UNIVERSIDADE DO EXTREMO SUL CATARINENSE - UNESC
PROGRAMA DE PÓS-GRADUAÇÃO EM DIREITO
MESTRADO EM DIREITO

FÁBIO JEREMIAS DE SOUZA

MÍDIAS SOCIAIS, COMBATE À DESINFORMAÇÃO E MODERAÇÃO DE CONTEÚDO: ANÁLISE DAS DIRETRIZES INTERNACIONAIS DO DIREITO HUMANO À LIBERDADE DE EXPRESSÃO

CRICIÚMA
2023

FÁBIO JEREMIAS DE SOUZA

MÍDIAS SOCIAIS, COMBATE À DESINFORMAÇÃO E MODERAÇÃO DE CONTEÚDO: ANÁLISE DAS DIRETRIZES INTERNACIONAIS DO DIREITO HUMANO À LIBERDADE DE EXPRESSÃO

Dissertação apresentada ao Programa de Pós-Graduação em Direito, Área de Concentração em Direitos Humanos e Sociedade, Linha de Pesquisa em Direitos Humanos, Cidadania e Novos Direitos da Universidade do Extremo Sul Catarinense - UNESC, como requisito parcial para a obtenção do título de Mestre em Direito.

Orientador: Prof. Dr. Gustavo Silveira Borges.

CRICIÚMA

2023

Dados Internacionais de Catalogação na Publicação

S729m Souza, Fábio Jeremias de.
Mídias sociais, combate à desinformação e moderação de conteúdo : análise das diretrizes internacionais do direito humano à liberdade de expressão / Fábio Jeremias de Souza. - 2023.
105 p. : il.

Dissertação (Mestrado) - Universidade do Extremo Sul Catarinense, Programa de Pós-Graduação em Direito, Criciúma, 2023.
Orientação: Gustavo Silveira Borges.

1. Liberdade de expressão. 2. Sociedade da informação. 3. Desinformação. 4. Gestão de conteúdo de internet. 5. Mídia social. I. Título.

CDD 23. ed. 341.2727

Bibliotecária Eliziane de Lucca Alosilla – CRB 14/1101
Biblioteca Central Prof. Eurico Back - UNESC

FÁBIO JEREMIAS DE SOUZA

““MÍDIAS SOCIAIS, COMBATE À DESINFORMAÇÃO E MODERAÇÃO DE CONTEÚDO: ANÁLISE DAS DIRETRIZES INTERNACIONAIS DO DIREITO HUMANO À LIBERDADE DE EXPRESSÃO”

Esta dissertação foi julgada e aprovada para obtenção do Grau de Mestre em Direito no Programa de Pós-Graduação em Direito da Universidade do Extremo Sul Catarinense.

Criciúma, 13 de dezembro de 2023.

Prof. Dr. Gustavo Silveira Borges
(Presidente e Orientador (a) – UNESC))

Profa. Dra. Cleide Calgaro
(Membro externo – PPGD/UCS)

Prof. Dr. Daniel Ribeiro Preve
(Membro – PPGD/UNESC

FABIO JEREMIAS DE SOUZA
Assinado de forma digital por FABIO JEREMIAS DE SOUZA
Dados: 2023.12.18 08:04:07 -03'00'

Fábio Jeremias de Souza
(Mestrando(a))

Prof. Dr. Yduan de Oliveira May
(Membro Suplente – PPGD/UNESC)

À Deus, à minha esposa Elke Minatto Steiner, aos meus filhos Igor Steiner de Souza e Caio Steiner de Souza e aos meus pais, Arnoldo Ido de Souza e Edinete Jeremias de Souza.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus pelo dom da vida, por me proporcionar tanta alegria, saúde e uma família linda, alicerces da realização desse grande objetivo.

Aos meus pais, Arnoldo Ido de Souza e Edinete Jeremias de Souza, que me ensinaram a jamais desistir, a alimentar a perseverança, me dotando dos elementos necessários para enfrentar os desafios e aproveitar as oportunidades da vida.

Um agradecimento especial à minha esposa Elke Minatto Steiner, amor da minha vida, minha colega de profissão, minha base. Sem ela essa tarefa seria impossível, pois os contratempos diários, a nossa rotina de trabalho e de compromissos familiares não seriam superados não fosse o seu amor e paciência. Aos meus filhos Igor e Caio, por serem a minha fonte de inspiração, os motivos de tudo, por terem me escolhido, por serem a prova da existência do Criador.

