataforma do Twitter impediu os tweets de Trump de serem compartilhados, receber curtidas ou serem respondidos, incluindo também tarjas indicando que as publicações continham desinformação sobre o processo eleitoral. Horas depois, como medida, o Facebook removeu o vídeo do discurso da plataforma e afirmou que a decisão foi tomada frente ao risco material de violência (LAVADO, 2021, online).

Decorrido esse episódio, o Facebook advertiu a conta de Trump por duas violações de termos de uso da rede social e bloqueou a conta por 24 horas, o que impediu de realizar novas postagens durante a punição. Antes mesmo de terminar o prazo da suspensão, a plataforma estendeu indefinitivamente o bloqueio nas contas do Trump no Facebook e Instagram, banindo o chefe de Estado dos Estados Unidos da plataforma (CHARLEAUX, 2021, online).

Seguindo o objetivo deste título em apresentar exemplos de moderação pelo Facebook, importa referenciar a moderação no contexto pandêmico. Conforme exposto, a pandemia da Covid-19 foi marcada pela profusão de notícias e informações falsas sobre a doença, os possíveis tratamentos e as vacinas disseminadas, principalmente, nas redes sociais. Assim, o Facebook instituiu regras de moderação mais rígidas quando se verifica um conteúdo contrário às recomendações oficiais de enfrentamento da pandemia.

O resultado disto implicou na remoção de diversas postagens, como as do presidente Donald Trump. Contudo, o presidente americano não foi a única personalidade política a ter publicações removidas por conterem desinformação sobre a Covid-19. No Brasil, o presidente Jair Bolsonaro, em março de 2020, postou em suas redes sociais um vídeo discursando contra as medidas de enfrentamento da pandemia recomendadas pela Organização Mundial da Saúde.

O referido vídeo foi removido das contas oficiais do presidente nas plataformas do Facebook, Instagram e Twitter. A decisão do Facebook fundamentou-se na violação dos Padrões da Comunidade em não permitir desinformação que possa

potencialmente causar danos reais às pessoas em razão da Covid-19 (SENRA, 2020, online).

Ainda no Brasil, em junho de 2021, o deputado federal, Eduardo Bolsonaro, publicou em sua conta um vídeo de seu pai afirmando que metade das mortes ocorridas por Covid-19 no Brasil não haviam ocorrido efetivamente por causa do vírus, implicando em uma afirmação falsa. O deputado teve acesso restringido à sua conta no Facebook por 7 dias (TILT, 2021, online).

Posteriormente, em julho de 2021, Eduardo Bolsonaro publicou na sua conta frases creditadas a Adolf Hitler o que resultou na exclusão do conteúdo pelo Facebook, bem como na aplicação da sanção de suspensão da conta por 30 dias, haja vista o histórico de violações das regras da rede social. Todavia, o Facebook revisou sua decisão, restaurou a publicação e retirou a suspensão na conta justificando-se que a publicação foi removida indevidamente (TILT, 2021, online).

Nesta senda, ao passo que as plataformas de comunicação crescem o caos e a discórdia crescem paralelamente. Isso acontece em razão de que os usuários possuem a necessidade de se expressarem, seja de forma inspiradora ou repreensiva, desde que “eu possa dizer onde outros vão me ouvir” (GILLESPIE, 2018, p. 5). Assim, cumpre abordar as implicações que decorrem da atividade de moderação de conteúdo.

2.3.1 Questões tecnológicas da moderação

Considerando que o que é publicado nas redes sociais exerce influência tanto nas relações dentro quanto fora do ambiente digital, cumpre realizar o alerta acerca das implicações tecnológicas e éticas que decorrem da utilização das redes sociais e da atividade de moderação de conteúdo.

Inicialmente serão abordadas as implicações técnicas, das quais consideram-se a moderação em massa por IA em relação à moderação pelo ser humano. Considera-se como implicações éticas e tecnológicas o enviesamento algoritmo, opacidade do algoritmo e nuances. Em relação às implicações de ordem ética, leva-se em conta a falta de implementação de preceitos éticos na tecnologia,

vigilância de dados provenientes do uso das redes sociais e a governança nas redes sociais.

Considerando que o que é publicado nas redes sociais exerce influência tanto nas relações dentro quanto fora do ambiente digital, cumpre realizar o alerta acerca das questões éticas do implemento de ferramentas automatizadas, bem como dos problemas que decorrem da não observância de preceitos éticos, tais como enviesamento e opacidade dos algoritmos.

Desta feita, “o emprego de ferramentas de inteligência artificial que interagem por meio de mídias sociais exige que se reflita sobre os requisitos éticos que devem acompanhar o desenvolvimento desse tipo de tecnologia” (MAGRANI, 2019, p. 220). Assim, a ética digital pode ser compreendida pela a ética dos dados, a ética de algoritmos e a ética das práticas.

A ética dos dados considera a “geração, gravação, curadoria, processamento, disseminação, compartilhamento e uso de dados” e as implicações morais decorrentes da coleta, análise e aplicação dos dados (CARINI; MORAIS, 2020, p. 5). No que tange à ética de algoritmos, considera-se a intenção e influência nos processos de designer algorítmico de qualquer ator, grupo, instituição ou agência que atue em qualquer fase desses processos (CARINI; MORAIS, 2020, p. 5), de modo a prevalecer a responsabilidade algorítmica (MAGRANI, 2018, p. 160).

Por sua vez, a ética das práticas está sob o comando das práticas de governança, o que permite observar a natureza constitutiva dos problemas decorrentes da aplicação algorítmica pelo conhecimento e subjetividades, provenientes da própria natureza do algoritmo (MAGRANI, 2018, p. 159).

Portanto, deve-se considerar na programação e implementação tecnológica, em especial nos sistemas de IA utilizados na moderação do conteúdo pelas redes sociais, uma “espécie de governador ético, que deverá reproduzir as normas éticas, legais e padrões técnicos predeterminados pela sociedade ou pelo país onde serão utilizados” (CARINI; MORAIS, 2020, p. 21).

A inobservância de preceitos éticos no processo de designer dos algoritmos e da inteligência artificial podem carregar distorções no resultado da aplicação decorrentes da elaboração pelo programador. Por refletir os padrões do programador, esses sistemas podem conter vieses [*bias*].

A discriminação algorítmica é um fenômeno ligado a falta de submissão aos princípios éticos na programação e implementação de tecnologias que resulta em tecnologia enviesadas, ecoando nas dimensões sociais, econômicas e jurídicas. A discriminação, *per si*, decorre do “fato de alguém pertencer a determinado grupo e ser julgado a partir das características desse grupo” tanto por afirmações generalizadoras quanto por afirmações estatisticamente inconsistentes (MENDES; MATTIUZZO, 2019, p. 47e 51).

As características aparentemente neutras e os dados-alvo determinados por um método estatístico explicitam o fenômeno da discriminação estatística ao condescender em tratamentos diferenciados com base em características pessoais relevantes para a tomada de decisões algorítmica (MENDES; MATTIUZZO, 2019, p. 53).

Nessa esteira, sobreleva ressaltar que a diferença entre o comportamento artificial e o humano consiste no fato de que “o resultado da atividade cognitiva artificial pode ser mais facilmente corrigida que os desvios e preconceitos do próprio ser humano” (HARTMANN PEIXOTO, 2020, p. 28).

Por conseguinte, tendo em vista que os dados são fundamentais para os algoritmos resta oportuno alertar acerca do tratamento legal imposto aos dados, visto que devem ser legítimos, corretos e livres de preconceitos, “isto porque métodos de lidar com os dados podem gerar discriminação e resultados tendenciosos” (MAGRANI, 2019, p. 205).

Feitas tais considerações, a coleta ilegal de dados e a vigilância de dados são fenômenos decorrentes do capitalismo do século XXI que encontrou uma nova e massiva matéria-prima para se apropriar: dados (SRNICEK, 2017, p. 49). O capitalismo de dados [*data capitalism*] ou capitalismo de vigilância [*surveillance capitalism*] fundam-se pela coleta incessável de dados, com o intuito de monetização.

Assim, as plataformas, especialmente as de redes sociais em razão da adesão, tornaram-se uma forma cada vez mais “dominante de organizar negócios de forma a monopolizar esses dados e, em seguida, extraí-los, analisá-los, usá-los e vendê-los” (SRNICEK, 2017, p. 49).

Desse modo, a monetização dos dados obtidos na internet é denominada de Capitalismo de vigilância [*Surveillance capitalism*], no qual é revivido a imagem de

Karl Marx do capitalismo como um vampiro que se alimenta do trabalho, mas ao invés de trabalho, o capitalismo de vigilância se alimenta de todos os aspectos da experiência de cada ser humano (ZUBOFF, 2019, p. 16).

O capitalismo de vigilância trata a experiência humana como material para tradução em dados comportamentais declarado como propriedade comportamental para a produção de previsões que antecipam comportamento os do que você fará agora, em breve e depois. Esse produto de previsão é negociado em um mercado de previsões comportamentais, no qual nós somos a fonte do capitalismo de vigilância, ou seja, o ser humano se tornou o objeto de matéria prima das empresas que comercializam o comportamento futuro (ZUBOFF, 2019, p. 15-16).

Em relação às implicações de ordem técnica, observa-se na aplicação algorítmica em geral, o próximo problema tem íntima relação com a moderação de conteúdo, qual seja a remoção de conteúdo em massa formalizada pela utilização de inteligência artificial.

Conforme exposto anteriormente, diariamente são publicados bilhões de novos conteúdos nas redes sociais, devido à crescente das plataformas. Ao passo que as redes sociais foram crescendo, houve a implementação de sistemas de inteligência artificial capazes de auxiliar o trabalho humano. Todavia, esses sistemas moderam atualmente todo o conteúdo que circula nas plataformas, o que gera uma crescente tendência pela remoção em massa de conteúdo pela IA.

Um estudo realizado pela Universidade de Nova Iorque aponta que os sistemas de IA tendem a remover muito do conteúdo publicado marcado como spam e nudez. Entretanto, para outras classificações de conteúdos, tais como discurso de ódio e assédio, não atendem positivamente, dado que requerem, por vezes, um discernimento humano em relação ao contexto e nuance para a compreensão do conteúdo publicado (BARRET, 2020, p. 3).

Por seu turno, o problema da opacidade dos algoritmos pode ser compreendido como de ordem ética e jurídica. A discriminação algorítmica pode levantar implicações éticas-jurídicas acerca da justiça social e desigualdade, tal como acentuação da discriminação existente de determinados grupos sociais, por mais estaticamente consistente que possa estar o uso de informações sensível na aplicação tecnológica. O que torna “mais difícil para os membros de tais agrupamentos

superarem determinada situação prejudicial” (MENDES; MATTIUZZO, 2019, p. 53-54).

Outrossim, os sistemas de IA podem produzir resultados que externem problemas dogmático-jurídicos eivados pela “opacidade, arbitrariedade de critérios e de conclusões, associada à discricionariedade, à discrepância com direitos fundamentais e outros princípios jurídicos” (HARTMANN PEIXOTO, 2020, p. 28).

O problema da opacidade, portanto, está associado ao esforço e a falta de habilidade humana em “decodificar o resultado gerado pelo algoritmo”, o que “pode criar problemas quando eles são usados para tomar decisões importantes que afetem nossas vidas” (MAGRANI, 2019, p. 201). Ainda, não se pode olvidar que a observâncias dessas questões destacadas servem para a implementação de medidas para o Facebook em promover racionalidade e explicabilidade das decisões de moderação de conteúdo.

Por fim, remonta-se ao problema da falta de transparência das plataformas tanto dos seus sistemas e algoritmos de moderação, quanto dos resultados da atividade moderadora. Sem embargo, a transparência vem sendo o princípio amplamente reconhecido pelas plataformas de modo a implementar a transparência quanto às suas regras, operações e resultados.

2.3.2 Governança na moderação de conteúdo pelo Facebook

A governança compreende o exercício da autoridade, controle e administração do poder. Assim, sobreleva indicar as soluções de governança pelo Facebook na atividade de moderação de conteúdo na sua plataforma. Desse modo, a internalização de princípios na rede social representa uma medida de governança da administração do poder que o Facebook possui sobre os usuários operacionalizados sob a forma de ferramentas tecnológicas do exercício do poder da rede.

Importa destacar que o meio acadêmico comumente apresenta como solução para as implicações destacadas nos títulos anteriores por meio da transparência, uma vez que a partir da evidência das informações permite-se a identificação dos problemas e apurar soluções, constituindo, portanto, o primeiro passo para a governança das redes sociais.

Outrossim, a transparência representa “deixar a informação clara antes da participação do usuário”. Todavia, é preciso que os termos de uso do ambiente digital devam tratar detalhadamente do conteúdo que pode ou não circular na plataforma, bem como o procedimento de toda verificação de violação a essas regras. Assim, “esse documento tem relevância como prova legal do que ficou combinado entre as partes” (PINHEIRO, 2013, p. 197).

Não obstante, caminha-se para fiscalização da atividade moderadora entre um panorama estritamente legal e regulatório e um puramente técnico, através da governança de algoritmos. Exige-se tanto uma regulação interna pelas companhias privadas quanto uma autorregulação por toda a indústria tecnológica, devendo, portanto, baseando-se no interesse público (MAGRANI, 2019, p. 201).

O interesse público firma-se tanto em medidas afirmativas que remonte o ser humano no centro da atividade das redes sociais, principalmente quanto à responsabilidade no designer dos sistemas algoritmos para que possam compreender aspectos éticos e morais da sua tomada de decisão (MENDES; MATTIUZZO, 2019, p. 60).

Consoante ao exposto, em uma tentativa de promover níveis mínimos de governança nas plataformas de redes sociais contra a moderação de conteúdo de forma agressiva, foi editada uma carta de princípios para servir como ponto de partida no diálogo com empresas de tecnologia envolvidas com moderação de conteúdo, como o Facebook, que é signatário da carta.

Os Princípios de Santa Clara [*The Santa Clara Principles*] sobre transparência e responsabilidade na moderação de conteúdo no âmbito digital, buscam delinear níveis mínimos de transparência significativa e responsabilidade pelo respeito dos direitos dos usuários, com o intuito de mitigar o impacto dos conteúdos nocivos propagados na internet (THE SANTA CLARA PRINCIPLES, 2020, online).

O primeiro dos três princípios propostos, denominado *Numbers*, corresponde a publicação dos números de postagens removidas ou contas suspensas por violação às diretrizes de conteúdos da plataforma. O segundo princípio, *Notice*, traduz o dever de notificação do usuário sobre o motivo pelo qual sua conta foi suspensão ou conteúdo removido. E, por fim, o *Appeal*, estabelece que as empresas devem oportunizar mecanismos de recursos aos usuários acerca das decisões de

remoção de conteúdo ou suspensão de contas (THE SANTA CLARA PRINCIPLES, 2020, online).

Congruente com as medidas de governança nas redes sociais, o Facebook desenvolveu uma plataforma digital de transparência, o Centro de Transparência [*Transparency Center*], no qual apresenta as políticas da rede social, as estratégias de execução dessas políticas e os dados da atividade moderadora. O tópico das políticas do Facebook será apresentado no capítulo subsequente.

A execução das políticas do Facebook parte da detecção de violações das regras pelos usuários e, posteriormente, a aplicação das políticas. Para verificar as violações dos termos, a plataforma utiliza mecanismos tecnológicos que detectam e removem proativamente os conteúdos violadores, isto é, o Facebook possui sistemas de inteligência artificial treinados para filtrar todas as publicações na rede e identificar as que possuem conteúdo potencialmente violador (FACEBOOK, 2021d, online).

A plataforma dispõe que na maioria dos casos a identificação da violação se dá de forma simples, considerando que de forma clara pode-se verificar se houve ou não violação. No entanto, determinadas publicações requerem uma análise complexa, passando a ser avaliada por pessoas e não mais IA. Assim, a decisão das equipes de revisão humana baseia-se nas políticas de conteúdo do Facebook (FACEBOOK, 2021d, online).

As decisões humanas que revisam as decisões de moderação da IA são utilizadas para treinar a inteligência artificial moderadora da plataforma, por meio do aprendizado de máquina, a fim de “criar modelos mais específicos de previsões sobre pessoas e conteúdos” (FACEBOOK, 2021d, online).

Outra medida de responsabilização pelo Facebook, pode ser verificada na abordagem da plataforma para questões relativas à eleição e desinformação. Essas medidas foram implementadas para as eleições presidenciais norte-americanas e atualizadas em julho de 2021.

Durante o pleito, o Facebook trabalhou para prevenir interferências na manipulação do debate público dentro da plataforma removendo contas falsas, combatendo a desinformação pela remoção de conteúdo que intenta violência eleitoral e pela redução da distribuição de notícias falsas sobre a eleição, por fim, aumentando

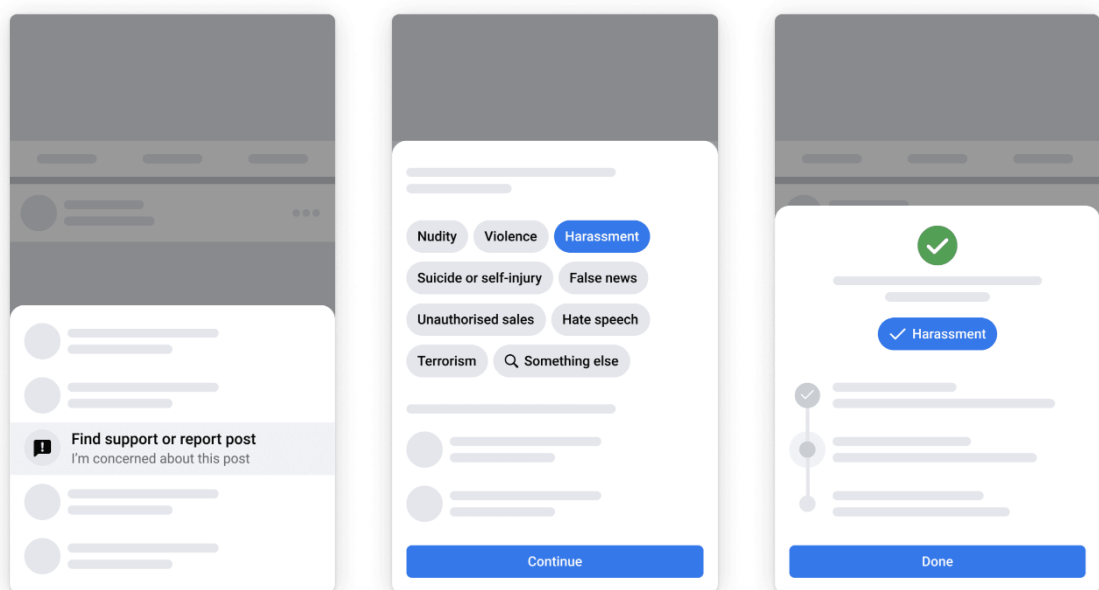
a transparência dos anúncios políticos para que fosse informado o patrocinador (FACEBOOK, 2021d, online).

No que tange às medidas de combate à disseminação de informações incorretas, o Facebook possui como estratégia “remover, reduzir, informar” da moderação de conteúdo pelo Facebook (FACEBOOK, 2021d, online). O Facebook possui ferramentas de monitoramento do conteúdo pelos próprios usuários através da experiência do usuário.

Essas ferramentas representam algumas das medidas de governança institucionalizadas pela plataforma, a fim de possibilitar ao usuário denunciar um conteúdo que entenda estar violando um direito seu ou as regras da plataforma, informar ao usuário os motivos pelos quais sua publicação foi removida da rede social, possibilitar a revisão da decisão moderativa e, ainda, indicar aos usuários que determinados conteúdos são questionáveis quanto à veracidade.

A primeira ferramenta a ser apresentada importa na possibilidade de um usuário denunciar conteúdos que entendam que não deveria estar na plataforma. Essa ferramenta permite uma comunicação direta entre os usuários e a plataforma no controle de conteúdo violador, conforme Figura 2.

Figura 2 - Telas de denúncia de conteúdo.

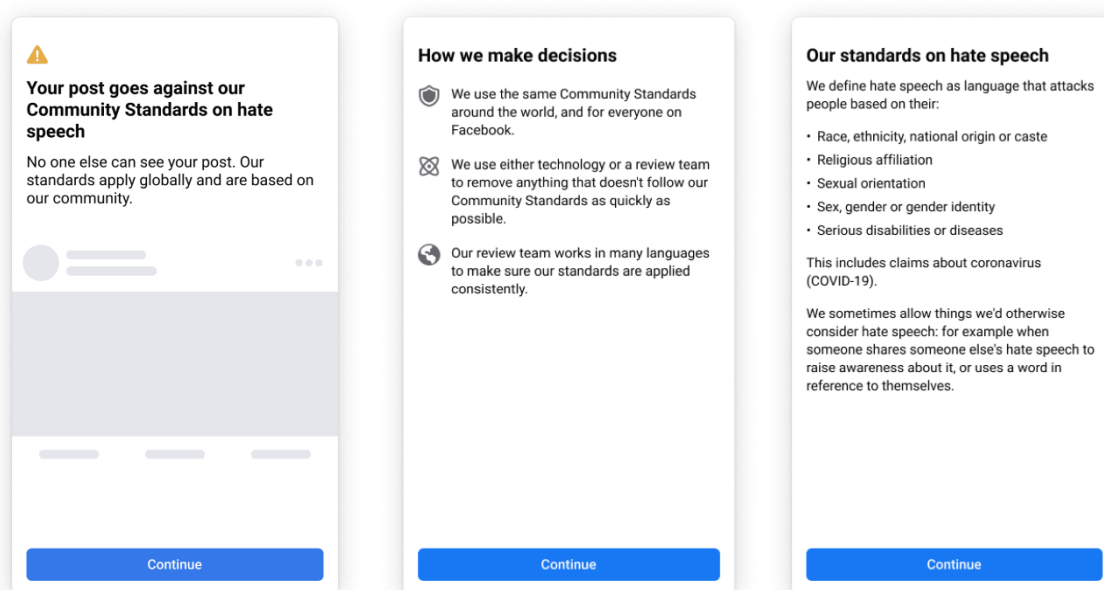


Fonte: Facebook, 2021d.