Ao meu orientador e amigo Gustavo Silveira Borges. Ao mesmo tempo em que é um grande pesquisador e estudioso do tema objeto dessa dissertação, com trabalhos atuais e relevantes no Brasil e no exterior, me presenteou em todos os momentos com sua generosidade, humildade e paciência.

Meus agradecimentos aos meus professores do Programa de Pós-graduação em Direito da Universidade do Extremo Sul Catarinense – UNESC, por proporcionar momentos de grande reflexão e aprendizado, uma verdadeira revolução na minha compreensão de mundo e do Direito.

Aos integrantes da banca de qualificação do projeto de dissertação, Professora Cleide Calgaro (PPGD/UCS) e Professor Daniel Ribeiro Preve (PPGD/UNESC) pelas considerações e apontamentos necessários para o desenvolvimento da pesquisa.

Aos meus amigos e colegas do Mestrado, pela parceria, cumplicidade e pelos momentos de troca de conhecimento e de experiências de vida.

“É o começo de uma nova existência e, sem dúvida, o início de uma nova era, a era da informação, marcada pela autonomia da cultura vis-à-vis as bases materiais de nossa existência. Mas este não é necessariamente um momento animador porque, finalmente sozinhos em nosso mundo de humanos, teremos de olhar-nos no espelho da realidade histórica. E talvez não gostemos da imagem refletida.”

Manuel Castells

RESUMO

A presente dissertação tem por objetivo geral estudar as mídias sociais, o combate à desinformação, sobretudo através da moderação de conteúdo, com uma análise das diretrizes internacionais do direito humano à liberdade de expressão. Como objetivos específicos, apresentam-se: a) pesquisa sobre a sociedade da informação e o direito à liberdade de expressão, avaliando os impactos tecnológicos e da desinformação na era da pós-verdade; b) a abordagem da emergência das mídias sociais, o tratamento jurídico da desinformação e o seu impacto; c) a apresentação das recomendações para o aprimoramento do combate à desinformação pelas principais mídias sociais, com foco na moderação de conteúdo, com análise das principais diretrizes internacionais e dos *standards* globais de direitos humanos. No que se refere ao problema de pesquisa, elege-se o seguinte questionamento, que se pretendeu responder com a presente dissertação: quais as diretrizes internacionais que podem ser utilizadas para o aprimoramento do combate à desinformação, com foco na moderação de conteúdo? A pesquisa adotou o método de abordagem dedutivo e o método de procedimento, o monográfico. Concluiu-se que as recomendações para o combate à desinformação com foco na moderação de conteúdo devem levar em consideração os *standards* de direitos humanos, especialmente o direito humano à liberdade de expressão. Ainda, que a moderação de conteúdo é importante instrumento de combate à desinformação, porém, há outras recomendações importantes, sobretudo o compromisso de combater o comportamento não autêntico, a definição universal do conceito de desinformação, o direito à privacidade no processo regulatório, a desmonetização de fornecedores que ampliam a desinformação e a capacitação do usuário, pesquisadores e comunidade de verificação de fatos.

Palavras-chave: Sociedade da Informação. Desinformação. Liberdade de Expressão. Moderação de Conteúdo.

ABSTRACT

This dissertation aims to study social media and the fight against misinformation, especially through content moderation, analyzing the international guidelines of the human right to freedom of expression. Specific objectives include: a) researching the information society and the right to freedom of expression, evaluating technological impacts and misinformation in the post-truth era; b) addressing the emergence of social media, the legal treatment of misinformation, and its impact; c) provide recommendations to enhance the fight against misinformation by major social media platforms, focusing on content moderation, with an analysis of key international guidelines and global human rights standards. Regarding the research problem, the following question is posed: what are the main recommendations to improve the fight against misinformation, with a focus on content moderation? The research adopted a deductive approach and a monographic procedure method. It was concluded that recommendations for combating misinformation with a focus on content moderation should consider human rights standards, especially the human right to freedom of expression. Furthermore, content moderation is an essential tool in combating misinformation, but there are other crucial recommendations, particularly the commitment to combat non-authentic behavior, the universal definition of misinformation, the right to privacy in regulatory processes, the demonetization of providers that amplify misinformation, and the training of users, researchers, and fact-checking communities.

Keywords: Information Society. Disinformation. Freedom of expression. Content Moderation.

LISTA DE ABREVIATURAS E SIGLAS

DUDH	Declaração Universal dos Direitos do Homem
PIDCP	Pacto Internacional de Direitos Civis e Políticos
TIC	Tecnologia da Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

A presente pesquisa dissertativa aborda a análise da discussão acerca da ascensão das mídias sociais, da desinformação e do seu combate através da moderação de conteúdo, numa incursão necessária nas diretrizes internacionais dos direitos humanos com foco no direito humano à liberdade de expressão.