As telas apresentadas representam o processo de denúncia de uma publicação pelo próprio usuário. Ao realizar a denúncia, o usuário deve informar qual o assunto que o conteúdo estaria violando, a fim de direcionar a análise de conteúdo de acordo com a diretriz violada.

Conforme exposto ao longo do trabalho, a moderação de conteúdo pelo Facebook pode ocorrer após a denúncia de um usuário ou pela moderação ativa da própria plataforma. Quando a moderação resulta na remoção de uma publicação, o usuário receberá a indicação representada na Figura 3:

Figura 3 - Telas de experiência de remoção: notificação.



Fonte: Facebook, 2021d.

Primeiramente, há a notificação do usuário de que sua publicação violaria determinado ponto dos Padrões da Comunidade, com a indicação de como se deu o processo de moderação e apontando as diretrizes de conteúdos que violam as regras da plataforma. Com a notificação, representada na Figura 4, a plataforma permite ao usuário aceitar a decisão de moderação ou discordar:

Figura 4 - Telas de experiência de remoção: revisão da decisão.

What would you like to do?
Because your post goes against our Community Standards on hate speech, no one can see it.

Accept Decision
No one else can see your post ☐

Disagree With Decision
We use your feedback to make improvements to future decisions ☒

Continue

Why do you disagree with the decision?
We'd like to learn more from you.

I think my post does follow the Community Standards ☐

I think Facebook misunderstood the intent or context of my post ☒

I've seen similar content on Facebook ☐

I think this restriction is too harsh ☐

I disagree with the Community Standards ☐

Something else ☐

Continue

You disagreed with the decision

We usually offer the chance to request a review, and follow up if we got decisions wrong. We have fewer reviewers available right now because of the coronavirus (COVID-19) pandemic.

We're trying hard to prioritize reviewing content with the most potential for harm.

This means we may not be able to follow up with you, though your feedback helps us do better in the future.

Thank you for understanding.

[See Community Standards](#)

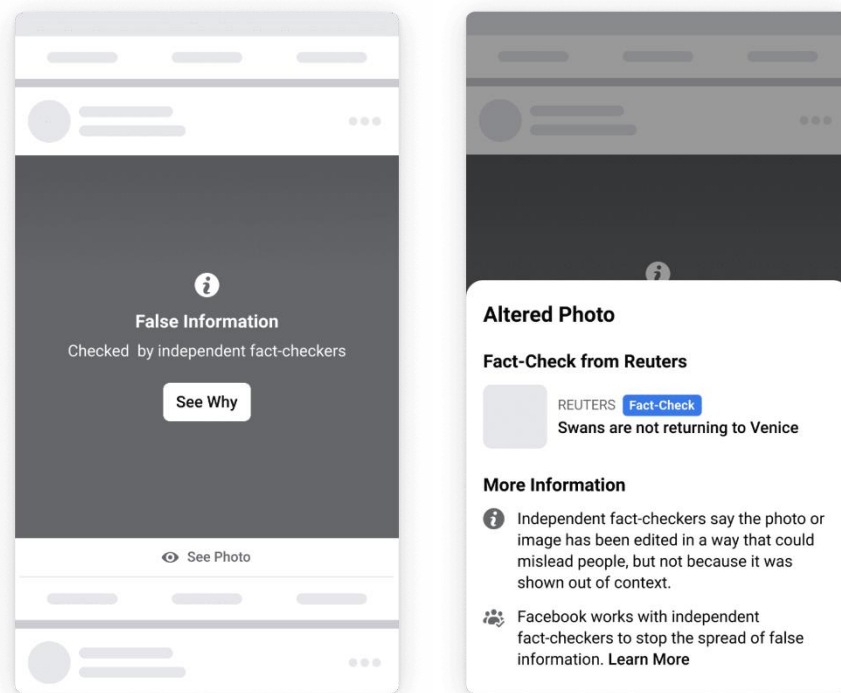
Close

Fonte: Facebook, 2021d.

Não havendo concordância com a decisão de remoção de conteúdo pelo Facebook, o usuário que teve a publicação moderada poderá solicitar revisão da decisão no aplicativo do Facebook, marcando o motivo pelo qual discorda da decisão moderadora. A decisão será revista pelos moderadores de conteúdo do Facebook e mantida ou reformada, quando equivocada.

Outra medida de governança na plataforma diz respeito a conteúdos questionáveis. Quando uma publicação moderada pelo Facebook é marcada como falsa, mas causa risco de dano, a plataforma rotula o *post* e sugere a verificação de mais informações sobre o assunto. Essa medida, na figura 5, representa-se:

Figura 5 - Telas de alerta e indicação de informações adicionais.



Fonte: Facebook, 2021d

Nessa representação pode-se verificar o resultado da atividade moderadora do Facebook pela identificação de uma publicação com conteúdo falso e a indicação para a verificação da informação constante com a checagem de fatos. O programa de checagem de fatos do Facebook trabalha para lidar com informações falsas na rede social e informar os usuários o motivo que o conteúdo foi classificado como falso (FACEBOOK, 2021d, online).

Por fim, uma medida adotada pelo Facebook para a implementação de princípios de governança na internet foi a criação do Comitê de Supervisão [*Oversight Board*], que tem como objetivo auxiliar o Facebook a resolver questões complexas sobre a liberdade de expressão online (FACEBOOK, 2021d, online).

O Comitê de Supervisão é um órgão externo e independente da rede social, que atua tanto como um conselho propondo melhorias e recomendações nas políticas da plataforma, como um tribunal de apelação das decisões de moderação de conteúdo (FACEBOOK, 2021d, online). Os julgamentos do *Oversight Board* são independentes e suas decisões de manter ou reformar as decisões de moderação do

Facebook são vinculantes, devendo, portanto, serem implementadas pelo Facebook (OVERSIGHT BOARD, 2021, online).

Desta feita, as proposições de melhorias e recomendações do Comitê para o Facebook pautam-se no direito à liberdade de expressão exercida no ambiente online, de modo a garantir o respeito a esse direito a partir dos princípios institucionais do *Oversight Board* podendo ser emitidas para melhorar a aplicação das políticas pelo Facebook (OVERSIGHT BOARD, 2021, online).

Dentre os princípios, ressalta-se a independência, a autoridade, a acessibilidade e a transparência. O princípio da independência, institucionalizado pelo Comitê de Supervisão, pressupõe tanto autonomia no julgamento e tomada de decisões, quanto desvinculação administrativa e financeira do Facebook (OVERSIGHT BOARD, 2021, online).

Nessa esteira, o princípio da autoridade compreende-se pelo poder do Comitê em decidir se a rede social deve permitir ou remover o conteúdo, provendo decisões vinculantes. Já o princípio da acessibilidade institucionaliza a possibilidade de o usuário apelar para o Comitê das decisões de moderação do Facebook, sendo-lhe oportunizado momento para se manifestar. Finalmente, o princípio da transparência sujeita o *Oversight Board* a divulgar as declarações sobre suas decisões e justificativas, bem como apresentar relatórios anuais sobre seu trabalho (OVERSIGHT BOARD, 2021, online).

No que tange ao sistema de apelação do *Oversight Board*, a apelação pode ser sobre conteúdo do Facebook ou Instagram. A possibilidade de apelação ao Comitê pressupõe a existência de 4 requisitos: 1) o Facebook ou Instagram tenham analisado a solicitação e emitido sua decisão final; 2) a apelação deve ser realizada pelo titular da conta e a conta deverá estar ativa; 3) a apelação deve ser enviada dentro de 15 dias, a contar da notificação do usuário pelo Facebook ou Instagram da decisão final; e 4) a decisão sobre o conteúdo precisa se qualificar para a apelação.

Dentre os requisitos de admissibilidade da apelação, percebe-se pressupostos de legitimidade, tempestividade e o prequestionamento do conteúdo, típicos de sistemas jurídicos democráticos. Contudo, o quarto requisito, de forma controversa, exige uma qualificação do conteúdo para poder ser apelado, estando à livre indicação da própria rede social de qual decisão poderá ou não ser apelada.

O processo de apelação para o *Oversight Board* possui 5 etapas. A primeira etapa consiste no envio da apelação ao comitê “explicando por que ele deve analisar o caso e por que a decisão do Facebook ou do Instagram sobre o conteúdo está errada”. Na segunda etapa, a seleção, a equipe do Comitê avalia os casos enviados e seleciona os qualificados que sejam difíceis, significativos e tenham relevância global, isto é, o comitê não analisa todos os casos que são submetidos à sua apreciação (OVERSIGHT BOARD, 2021, online).

A terceira etapa, o painel, limita-se à emissão de uma decisão preliminar por um subgrupo do comitê. Já na quarta etapa, constitui-se a decisão final pelo comitê inteiro, podendo rever a decisão preliminar ou firmá-la. Por fim, a quinta etapa consiste na implementação imediata da decisão proferida pelo comitê por parte do Facebook (OVERSIGHT BOARD, 2021, online).

O Comitê de Supervisão idealizado pelo Facebook importa um “tribunal” com o objetivo de desobrigar a responsabilidade das decisões da plataforma sobre moderação de conteúdo pela justificativa de existir um tribunal para que os usuários possam apelar o que representa um avanço nos mecanismos de governança e segurança aos usuários.

Todavia, a utilização de alguns conceitos indeterminados nas regras de procedimentos e justificativas, ainda admite a escolha discricionária das decisões que serão analisadas pelo comitê, bem como na justificação de determinadas decisões.

Portanto, a moderação de conteúdo não conseguirá prevenir todos os casos das patologias da vida online, por isso é necessário estabelecer limites toleráveis, sem, até mesmo, elevar os custos jurídicos-sociais da moderação de conteúdo a um nível inaceitável (GRIMMELMANN, 2015, p. 53).

Desta feita, cumpre exigir das plataformas a implementação de medidas de governança na internet para que os problemas de decorrem da atividade de moderação de conteúdo, considerando ser uma atividade essencial para as redes sociais, seja desenvolvida com o mínimo de impacto negativo aos usuários e a máxima observância dos preceitos de transparência e governança nas redes sociais, firmando o dever de racionalidade, explicabilidade e justificação das decisões moderativas.

2.3.3 Termos de Serviço e Padrões da Comunidade: regras da moderação de conteúdo pelo Facebook

Os Termos do Facebook são compreendidos como o instrumento que legitima a utilização da plataforma pelos usuários, em virtude disso é importante ter uma visão dessas regras que dão todo o suporte para moderação do conteúdo na plataforma. Nesta seção serão apresentadas as regras e diretrizes que orientam o funcionamento da plataforma do Facebook.

Por sua vez, cumpre destacar a forma que se dá a regulamentação do Facebook, pela análise dos Termos de uso e Padrões da Comunidade, como o instrumento que legitima de utilização da plataforma pelos usuários. No presente título, ainda, será abordado as regras que legitimam a moderação de conteúdo pelo Facebook de determinar o que pode ou não ser publicado.

Os termos de uso das redes sociais são tidos por algumas doutrinas como documentos contratuais entre partes privadas que regulam transação do consumidor de trocar o acesso à plataforma com a vinculação aos termos e condições estabelecidas, salvaguardando os interesses econômicos envolvidos (SHENGELIA, 2020, p. 6; SUZOR, 2018, p. 3). De outra forma, tem-se que os termos do Facebook não se manifestam como um contrato tradicional entre partes privadas, dado que se percebe a disposição do Facebook em dar um tom constitucional aos seus termos (CELESTE, 2018, p. 3).

A relação entre o usuário e o Facebook é selada pelo aceite adesivo ou pela outorga dos Termos de Serviço ao realizar o cadastro na plataforma. Fala-se em aceite adesivo e outorga, uma vez que ao usuário não lhe é facultado a possibilidade de discussão dos termos, sendo seu cadastro na rede social vinculado à imposição das regras da plataforma, conforme Figura 6 da tela de cadastro no Facebook:

Figura 6 - Tela de cadastro no Facebook.

Cadastre-se

É rápido e fácil.

Nome

Sobrenome

Celular ou email

Nova senha

Data de nascimento

Dia

Mês

Ano

Gênero

Feminino

Masculino

Personalizado

Ao clicar em Cadastre-se, você concorda com nossos [Termos](#), [Política de Dados](#) e [Política de Cookies](#). Você poderá receber notificações por SMS e cancelar isso quando quiser.

Cadastre-se

Fonte: Facebook.com

Não obstante, caso o usuário discorde dos Termos impostos para a plataforma, o Facebook apresenta como solução a possibilidade de exclusão da conta a qualquer momento.

Os Termos de Serviço importam em diretrizes que regem o uso da plataforma do Facebook, juntamente com outros termos específicos para determinados serviços fornecidos, como Política de Cookies e Política de Dados. Os Termos de Serviço, portanto, representam a descrição dos serviços prestados pelo Facebook e dos compromissos que o usuário deve assumir com a plataforma.

O primeiro ponto a ser analisado dos Termos de Serviço diz respeito à legitimidade de estar na rede social. Sobre quem pode usar o Facebook, e proíbe a utilização da plataforma se: a) a pessoa for menor de 13 anos; b) a pessoa tiver cometido crime sexual; c) a conta tiver sido desativada anteriormente por violações aos Termos e Políticas do Facebook; e d) a pessoa ter sido proibida de receber os serviços do Facebook de acordo com as leis aplicáveis (FACEBOOK, 2020, online).

Primeiramente ressalta-se o fator etário da rede, impedindo que crianças tenham contas e sejam expostas na plataforma. Na sequência, não há justificativa legal para a restrição de acesso por pessoas que cometeram crimes sexuais, frente aos direitos e garantias constitucionais que norteiam o Direito Penal brasileiro.

O terceiro impedimento decorre de uma sanção por violação das regras da plataforma, como visualiza-se no caso do ex-presidente norte-americano, Donald Trump, que foi impedido de criar novas contas no Facebook. Por fim, o último impedimento resulta da tentativa de adequação dos serviços prestados pelo Facebook com as legislações vigentes dos países dos usuários.

Outro ponto substancial para a análise do Facebook representa o controle do conteúdo e utilização do Facebook. Os Termos de Serviço descrevem que o usuário é coibido a usar a plataforma ou compartilhar algo que viole os termos, as políticas e os Padrões da Comunidade do Facebook, tampouco algo que seja ilegal, enganoso, discriminatório ou fraudulento, ou, ainda, algo que viole o direito de terceiros (FACEBOOK, 2020, online).

Outras proibições de conteúdo dizem respeito ao carregamento de vírus ou códigos maliciosos pelos usuários da plataforma e à impossibilidade de acessar e coletar dados do Facebook por meios automatizados sem prévia permissão da plataforma (FACEBOOK, 2020, online).

Ainda, nesse ponto dos Termos de Serviço, a plataforma legitima a moderação de conteúdo por explicitar a possibilidade de remover ou restringir o acesso a conteúdo que viole as disposições expostas, caso o Facebook entender que as medidas de remoção ou restrição são razoáveis e necessárias “para evitar ou reduzir impactos jurídicos ou regulatórios adversos para o Facebook” (FACEBOOK, 2020, online). Nos termos, não há explicação dos critérios utilizados pelo Facebook para determinar o que é razoável e necessário para coibir um usuário.

Na sequência, os Termos de Serviços apresentam as regras para encerramento ou suspensão das contas decorrente de violações “de forma manifesta, grave ou recorrente” (FACEBOOK, 2020, online).

Por fim, os Termos de Serviço impõem uma limitação à responsabilidade do Facebook quanto à segurança da rede e aos conteúdos publicados, limitando-se à extensão da aplicação legislativa em relação ao serviço fornecido (FACEBOOK, 2020, online). Ou seja, o Facebook isenta-se de qualquer responsabilização de conteúdos publicados em sua plataforma pelos usuários, mesmo que possam implicar danos de qualquer natureza.

O fundamento da moderação de conteúdo pelo Facebook pauta-se também nos Padrões da Comunidade desenvolvidos pela plataforma para estipular diretrizes de conteúdos de podem circular na plataforma, assim, a moderação de conteúdo ensinará sobre o conteúdo contrário aos Padrões da Comunidade do Facebook.

Reconhecendo a importância do Facebook como um local de comunicação, com o intuito de eliminar dos serviços todo tipo de abuso, a empresa desenvolveu um conjunto de Padrões da Comunidade que detalham o que é ou não permitido no Facebook, baseado nas áreas de tecnologia, segurança pública e direitos humanos (FACEBOOK, 2021a).

Nessa esteira, o objetivo da criação dos Padrões da Comunidade pelo Facebook justifica-se:

O objetivo de nossos Padrões da Comunidade sempre foi criar um espaço para expressão e dar voz às pessoas. Isso não mudou e não mudará. Criar uma comunidade e unir pessoas depende da capacidade de compartilhar diversos pontos de vista, experiências, ideias e informações. Queremos que as pessoas possam falar abertamente sobre os assuntos importantes para elas, ainda que sejam temas que geram controvérsias e objeções. Em alguns casos, permitimos conteúdo que poderia ir contra os nossos Padrões da Comunidade para fins de conscientização pública, caso ele seja interessante e de utilidade pública. Para fazer esses julgamentos, consideramos o valor do interesse público e o risco de dano. Também observamos os padrões internacionais relativos a direitos humanos. Dessa forma, avaliamos a importância do conteúdo publicado por alguém, incluindo organizações de notícias e usuários individuais. Por exemplo, permitimos conteúdo que retrate graficamente uma guerra ou as consequências dela quando isso é importante para o discurso público (FACEBOOK, 2021a).

Malgrado a plataforma reconhecer que a internet cria novas e maiores oportunidades de abusos, o Facebook assegura o comprometimento com a livre expressão, sendo as limitações baseadas em valores de autenticidade, segurança, privacidade e dignidade (FACEBOOK, 2021).

Esses valores representam, respectivamente, o compromisso com a autenticidade do conteúdo publicado e da identidade os usuários; o compromisso em manter a plataforma segura e livre de ameaças, intimidações e silenciamento de usuários; o compromisso em proteger as informações pessoais; e, o respeito a igualdade de direitos e dignidade (FACEBOOK, 2021).

Por mais que a plataforma objetiva a promoção da comunicação online, as limitações de conteúdo possuem, muito mais, um interesse comercial. Esse interesse

“é uma razão central para que empresas privadas de mídia social proíbam manifestação que não é ilegal e, portanto, limitem a liberdade de expressão muito além do exigido pela lei” (HARTMANN; SARLET, 2019, p. 96).

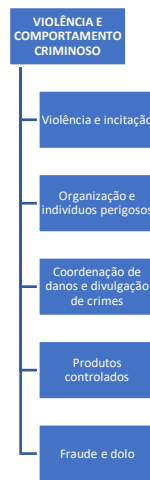
Vale explicar que os Padrões da Comunidade classificam os conteúdos publicados em 25 tópicos dentro de seis diretrizes, dentre elas, representa-se na Figura 7:

Figura 7 - Padrões da Comunidade do Facebook.



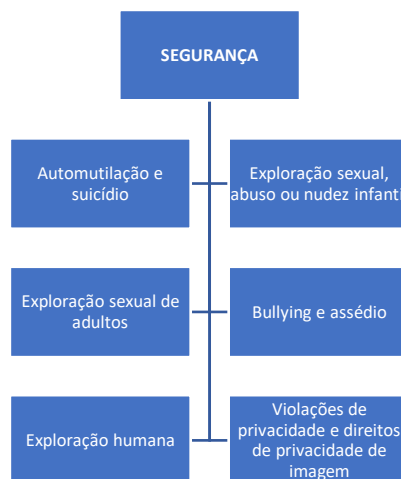
Fonte: Elaborado pelo autor, 2021.

A primeira diretriz, deste modo, engloba todas as publicações na plataforma que contenham violência e comportamento criminosos de modo a evitar a circulação desse tipo de conteúdo, a fim de evitar danos reais e potenciais aos usuários. (FACEBOOK, 2021a). Essa diretriz subdivide-se, na figura 8, em:

Figura 8 - Violência e comportamento criminoso.

Fonte: Elaborado pelo autor, 2021.

Na diretriz de Segurança, o Facebook afirma seu compromisso em fazer da plataforma um lugar seguro, onde conteúdos com ameaças, intimidações e exposição dos usuários à determinados conteúdos podem causar riscos à própria segurança física ou mental, dos quais representa-se na Figura 9:

Figura 9 - Segurança.

Fonte: Elaborado pelo autor, 2021.

Já na diretriz dos conteúdos questionáveis encontram-se publicações com conteúdos controversos que requerem uma análise profunda de nuances e contextos

da publicação, a fim de realizar distinções entre o debate e a violação dos limites de assuntos, nos termos da figura 10:

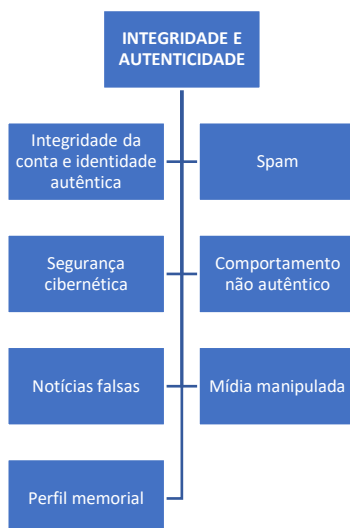
Figura 10 - Conteúdo questionável.



Fonte: Elaborado pelo autor, 2021.

Outra importante diretriz diz respeito à integridade e autenticidade, a qual compreende a atuação da plataforma em garantir a segurança cibernética dos usuários e das contas contra perfis falsos, spam e manipulações de conteúdos. Essa diretriz é representada pela Figura 11 e subdivide-se em:

Figura 11 - Integridade e autenticidade.