A sociedade informacional é fruto da recente evolução histórica, da quarta revolução, da Tecnologia da Informação (TIC), que trouxe, dentre as suas inúmeras consequências, a ascensão das mídias sociais com a reunião de pessoas em rede, em número comparado à população dos maiores países do mundo, se considerarmos as maiores plataformas. Apesar da democratização do acesso, o aumento exponencial do fluxo de informação e a reunião de pessoas em grupos de afinidades algorítmicas relevou grandes problemas coletivos, destacando-se os impactos nocivos da pós-verdade e da desinformação.

O cenário enfraquece as instituições e ameaça a própria democracia, pois o ritmo frenético de dados e a intencionalidade de causar danos que caracteriza a desinformação, potencializada através dos algoritmos, do uso de robôs e contas falsas, faz com que as pessoas não tenham condições de processar toda a informação despejada de forma crítica. A democracia é ameaçada pois as mídias sociais formaram um campo fértil para discutir e definir assuntos que não passaram pelo crivo das urnas, influenciando questões relacionadas à segurança, à violência e à própria soberania. São muitos e recentes os exemplos de eleições e campanhas de saúde pública em que o despejo intencional de desinformação causou prejuízos inegáveis.

Dentre os mecanismos para o combate à desinformação, existe a moderação de conteúdo pelas plataformas, uma espécie de instância decisória das empresas de mídias sociais sobre o que pode ou não ser postado. A moderação é feita a partir das políticas e termos de uso, organizada muitas vezes pelos algoritmos, com uma tímida revisão humana, a depender do tamanho da plataforma. De toda forma, o combate à desinformação é algo extremamente sensível, pois lida com a democratização da *internet*, com os direitos humanos, sobretudo, com o direito à liberdade de expressão.

Diante disso, a dissertação tem como objetivo geral estudar as mídias sociais, o combate à desinformação, especialmente através da moderação de

conteúdo, com uma análise das diretrizes internacionais e do direito humano à liberdade de expressão.

Quanto aos objetivos específicos, tem-se que o primeiro será o de pesquisar a sociedade informacional e o direito à liberdade de expressão, avaliando os impactos tecnológicos e da desinformação na era da pós-verdade. Ainda, procurar-se-á apresentar um histórico conceitual de desinformação e o seu tensionamento com o direito humano à liberdade de expressão. Já o segundo objetivo específico abordará a emergências das mídias sociais, o tratamento jurídico da desinformação e o seu impacto. Serão apresentadas as principais iniciativas de governança e estratégias de combate à desinformação, além de apresentar a moderação de conteúdo pelas principais mídias sociais, com base nos padrões das comunidades, *guidelines*, cartas de princípios e iniciativas governamentais. Por fim, o último objetivo apresentará recomendações para o aprimoramento do combate à desinformação pelas principais mídias sociais, com foco na moderação de conteúdo, analisando as principais diretrizes internacionais e os *standards* globais de direitos humanos.

A justificativa do presente trabalho parte da premissa de que a humanidade está conectada em rede, vivendo a sociedade informacional. O avanço das Tecnologias da Informação (TIC) demonstra de forma inequívoca que a revolução ainda está em curso. Especificamente sobre as mídias sociais, no início de 2023 o *Facebook* possuía 2,989 bilhões de usuários, o *Youtube* 2,527 bilhões, o *Instagram* 1,628 bilhões, o *TikTok* 1 bilhão, o *LinkedIn* 922,3 milhões e o *Twitter* 372,9 milhões. Os usuários que formam as comunidades das principais plataformas de mídias sociais de fato superam, em números, a população de vários países. No Brasil, no início de 2023 haviam 181,8 milhões de usuários de *internet*, numa penetração de 84,3%, com 152,4 milhões de usuários de mídia social, equivalente a 70,6% da população total (DataReportal, 2023). É evidente que a ascensão das mídias sociais impactou de forma considerável o tecido social, com transformações culturais e sociais profundas. Sobre a desinformação nas redes, em 2016 os Dicionários Oxford nomearam “pós-verdade” como a palavra do ano, como consequência dos eventos que marcaram a votação do *Brexit* e a eleição presidencial dos Estados Unidos, caracterizados pela utilização de campanha massiva de desinformação. De outro lado, o combate à desinformação possui um natural tensionamento com o direito à liberdade de expressão e, de acordo com as principais cartas de princípios, autores e diretrizes

internacionais, o fato de que as principais plataformas operam em escala global, faz com que os direitos humanos sirvam de bússola para nortear a atuação das mídias sociais e dos Estados nessa difícil missão.