Fonte: Elaborado pelo autor, 2021.

Em atenção aos direitos autorais, a quarta diretriz disciplina as publicações sobre uso de marcas comerciais e outros direitos legais alheios, bem como problemas relativos à propriedade intelectual que possam decorrer do uso da rede social. Por fim, o último ponto dos Padrões da Comunidade refere-se a “solicitações e decisões relativas a conteúdo”, como requerimentos ao Facebook para a revisão das decisões de moderação de conteúdo.

Dentre as recentes atualizações nas políticas do Facebook, a plataforma dedicou-se em tornar mais clara os Padrões da Comunidade detalhando cada diretriz do não é permitido publicar na plataforma. Desta feita, sobrealva exemplificar como o Facebook orienta os usuários em relação ao que pode ou não ser publicado.

Parte-se da política do Conteúdo Questionável, limitada ao discurso de ódio, dada a natureza sensível dos conteúdos da diretriz associados aos resultados da moderação de conteúdo apresentados na seção 2.2.4, em contraste com os direitos fundamentais exercidos no ambiente da rede social, especialmente quanto à liberdade de expressão.

A política de moderação de conteúdo classificado como discurso de ódio inicia pela visão do Facebook de que permitir discursos de ódio na plataforma criam um ambiente de intimidação e exclusão podendo ocorrer casos de violência no mundo físico (FACEBOOK, 2021d, online). Assim, a definição de como o Facebook compreende esse fenômeno, conforme o fundamento da política, traduz-se:

Definimos discurso de ódio como um ataque direto a pessoas, e não a conceitos e instituições, baseado no que chamamos de características protegidas: raça, etnia, nacionalidade, religião, orientação sexual, casta, sexo, gênero, identidade de gênero e doença grave ou deficiência. Definimos ataques como discursos violentos ou desumanizantes, estereótipos prejudiciais, declarações de inferioridade, expressões de desprezo, repugnância ou rejeição, xingamentos e apelos à exclusão ou segregação. Também proibimos o uso de estereótipos prejudiciais, que definimos como comparações desumanizantes que têm sido historicamente usadas para atacar, intimidar ou excluir grupos específicos, e que muitas vezes estão ligadas à violência no meio físico (FACEBOOK, 2021d, online).

Percebe-se que o Facebook estabeleceu os limites da classificação com base na proteção das características individuais dos usuários considerando também o contexto regional e linguístico da publicação. A classificação do que se enquadra como discurso de ódio parte de 3 níveis de violação, dado que as pessoas possuem diferentes níveis de tolerância para falas sobre características protegidas (FACEBOOK, 2021d, online).

Desta feita, o Facebook (2021d, online) recomenda aos usuários que não publique, sob pena de remoção da publicação, em primeiro nível, conteúdo “visando um indivíduo ou grupo de pessoas (incluindo todos os subconjuntos, salvo os que descrevem o cometimento de crimes violentos ou ofensas sexuais)”, como:

- Discurso violento ou apoio de forma escrita ou visual
- Imagem ou discurso degradante sob a forma de comparações, generalizações ou declarações de comportamento não qualificadas (de forma visual ou escrita) voltadas para ou sobre:
 - Insetos.
 - Animais culturalmente percebidos como inferiores física ou intelectualmente.
 - Sujeira, bactérias, doenças e excrementos.
 - Predadores sexuais.
 - Subumanidade.
 - Criminosos sexuais e violentos.
 - Outros criminosos (incluindo, entre outros, “ladrões” e “assaltantes de banco” ou afirmando que “todo [característica protegida ou semiprotegida] é ‘criminoso’”).
 - Declarações negando a existência.
- Deboche do conceito, de eventos ou de vítimas de crimes de ódio, mesmo que nenhuma pessoa real apareça na imagem.
- Comparações ou generalizações desumanizantes designadas, ou afirmações comportamentais (por escrito ou visuais) que incluam:

- Pessoas negras e macacos ou seres semelhantes a macacos.
- Pessoas negras e equipamentos agrícolas.
- Caricaturas de pessoas negras com prática de blackface.
- Judeus e ratos.
- Judeus comandando o mundo ou controlando grandes instituições, como redes de mídia, a economia ou o governo.
- Negação ou distorção de informações sobre o Holocausto.
- Muçulmanos e porcos.
- Muçulmanos e relações sexuais com cabras ou porcos.
- Mexicanos e seres semelhantes a vermes.
- Mulheres como objetos domésticos ou referência à mulher como propriedade ou “objeto”.
- Pessoas transgêneras ou não binárias sendo chamadas de “isso”.
- Dalits, pessoas de casta registrada ou de "casta inferior" como trabalhadores braçais.

Em um segundo nível de tolerância, o Facebook (2021d, online) orienta que não seja publicado conteúdo, visando um usuário ou grupo com base em características protegidas, que contenha:

- Generalizações afirmando inferioridade (por escrito ou visuais) sob as seguintes formas:
 - Deficiências físicas são definidas em termos de:
 - Higiene, incluindo, entre outros, sujo, imundo, fedorento.
 - Aparência física, incluindo, entre outros, feio, medonho.
 - Deficiências mentais são definidas em termos de:
 - Capacidade intelectual, incluindo, entre outros, burro, imbecil, idiota.
 - Educação, incluindo, entre outros, analfabeto, atrasado.
 - Saúde mental, incluindo, entre outros, doente mental, retardado, louco, maluco.
 - Deficiências morais são definidas em termos de:
 - Traços de personalidade culturalmente tidos como negativos, incluindo, entre outros, covarde, mentiroso, arrogante e ignorante.
 - Termos pejorativos relacionados a atividades sexuais, incluindo, entre outros, vagabunda, vadia, perverso.
- Outras declarações de inferioridade, definidas em termos de:
 - Expressões sobre a falta de adequação, incluindo, entre outras, inútil, incapaz.
 - Expressões sobre ser melhor/pior do que outra característica protegida, incluindo, entre outras: “Eu acredito que os homens são superiores às mulheres.”
 - Expressões sobre desvio das normas, incluindo, entre outras, esquisito, anormal.
- Expressões de desprezo (de forma visual ou escrita), definidas como:

- Autoadmissão de intolerância com base em características protegidas, incluindo, entre outras, homofóbico, islamofóbico, racista.
- Expressões indicando que uma característica protegida não deveria existir.
- Expressões de ódio, incluindo, entre outras, desprezo, repulsa.
- Expressões de desaprovação, incluindo, entre outras, não respeito, não gosto, não me importo.
- Expressões de repulsa (de forma visual ou escrita), definidas como:
 - Expressões que sugiram que o alvo causa náusea, incluindo, entre outras, vômito, regurgitação.
 - Expressões de repulsa ou nojo, incluindo, entre outras, nojento, abjeto, eca.
- Xingamentos, definidos como:
 - Referir-se ao alvo como genitália ou ânus, incluindo, entre outras, arrombada, pau no cu, cuzão, escroto.
 - Termos profanos ou frases com a intenção de insultar, incluindo, entre outros, corno, puta, piranha, arrombado, fodido.
 - Termos ou frases solicitando participação em atividade sexual ou contato com a genitália, o ânus, as fezes ou a urina, incluindo, entre outros, chupa meu pau, lambe meu cu, come merda.

Já no terceiro nível da política de discurso de ódio, a rede social proíbe que seja publicado conteúdo, sob pena de remoção, conteúdo que “descreva ou prejudique pessoas com calúnias, definidas como palavras intrinsecamente ofensivas e usadas como frases ofensivas” (FACEBOOK, 2021d, online) com qualquer dos seguintes termos:

- Segregação sob a forma de conclamações, declarações de intenção, declarações intencionais ou condicionais, ou declarações que defendem ou apoiem a segregação.
- Exclusão por meio de conclamações, declarações de intenção, declarações intencionais ou condicionais, ou declarações que defendam ou apoiem a exclusão, definidas como:
 - Exclusão explícita, como expulsar certos grupos ou dizer que eles não são permitidos.
 - Exclusão política, como negar o direito à participação política.
 - Exclusão econômica, no sentido de negar o acesso aos direitos econômicos e limitar a participação no mercado de trabalho.
 - Exclusão social, como negar o acesso a espaços (físicos e online) e serviços sociais.

Por fim, o Facebook (2021d, online) exige informações e/ou contexto adicional dos usuários para a aplicação os Padrões da Comunidade, a fim de, consequentemente, moderar determinados conteúdos que:

- Conteúdo que, de forma explícita, forneça ou ofereça produtos ou serviços com o objetivo de mudar a orientação sexual ou a identidade de gênero das pessoas.
- Conteúdo que ataca conceitos, instituições, ideias, práticas ou crenças associadas a características protegidas, que provavelmente contribuem para danos corporais iminentes, intimidação ou discriminação contra as pessoas associadas a essa característica protegida. O Facebook considera vários sinais para determinar se há uma ameaça de dano no conteúdo. Isso inclui, entre outros, conteúdo que poderia incitar violência ou intimidação iminente; se existe um período de alta tensão, como uma eleição ou conflito em andamento; e se existe um histórico recente de violência contra o grupo protegido. Em alguns casos, podemos também considerar se o orador é uma figura pública ou ocupa uma posição de autoridade.

Em suma, os Padrões da Comunidade importam em regras para orientar toda a moderação de conteúdo na rede social do que estaria violando ou não as regras de uso da plataforma, isto é, os Padrões são normas que dispõem o conteúdo que pode estar na rede.

Cada diretriz possui orientações para os usuários do conteúdo violador que se enquadra na determinada política. Portanto, se determinado usuário publicar algo que se amolde como conteúdo proibido pelos Padrões da Comunidade do Facebook terá sua publicação moderada e restringida.

Os termos e os Padrões da Comunidade são documentos de elevada importância para o funcionamento da rede, dado que constituem as regras de operacionalização da plataforma e as regras do que é legítimo ao usuário. Entretanto, a construção dessas políticas é procedida de forma outorgada e imposta aos usuários, sem qualquer possibilidade de discussão democrática.

A relevância jurídica dos Padrões da Comunidade importa legitimação e fundamento da atividade de moderação de conteúdo, seja para a remoção de conteúdo violador das políticas ou a suspensão de conta, devendo, portanto, essas regras estarem de acordo com o ordenamento jurídico do Estado o qual o usuário é jurisdicionado.

Sobreleva destacar que os Termos de Serviço e os Padrões da Comunidade possuem um “tom constitucional” (CELESTE, 2018, p. 3) por fazerem uso do tradicional estilo de textos constitucionais e à adoção da configuração peculiar que articula conteúdos em termos de direitos, princípios e deveres.

Isso fica evidente na análise formal e material dos termos da rede social. Do ponto de vista formal, o Facebook liberou seus termos da retórica característica dos acordos privados para adotar a redação tradicional do direito. Do ponto de vista

material, o Facebook adapta os princípios constitucionais tradicionais, como a liberdade de expressão ou o direito à privacidade e à proteção de dados, ao contexto de sua plataforma, sugerindo uma ideia de espaço democrático (CELESTE, 2018, p. 4). Isto representa um ponto chave na garantia da transparência, através da explicabilidade das decisões trazendo mais clareza.

Na perspectiva democrática, as plataformas digitais acenderam como um espaço no qual direitos fundamentais são exercidos, incidindo no aumento da moderação do conteúdo nas redes sociais pelo aumento do fluxo informacional. Nessa esteira, as regulamentações de ambientes digitais e como essas regras são aplicadas têm impacto relevante sobre os direitos fundamentais de seus usuários.

O reconhecimento desse impacto implica novas formas de pensar sobre a governança da plataforma digital e uma melhor compreensão de como a governança online e os direitos fundamentais convergem. Posto isso, a partir do exercício de direitos fundamentais no ambiente democrático demanda uma necessidade constitucional de proteção desses direitos, decorrente dos conflitos no espaço virtual pela ampliação do poder das plataformas sobre os usuários.

REFERÊNCIAS

ALEMANHA. Ministério Federal da Justiça e Defesa do Consumidor. **Network Enforcement Act (Netzwerkdurchsetzungsgesetz -NetzDG)**, de 01 de setembro de 2017. Disponível em: <https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html>. Acesso em: 02 mar. 2021.

ALEXY, Robert. Formal principles: Some replies to critics. **Oxford University Press and New York University School of Law**, v. 12, n. 3, 511–524, 2014.

ALEXY, Robert. **Teoria da Argumentação Jurídica**. São Paulo: Landy, 2001.

ALEXY, Robert. **Teoria dos Direitos Fundamentais**. Trad. Virgílio Afonso da Silva. São Paulo: Malheiros, 2008.

ARRIETA, Alejandro; et.al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**. v. 58, 2019.

BALKIN, Jack M., How to Regulate (and Not Regulate) Social Media. **Knight Institute Occasional Paper Series**, n. 1 mar. 2020. Disponível em: <https://knightcolumbia.org/content/how-to-regulate-and-not-regulate-social-media>. Acesso em: 13 nov. 2020.

BALKIN, Jack M. The Future of Free Expression in a Digital Age. **Pepperdine Law Review**, [s.l.], v. 36, n. 2, 2009. Disponível em: <http://digitalcommons.pepperdine.edu/plr/vol36/iss2/9> Acesso em: 06 dez. 2021.

BARRETT, Paul M. **Who Moderates the Social Media Giants? A Call to End Outsourcing**. Nova Iorque: New York University, 2020.

BARROSO, Luís Roberto. **Curso de direito constitucional contemporâneo: os conceitos fundamentais e a construção do novo modelo**. São Paulo: Saraiva, 2009.

BARROSO, Luís Roberto. Hate, Lies, and Democracy. **Int'l J. Const. L. Blog**, 22 jan. 2022. Disponível em: <http://www.iconnectblog.com/2022/01/hate-lies-and-democracy> Acesso em: 03 fev. 2022.

BASSINI, Marco. Fundamental rights and private enforcement in the digital age. **Eur Law J.**, v. 25, 182–197, 2019.

BINENBOJM, Gustavo. Direito ao esquecimento: a censura no retrovisor. **Jota** [s.l.], 16 out. 2016. Disponível em: <https://www.jota.info/opiniao-e-analise/artigos/direito-ao-esquecimento-censura-retrovisor-16102014> Acesso em: 3 fev. 2022.

BRANCO, Paulo Gustavo Gonet. Aspectos da Teoria Geral dos Direitos Fundamentais. IN: MENDES, Gilmar F. **Hermenêutica constitucional e direitos fundamentais**. Inocência Mártires Coelho, Paulo Gustavo Gonet Branco, Gilmar Ferreira Mendes Brasília: Brasília Jurídica, 2000.

BRASIL. Câmara dos Deputados. **Projeto de Emenda à Constituição 17/2019**.

Disponível em:

<https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2210757> Acesso em: 20 dez. 2021.

BRASIL. Câmara dos Deputados. Projeto de Lei nº 2.630, de 03 de julho de 2020.

Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet.

Disponível em:

<https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735&fichaAmigavel=nao>. Acesso em: 03 mar. 2021.

BRASIL. Câmara dos Deputados. **Projeto de Lei nº 3.227**, de 20 de setembro de 2021. Disponível em:

https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2076539&filename=PL+3227/2021. Acesso em: 20 dez. 2021.

BRASIL. **Constituição Federal (1988)**. Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal: Centro Gráfico, 1988. Disponível em:

http://www.planalto.gov.br/ccivil_03/constituicao/constituicaocompilado.htm Acesso em: 21 dez. 2021

BRASIL. Justiça Federal Seção Judiciária de São Paulo. **Procedimento Comum**

Cível 5004924-79.2020.4.03.6100. Julgado em 27 de abril de 2020a. Disponível em:

<https://pje1g.trf3.jus.br/pje/ConsultaPublica/DetalheProcessoConsultaPublica/listView.seam?ca=46350c235410edbca349563b97ce43cf845d08a807f9fbba> Acesso em: 28 jan. 2022.

BRASIL. Lei 13.709 de 14 de agosto de 2018. **Lei Geral de Proteção de Dados (LGPD)**. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/L13709.htm. Acesso em: 03 mar. 2021.

BRASIL. Lei nº 12.965, de 23 de abril de 2014. **Marco Civil da Internet**. Disponível em:

http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 01 mar. 2021.

BRASIL. Poder Judiciário. Conselho da Justiça Federal. **Enunciado nº 613**. VIII

Jornada de Direito Civil. Brasília, 2018a. Disponível em:

<https://www.cjf.jus.br/enunciados/enunciado/1161> Acesso em: 14 dez. 2021.

BRASIL. Superior Tribunal de Justiça. **Embargos de Declaração no Recurso Especial nº 1568935** - RJ, Voto do Rel. Min. Ricardo Villas Bôas Cueva, julgado em

04 de agosto de 2016. Disponível em:

https://processo.stj.jus.br/processo/revista/documento/mediado/?componente=ATC&sequencial=59584867&num_registro=201501011370&data=20160413&tipo=51&formato=PDF Acesso em: 22 dez. 2021.

BRASIL. Superior Tribunal de Justiça. **Recurso Especial nº 1.316.921** – RJ.

Relatora: Ministra Nancy Andrighi, julgado em 26 de junho de 2012. Disponível em:

<https://processo.stj.jus.br/processo/pesquisa/?src=1.1.3&aplicacao=processos.ea&tip>

oPesquisa=tipoPesquisaGenerica&num_registro=201103079096 Acesso em: 28 jan. 2022.

BRASIL. Superior Tribunal de Justiça. **Recurso Especial nº 1381610** – RS. Relatora: Ministra Nancy Andrighi, julgado em 3 de setembro de 2013. Disponível em: https://scon.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=201300613536&dt_publicacao=12/09/2013 Acesso em: 22 dez. 2021.

BRASIL. Superior Tribunal de Justiça. **Recurso Especial nº 1642997** – RJ. Relatora: Ministra Nancy Andrighi, julgado em 12 de setembro de 2017. Disponível em: https://scon.stj.jus.br/SCON/GetInteiroTeorDoAcordao?num_registro=201602722634&dt_publicacao=15/09/2017 Acesso em: 22 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Ação Direita de Inconstitucionalidade nº 4451** – DF, Relator: Ministro Alexandre de Moraes, julgada em 26 de junho de 2018b. Disponível em: <https://jurisprudencia.stf.jus.br/pages/search/sjur399151/false> Acesso em: 14 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Arguição de Descumprimento de Preceito Fundamental nº 130**- DF. Relator: Ministro Carlos Britto, julgada em 30 de abril de 2009. Disponível em: <https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=AC&docID=605411> Acesso em: 14 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Arguição de Descumprimento de Preceito Fundamental nº 187**- DF. Relator: Ministro Celso de Mello, julgada em 15 de junho 2011. Disponível em: <https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=5956195> Acesso em: 14 dez. 2021

BRASIL. Supremo Tribunal Federal. **Habeas Corpus 8.242**. Relator: MOREIRA ALVES, Relator p/ Acórdão: MAURÍCIO CORRÊA, julgado em 17 de setembro de 2003. Disponível em: <<https://jurisprudencia.stf.jus.br/pages/search/sjur96610/false>>. Acesso em 06 dez 2021.

BRASIL. Supremo Tribunal Federal. **Recurso Extraordinário 989.450** – SP. Relator: Ministro Luiz Fux, julgado em 17 agosto 2016. Disponível em: <https://redir.stf.jus.br/paginadorpub/paginador.jsp?docTP=TP&docID=12977132> Acesso em: 14 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Recurso Extraordinário 1.010.606** – RJ. Relator: Ministro Dias Toffoli. Processo eletrônico, 2021e. Disponível em: <https://portal.stf.jus.br/processos/detalhe.asp?incidente=5091603> Acesso em: 28 jan. 2022.

BRASIL. Supremo Tribunal Federal. **Recurso Extraordinário 1.037.396** – SP. Relator: Ministro Luiz Fux. Processo eletrônico, 2021d. Disponível em:

<http://portal.stf.jus.br/processos/detalhe.asp?incidente=5160549> Acesso em: 22 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Recurso Extraordinário 1.057.258** – MG. Relator: Ministro Luiz Fux. Processo eletrônico, 2021c. Disponível em: <http://portal.stf.jus.br/processos/detalhe.asp?incidente=5217273> Acesso em: 22 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Recurso Ordinário em Habeas Corpus 134.682**. Relator: EDSON FACHIN, julgado em 29 de novembro de 2016a. Disponível em: <<https://jurisprudencia.stf.jus.br/pages/search/sjur372168/false>>. Acesso em 06 dez 2021.

BRASIL. Supremo Tribunal Federal. **Tema de Repercussão Geral nº 533**. Dever de empresa hospedeira de sítio na internet fiscalizar o conteúdo publicado e de retirá-lo do ar quando considerado ofensivo, sem intervenção do Judiciário. Relator: Min. Luiz Fux, 2021. Disponível em: <http://www.stf.jus.br/portal/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=4155926&numeroProcesso=660861&classeProcesso=ARE&numeroTema=533>. Acesso em: 20 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Tema de Repercussão Geral nº 987**. Discussão sobre a constitucionalidade do art. 19 da Lei n. 12.965/2014 (Marco Civil da Internet) que determina a necessidade de prévia e específica ordem judicial de exclusão de conteúdo para a responsabilização civil de provedor de internet, websites e gestores de aplicativos de redes sociais por danos decorrentes de atos ilícitos praticados por terceiros. Relator: Min. Dias Toffoli, 2021b. Disponível em: <http://www.stf.jus.br/portal/jurisprudenciaRepercussao/verAndamentoProcesso.asp?incidente=5160549&numeroProcesso=1037396&classeProcesso=RE&numeroTema=987#>. Acesso em: 20 dez. 2021.

BRASIL. Supremo Tribunal Federal. **Tema de Repercussão Geral nº 1141**. Responsabilidade civil por disponibilização na internet de informações processuais publicadas nos órgãos oficiais do Poder Judiciário, sem restrição de segredo de justiça ou obrigação jurídica de remoção. Relatora: Ministra Cármen Lúcia, 2021f. Disponível em: <https://portal.stf.jus.br/jurisprudenciaRepercussao/tema.asp?num=1141> Acesso em: 28 jan. 2022.