A dissertação alinha-se à temática proposta pelo Programa de Pós-Graduação em Direito da Universidade do Extremo Sul Catarinense – UNESC, área de Concentração em Direitos Humanos e Sociedade, na linha de Pesquisa em Direitos Humanos, Cidadania e Novos Direitos, considerando que pretende promover um estudo crítico sobre as recomendações das diretrizes internacionais que podem servir de base para aprimorar o combate à desinformação e a moderação de conteúdo aplicada pelas principais mídias sociais. A pesquisa também está relacionada e direcionada aos estudos desenvolvidos pelo Orientador Professor Dr. Gustavo Silveira Borges, coordenador do Grupo de Pesquisa “Direitos Humanos e Novas Tecnologias” que faz parte, inclusive do Grupo Permanente de Combate à Desinformação do Tribunal Superior Eleitoral.

Nesse sentido, a fim de se alcançar o objetivo geral e principal do presente estudo, elege-se o seguinte questionamento como problema de pesquisa: quais as diretrizes internacionais podem ser utilizadas para o aprimoramento do combate à desinformação, com foco na moderação de conteúdo?

Parte-se da hipótese de que a desinformação é um fenômeno que ganha cada vez mais espaço na era da informação, sendo potencializada pelo uso de algoritmos, inteligência artificial, contas falas e robôs. Parte-se igualmente da premissa de que o aumento do fluxo de informações, a crescente adesão dos usuários nas plataformas de mídias sociais e o modelo de negócio das empresas de tecnologia formam um ambiente em que as notícias sensacionalistas, negativas e de confirmação dos ideais de cada grupo chamem mais a atenção do usuário. As plataformas, como dito, operam em vários países e apesar de serem empresas privadas, a sua atuação possui consequências estruturais e coletivas que exigem a busca de alternativas para o combate à desinformação. Desta forma, para que se definam as diretrizes de combate à desinformação, com foco na moderação de conteúdo, deve-se ter como norte um direito internacional que possa ser aplicado em escala mundial, qual seja, os direitos humanos.

Dividiu-se a dissertação em três capítulos a partir dos objetivos específicos elencados. O primeiro capítulo, intitulado: *“Sociedade da (des) informação e direito*

humano à liberdade de expressão", abordará num primeiro momento o itinerário histórico das grandes revoluções da humanidade, iniciando pela agrícola e culminando com a terceira revolução industrial (revolução digital), com o desenvolvimento dos semicondutores, da computação *mainframe* (1960), do computador pessoal (1970-1980) e da internet na década de 1990. Estamos, porém, diante de uma quarta revolução, não mais restrita à indústria, posto que as novas descobertas, tais como o sequenciamento genético, a nanotecnologia, as energias renováveis e a computação quântica, refletem em várias áreas (Schwab, 2016, p. 19). E a expressão sociedade informacional ou sociedade da informação foi utilizada para substituir o conceito de sociedade pós-industrial, referindo-se às transformações técnicas, organizacionais e administrativas com foco não mais nos insumos baratos de energia e sim nos insumos oportunizado pelos avanços tecnológicos (Werthein, 2000, p. 71). Nesse cenário foi que emergiram as plataformas de mídias sociais, solo fértil para a propagação da desinformação, tanto que em 2016 os Dicionários Oxford elegeram a palavra "pós-verdade" como a palavra do ano, indicando o momento em que fatos objetivos possuem importância menor do que os apelos emocionais e as crenças pessoais (McIntyre, 2018, p. 5). O fato é que diante da complexidade e da nova realidade, Wardle e Derakhshan (2017, p. 5) introduzem três conceitos que caracterizariam a desinformação: *misinformation* (informações falsas compartilhadas sem a intenção de causar dano); *disinformation (*informações falsas compartilhadas com a intenção de causar dano) e *malinformation* (informações verdadeiras ou baseadas na realidade, mas indevidamente tornadas públicas, com a intenção de causar dano). A potencialização da desinformação a partir da atuação das mídias sociais é um problema a ser enfrentado, devendo ser considerado na discussão, no entanto, o tensionamento com o direito à liberdade de expressão, que deve ser respeitado com base no sistema universal de direitos humanos (Osório, 2022, p. 42).