CALLEJÓN, Francisco Balaguer. REDES SOCIAIS, COMPANHIAS TECNOLÓGICAS E DEMOCRACIA. **REI - REVISTA ESTUDOS INSTITUCIONAIS**, [S.l.], v. 6, n. 2, p. 579-599, set. 2020.

CAPLAN, Robyn. CONTENT OR CONTEXT MODERATION? Artisanal, Community-Reliant, and Industrial Approaches. **Data & Society**. [s.l.], nov. 2018. Disponível em: <https://datasociety.net/library/content-or-context-moderation/> Acesso em: 15 jun. 2021.

CARINI, Lucas; MORAIS, Fausto Santos de. Governança Ética Para Construção De Confiança Em Sistemas De Inteligência Artificial. **PRIM@ FACIE**. João Pessoa: PPGCJ, v. 19, n. 40, 2020.

CASTELLS, Manuel. **A Galáxia da Internet**: reflexões sobre a internet, os negócios e a sociedade. Rio de Janeiro: Jorge Zahar Ed, 2003

CELESTE, Edoardo. Digital Constitutionalism: Mapping the Constitutional Response to Digital Technology's Challenges. **HIIG Discussion Paper Series**. v. 2018-02, 2018.

CHARLEAUX, Lupa. Facebook bane Donald Trump 'indefinitivamente' de suas plataformas. **TECMUNDO**, [s.l.], 7 jan. 2021. Disponível em: <https://www.tecmundo.com.br/redes-sociais/209084-facebook-bane-donald-trump-indefinitivamente-plataformas.htm> Acesso em: 25 ago. 2021.

CONTA de Eduardo Bolsonaro é suspensa no Facebook; entenda o que rolou. **TILT**, São Paulo, 20 jul. 2021. Disponível em: <https://www.uol.com.br/tilt/noticias/redacao/2021/07/20/eduardo-bolsonaro-tem-conta-suspensa-no-facebook-politica-tenta-reverter.htm> Acesso em: 26 ago. 2021.

CULLIFORD, Elizabeth. Twitter e Facebook rotulam publicações de Trump nas eleições dos EUA. **Agência Brasil**, [s.l.], 3 nov. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/internacional/noticia/2020-11/twitter-e-facebook-rotulam-publicacoes-de-trump-nas-eleicoes-dos-eua> Acesso em: 25 ago. 2021.

DE GREGORIO, Giovanni. The Rise of Digital Constitutionalism in the European Union. **International Journal of Constitutional Law**, v. 19(1), p. 41-70, 2020.

DOMINGOS, Pedro. **O Algoritmo Mestre**: como a busca pelo algoritmo de *machine learning* definitivo recriará nosso mundo. São Paulo: Novatec, 2017.

ESTADOS UNIDOS DA AMÉRICA. **Constitution of the United States**. [s.l.] 1788. Disponível em: <https://constitution.congress.gov/constitution/> Acesso em: 04 nov. 2021.

ESTADOS UNIDOS DA AMÉRICA. **Digital Millennium Copyright Act**. [s.l.] 1998. Disponível em: <https://www.copyright.gov/legislation/dmca.pdf> Acesso em: 20 dez. 2021.

ESTADOS UNIDOS DA AMÉRICA. **United States Code**. 2021. Disponível em: <https://uscode.house.gov/browse/prelim@title47/chapter5/subchapter2/part1&edition=prelim>. Acesso em: 04 nov. 2021.

EU: Due diligence obligations in the proposed Digital Services Act. **ARTICLE 19**, [s.l.], 21 mai. 2021a. Disponível em: <https://www.article19.org/resources/eu-due-diligence-obligations-in-the-proposed-digital-services-act/> Acesso em: 10 out. 2021.

FACEBOOK altera regras para permitir fotos de mães amamentando. **G1**, [s.l.], 15 jun. 2014. Disponível em: <http://g1.globo.com/tecnologia/noticia/2014/06/facebook->

altera-regras-para-permitir-fotos-de-maes-amamentando.html Acesso em: 22 ago. 2021.

FACEBOOK e Instagram bloqueiam conta de Trump por tempo indeterminado. **G1**, [s.l.], 07 jan. 2021. Disponível em: <https://g1.globo.com/economia/tecnologia/noticia/2021/01/07/facebook-bane-conta-de-donald-trump-por-tempo-indeterminado-diz-mark-zuckerberg.ghtml>. Acesso em: 01 mar. 2021.

FACEBOOK e Twitter excluem postagens de Trump com 'desinformação nociva sobre covid-19'. **BBC NEWS**, [s.l.], 5 ago 2020. Disponível em: <https://www.bbc.com/portuguese/internacional-53674195> Acesso em: 25 ago. 2021.

FACEBOOK remove publicação de Trump que diz que gripe é mais letal que Covid-19. **CNN BRASIL**, [s.l.], 6 out. 2020. Disponível em: <https://www.cnnbrasil.com.br/internacional/facebook-remove-publicacao-de-trump-dizendo-que-gripe-e-mais-letal-que-covid-19/> Acesso em: 25 ago. 2021.

FACEBOOK. **About**. [s.l.] 2021. Disponível em: <https://about.facebook.com/company-info/> Acesso em: 01 jun. 2021.

FACEBOOK. **Central de Ajuda**. [s.l.] 2021c. Disponível em: <https://pt-br.facebook.com/help> Acesso em: 09 jun. 2021.

FACEBOOK. **Padrões da Comunidade**. [s.l.] 2021a. Disponível em: <https://www.facebook.com/communitystandards/>. Acesso em: 01 mar. 2021.

FACEBOOK. **Política de dados**. [s.l.] 2021b. Disponível em: <https://www.facebook.com/policy.php>. Acesso em: 01 mar. 2021.

FACEBOOK. **Termos de Serviço**. [s.l.] 2020. Disponível em: <https://www.facebook.com/legal/terms/update>. Acesso em: 01 mar. 2021.

FACEBOOK. **Transparency Center**. [s.l.] 2021d. Disponível em: <https://transparency.fb.com/> Acesso em: 16 jun. 2021.

FERRAJOLI, Luigi. **Derechos y garantías**: La ley del más débil. Madri: Trotta, 2016.

FIORILLO, Celso Antonio Pacheco. **Curso de Direito Ambiental Brasileiro**. São Paulo: Saraiva, 2014.

FIORILLO, Celso Antonio Pacheco; SOUZA, Carolina Ferreira. O DIREITO À COMUNICAÇÃO E A TUTELA DO MEIO AMBIENTEDIGITAL. **Revista de Direito Ambiental e Socioambientalismo**, v. 2, n. 2, 186-206, jul/dez. 2016.

FRANÇA. Assemblée nationale. **PROPOSITION DE LOI visant à lutter contre la haine sur internet n° 1785** de 20 mar. 2019. Disponível em: https://www.assemblee-nationale.fr/dyn/15/textes/l15b1785_proposition-loi#D_Article_6 Acesso em: 10 out. 2021.

FRANÇA. Conseil Constitutionnel. **Décision n° 2020-801 DC** Loi visant à lutter contre les contenus haineux sur internet de 18 jun. 2020. Disponível em: <https://www.conseil-constitutionnel.fr/decision/2020/2020801DC.htm> Acesso em: 04 nov. 2021.

FREITAS, Luiz Fernando Calil de. **Direitos fundamentais**: limites e restrições. Porto Alegre: Livraria do Advogado, 2007.

GARDBAUM, Stephen. THE "HORIZONTAL EFFECT" OF CONSTITUTIONAL RIGHTS. **Michigan Law Review**. v.102, p. 387-459, 2003.

GILLESPIE, Tarleton. **Custodians of the Internet**: platforms, content moderation, and the hidden decisions that shape social media. [s.l.] Yale University Press, 2018.

DA SILVA, LUCAS GONÇALVES; CARVALHO, Mariana Amaral. DIREITO AO ESQUECIMENTO NA SOCIEDADE DA INFORMAÇÃO: análise dos direitos fundamentais no meio ambiente digital. **Revista Brasileira de Direitos e Garantias Fundamentais**, v. 3, p. 66-86, 2017.

GRIMMELMANN, James. The Virtues of Moderation. **Yale Journal of Law & Technology**. v. 17, p. 41-109, 2015.

GRYGIEL, Jennifer. Após massacre em Christchurch, chegou a hora de acabar com o Facebook Live? **Poder360**, [s.l.], 31 mar. 2019. Disponível em: <https://www.poder360.com.br/nieman/apos-massacre-em-christchurch-chegou-a-hora-de-acabar-com-o-facebook-live/> Acesso em: 24 ago. 2021.

HARTMANN PEIXOTO, Fabiano. **Direito e Inteligência Artificial**: Referenciais básicos com comentários à resolução CNJ 332/2020. Brasília: DR.IA, 2020.

HARTMANN, Ivar Alberto M.; ABBAS DA SILVA, Lorena. INTELIGÊNCIA ARTIFICIAL E MODERAÇÃO DE CONTEÚDO: O SISTEMA CONTENT ID E A PROTEÇÃO DOS DIREITOS AUTORAIS NA PLATAFORMA YOUTUBE. **Ius Gentium**. Curitiba, vol. 10, n. 3, p. 145-165, set./dez. 2019.

HARTMANN, Ivar Alberto M.; MONTEIRO, Julia Iunes. Fake News no Contexto de Pandemia e Emergência Social: os Deveres e Responsabilidades das Plataformas de Redes Sociais na Moderação de Conteúdo Online: entre a Teoria e as Proposições Legislativas. **Revista Direito Público**, Brasília, v. 17, n. 94, 388-414, jul./ago. 2020. Disponível em: <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/4607> Acesso em: 20 dez. 2021.

HARTMANN, Ivar Alberto M.; SARLET, Ingo Wolfgang. DIREITOS FUNDAMENTAIS E DIREITO PRIVADO: A PROTEÇÃO DA LIBERDADE DE EXPRESSÃO NAS MÍDIAS SOCIAIS. **Revista Direito Público**, Brasília, v. 16, n. 90, dez. 2019. Disponível em: <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/3755>. Acesso em: 07 mar. 2021.

KEMP, Simon. **Digital 2020**: Global Digital Overview Report. [s.l.] We Are Social & Hootsuite, 2020. Disponível em: <https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media> Acesso em: 28 mai. 2021.

KEMP, Simon. **Digital 2021**: BRAZIL. [s.l.] We Are Social & Hootsuite, 2021. Disponível em: <https://datareportal.com/reports/digital-2021-brazil> Acesso em: 28 mai. 2021.

LAVADO, Thiago. Facebook e Twitter removem posts de Trump após invasão do Capitólio. **EXAME**, [s.l.], 6 jan. 2021. Disponível em: <https://exame.com/tecnologia/facebook-remove-video-de-trump-apos-invasao-do-capitolio-twitter-impede-compartilhamento-de-publicacoes/> Acesso em: 25 ago. 2021.

LEAL DA SILVA, Rosane; BOLSON DALLA FAVERA, Rafaela. Estudo do caso Klayman v. Zuckerberg and facebook: da liberdade de expressão ao discurso do ódio/Study of the case klayman v. zuckerberg and facebook: from freedom of speech to hate speech. **Revista Brasileira de Direito**, Passo Fundo, v. 13, n. 2, p. 273-292, ago. 2017. Disponível em: <https://seer.imed.edu.br/index.php/revistadedireito/article/view/923>. Acesso em: 08 nov. 2021.

LEMOES, Ronaldo. **Direito, tecnologia e cultura**. E-book publicado pela licença Creative Commons na plataforma Google Books. 2005. Disponível em: https://books.google.com.br/books?hl=en&lr=&id=xG-2W7YnTfUC&oi=fnd&pg=PA7&dq=info:m7FSyJFgN1IJ:scholar.google.com&ots=P_B-p8yRem&sig=FyYOhKI64B7lfbmASWmqTD5nyyE&redir_esc=y#v=onepage&q&f=false. Acesso em: 28 mar. 2022.

MAGRANI, Eduardo. **Entre dados e robôs**: ética e privacidade na era da hiperconectividade. 2 ed. Porto Alegre: Arquipélago Editorial, 2019.

MAGRANI, Eduardo. Governance Of Internet Of Things And Ethics Of Artificial Intelligence. **Revista Direitos Culturais**, Santo Ângelo, v. 13, n. 31, p. 153-190, set./dez. 2018.

MARTINO, Luís Mauro Sá. **Teoria das Mídias Digitais**: linguagens, ambientes, redes. Petrópolis: Vozes, 2014.

MENDES, Gilmar F. **Hermenêutica constitucional e direitos fundamentais**. Inocência Mártires Coelho, Paulo Gustavo Gonet Branco, Gilmar Ferreira Mendes Brasília: Brasília Jurídica, 2000.

MENDES, Gilmar Ferreira; FERNANDES, Victor Oliveira. Constitucionalismo digital e jurisdição constitucional: uma agenda de pesquisa para o caso brasileiro. **Revista Brasileira de Direito**, Passo Fundo, v. 16, n. 1, p. 1-33, out. 2020. Disponível em: <https://seer.imed.edu.br/index.php/revistadedireito/article/view/4103>. Acesso em: 01 mar. 2021.

MORAES, Alexandre de. **Direito constitucional**. São Paulo: Atlas, 2020.

MORAIS, Fausto Santos de. O USO DA INTELIGÊNCIA ARTIFICIAL NA REPERCUSSÃO GERAL: DESAFIOS TEÓRICOS E ÉTICOS. **Direito Público**, [S.l.], v. 18, n. 100, jan. 2022. ISSN 2236-1766. Disponível em: <https://www.portaldeperiodicos.idp.edu.br/direitopublico/article/view/6001> Acesso em: 04 fev. 2022.

MORAIS, Fausto Santos de. **Ponderação e Arbitrariedade: A Inadequada Recepção de Alexy Pelo STF**. Salvador: Juspodivm, 2018.

MOZILLA. **Internet Health Report v0.1**. [s.l.] 2017. Disponível em: https://d20x8vt12bnfa2.cloudfront.net/InternetHealthReport_v01.pdf Acesso em 04 jun 2021.

NAÇÕES UNIDAS. **Declaração Universal dos Direitos Humanos de 1948**. Disponível em: <https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=por> Acesso em 06 dez 2021.

NOVAIS, Jorge Reis. **As restrições aos Direitos Fundamentais não expressamente autorizadas pela Constituição**. Coimbra: Coimbra Editora, 2010.

O'SULLIVAN, Donie. Facebook remove anúncios da campanha de Trump por violação da política de ódio. **CNN BRASIL**, [s.l.], 19 jun. 2020. Disponível em: <https://www.cnnbrasil.com.br/internacional/facebook-remove-anuncios-da-campanha-de-trump-por-violacao-da-politica-de-odio/> Acesso em: 25 ago. 2021.

ORTUTAY, Barbara. Zuckerberg still under fire over inflammatory Trump posts. **AP NEWS**, [s.l.], 2 jun. 2020. Disponível em: <https://apnews.com/article/violence-donald-trump-us-news-ap-top-news-mark-zuckerberg-3f341e0945436db3b01fc82d1b760fef> Acesso em: 25 ago. 2021.

PAYÃO, Felipe. Facebook apaga foto famosa da Guerra do Vietnã e recebe críticas pelo mundo. **TECMUNDO**, [s.l.], 09 set. 2016. Disponível em: <https://www.tecmundo.com.br/facebook/109371-facebook-apaga-foto-famosa-guerra-vietna-recebe-criticas-mundo.htm> Acesso em: 22 ago. 2021.

PAYÃO, Felipe. Por que o Facebook não bloqueou a live do massacre na Nova Zelândia. **TECMUNDO**, [s.l.], 22 mar. 2019. Disponível em: <https://www.tecmundo.com.br/seguranca/139761-facebook-nao-conseguiu-bloquear-livestream-massacre-terrorista.htm> Acesso em: 24 ago. 2021.

PIERRE, Lévy, **Cibercultura**. São Paulo: Ed. 34, 1999.

PINHEIRO, Patricia Peck. **Direito Digital**. 5. ed. São Paulo: Saraiva, 2013.

PORTUGAL. Lei n. 27/2021 de 17 de maio de 2021. **Carta Portuguesa de Direitos Humanos na Era Digital**. Disponível em: <https://dre.pt/application/conteudo/163442504> Acesso em: 24 jun. de 2021.

PRIVACIDADE Hackeada. Direção de Karim Amer e Jehane Noujaim. Netflix, 2019. (139 min).

REDEKER, Dennis; GILL, Lex; GASSER, Urs. Towards digital constitutionalism? Mapping attempts to craft an Internet Bill of Rights. **International Communication Gazette**, v. 80, n. 4, jun. 2018.

REINO UNIDO. Government Digital Service. **Online Harms White Paper** de abril de 2019. Disponível em: <https://www.gov.uk/government/consultations/online-harms-white-paper>. Acesso em: 02 mar. 2021.

REINO UNIDO. Department for Digital, Culture, Media & Sport. **Draft Online Safety Bill** de 12 de mai. de 2021. Disponível em: <https://www.gov.uk/government/publications/draft-online-safety-bill> Acesso em: 05 out. 2021.

SARLET, Ingo Wolfgang. **A eficácia dos direitos fundamentais**: uma teoria geral dos direitos fundamentais na perspectiva constitucional. 11. ed. rev. atual. Porto Alegre: Livraria do Advogado Editora, 2012.

SARLET, Ingo Wolfgang. Proteção de dados pessoais como direito fundamental na Constituição Federal brasileira de 1988: contributo para a construção de uma dogmática constitucionalmente adequada. **Direitos Fundamentais & Justiça**, Belo Horizonte, ano 14, n. 42, p. 179-218, jan./jun. 2020.

SARLET, Ingo; ROBL FILHO, Ilton. Estado Democrático de Direito e os limites da liberdade de expressão na Constituição Federal de 1988, com destaque para o problema da sua colisão com outros direitos fundamentais, em especial, com os direitos de personalidade. **Revista da Academia Brasileira de Direito Constitucional**, v. 08, n. 14, Constituição, Economia e Desenvolvimento, 2016. Disponível em: <http://abdconst.com.br/revista15/estadollton.pdf>. Acesso em: 20 dez. 2021.

SARLET, Ingo Wolfgang; SIQUEIRA, Andressa De Bittencourt. LIBERDADE DE EXPRESSÃO E SEUS LIMITES NUMA DEMOCRACIA: O CASO DAS ASSIM CHAMADAS “FAKE NEWS” NAS REDES SOCIAIS EM PERÍODO ELEITORAL NO BRASIL. **Revista Estudos Institucionais**, v. 6, n. 2, p. 534-578, maio/ago. 2020.

SARMENTO, Daniel. **Direitos fundamentais e relações privadas**. 2. ed. Rio de Janeiro: Lumen Juris, 2010.

SARMENTO, Daniel. LIBERDADES COMUNICATIVAS E “DIREITO AO ESQUECIMENTO” NA ORDEM CONSTITUCIONAL BRASILEIRA. **Revista Brasileira de Direito Civil**. v.7, jan-mar, 2016.

SENRA, Ricardo. Após Twitter, Facebook e Instagram excluem vídeo de Bolsonaro por 'causar danos reais às pessoas'. **BBC NEWS**, Londres, 30 mar. 2020. Disponível em: <https://www.bbc.com/portuguese/brasil-52101240> Acesso em: 26 ago. 2021.

SILVA, José Afonso da. **Curso de Direito Constitucional Positivo**. São Paulo: Malheiros, 2011.

SILVA, Virgílio Afonso da. **Direitos Fundamentais**: conteúdo essencial, restrições e eficácia. São Paulo: Malheiros, 2010.

SUSI, Mart. The internet balancing formula. **Eur Law J.** v. 25, p. 198–212, 2019.

SUZUR, Nicolas. Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms. **Social Media + Society**, p. 1-11, jul. 2018

TAKANO, Camila Cardoso; GONÇALVES DA SILVA, Lucas. O CONSTITUCIONALISMO DIGITAL E AS NOVASTECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO(TIC). **Revista de Direito, Governança e Novas Tecnologias**, v. 6, n. 1, 1-15, jan/jun. 2020

TEUBNER, Gunter. Horizontal Effects of Constitutional Rights in the Internet: A Legal Case on the Digital Constitution. **THE ITALIAN LAW JOURNAL**, v. 3, n. 1, jul. 2017. Disponível em: <https://www.theitalianlawjournal.it/teubner-1/> Acesso em: 18 out. 2021.

TEUBNER, Gunter. **Constitutional Fragments**. Reino Unido: Oxford University Press, 2012.

THE SANTA CLARA PRINCIPLES on Transparency and Accountability in Content Moderation. Disponível em: <https://santaclaraprinciples.org/>. Acesso em: 01 mar. 2021.

UNIÃO EUROPEIA. Comissão Europeia. **REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL** on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC de 2020. Disponível em: <https://eur-lex.europa.eu/legal-content/en/TXT/?qid=1608117147218&uri=COM%3A2020%3A825%3AFIN> . Acesso em: 10 out. 2021.

UK: Draft Online Safety Bill poses serious risk to free expression. **ARTICLE 19**, [s.l.], 26 jul. 2021. Disponível em: <https://www.article19.org/resources/uk-draft-online-safety-bill-poses-serious-risk-to-free-expression/> Acesso em: 05 out. 2021.

VEJA a cronologia da invasão do Congresso dos Estados Unidos por apoiadores de Trump. **G1**, [s.l.], 6 jan. 2021a. Disponível em: <https://g1.globo.com/mundo/noticia/2021/01/06/veja-a-cronologia-da-invasao-do-congresso-dos-estados-unidos-por-apoiadores-de-trump.ghml> Acesso em: 25 ago. 2021.