O segundo capítulo, denominado: *"Mídias sociais, desinformação e o seu tratamento jurídico"*, parte da constatação de que o aumento exponencial da adesão dos seres humanos nas plataformas de mídias sociais, com interação através de compartilhamento de fotos, vídeos e participação em comunidades, fez da desinformação um grande problema a ser enfrentado. O vertiginoso crescimento das mídias sociais apresentou um novo paradigma de comunicação, transformando a sociedade contemporânea (Madakam; Tripathi, 2021, p. 7). O surgimento da *Internet*

representou a possibilidade de realização plena do direito de expressão livre, propiciando que todos possam ser jornalistas, formadores de opinião e editores de conteúdo (Bento, 2016, p. 102). E, para que se tenha ideia do impacto que a desinformação pode causar para a democracia, estudos em campanhas eleitorais constataram que as contas de *bots* são formadas por aproximadamente um quinto das conversas sobre política (Kumar; Shah, 2018, p. 6). Nesse contexto, foram várias as iniciativas governamentais e da iniciativa privada para conter a desinformação. Cita-se o Relatório Frances, o White Paper "*Online Harms White Paper*: *Full government response to the consultation*", do Reino Unido e as Cartas de Princípios, iniciativas da sociedade civil. As plataformas, de sua vez, editaram os padrões de cada comunidade e seus *guidelines*, linhas dentre as quais passaram a moderar o conteúdo do que é publicado.

No terceiro capítulo, intitulado: *"Diretrizes internacionais no combate à desinformação nas mídias sociais com foco na Moderação de Conteúdo"*, serão abordados os *standards* globais de direitos humanos relacionados ao tema, uma vez que é necessário estabelecer uma solução linear para responder aos enormes desafios impostos por essa nova realidade. Ainda, para compreender o estágio da discussão internacional sobre o combate à desinformação, serão abordados dois documentos recentes e de extrema relevância, o Relatório da ONU sobre Desinformação, Liberdade de Opinião e Expressão de 2021 e o Código de Conduta Reforçado da União Europeia de 2022. Por fim, buscar-se-á estabelecer recomendações no combate à desinformação, com foco na moderação de conteúdo e no direito à liberdade de expressão.

No que se refere à metodologia, a pesquisa adotará o método de abordagem dedutivo, partindo de premissas gerais com o fim de chegar a uma conclusão particular. Nesse sentido o método dedutivo "parte de princípios reconhecidos como verdadeiros e indiscutíveis e possibilita a chegar a conclusões de maneira puramente formal, isto é, em virtude unicamente de sua lógica" (Gil, 2008, p. 9). Parte-se do itinerário histórico que culmina com a sociedade informacional, dos conceitos de pós-verdade e desinformação, contextualizando com as principais iniciativas governamentais, da sociedade civil e das próprias mídias sociais no combate à desinformação com foco na moderação de conteúdo e nos direitos humanos. Ao final, buscar-se-á apresentar recomendações para o combate à

desinformação, com foco na moderação de conteúdo e no direito humano à liberdade de expressão. Em relação ao método de procedimento será adotado o monográfico considerando que “consiste no estudo de determinados indivíduos, profissões, condições, instituições, grupos ou comunidades, com a finalidade de obter generalizações” (Marconi; Lakatos, 2007, p. 93). Por sua vez, a técnica de pesquisa a ser desenvolvida no presente estudo será a bibliográfica uma vez que será “desenvolvida com base em material já elaborado, constituído principalmente de livros e artigos científicos” (Gil, 2008, p. 44).

Assim, a presente dissertação buscará, sem a intenção de esgotar todos as linhas de pensamento e inquietações pessoais, apresentar recomendações para amenizar os efeitos nocivos da desinformação, sobretudo no que se refere à moderação de conteúdo e sempre tendo como base os princípios universais de direitos humanos.

2 SOCIEDADE DA (DES)INFORMAÇÃO E O DIREITO HUMANO À LIBERDADE DE EXPRESSÃO

O presente capítulo descreve o contexto da sociedade da informação, a era da pós-verdade e a potencialização da desinformação com a evolução e ascensão das Tecnologias de Informação e Comunicação - TIC, bem como a relação com o direito humano à liberdade de expressão.

Em um primeiro tópico serão contextualizados a sociedade da informação e os impactos tecnológicos, sobretudo no que diz respeito à comunicação. Já em um segundo tópico, apresenta-se o fenômeno da pós-verdade, com o surgimento das chamadas *fake news*, enquanto o terceiro tópico trará o histórico-conceitual da desinformação e o tensionamento com o direito humano à liberdade de expressão.

2.1 SOCIEDADE INFORMACIONAL E OS IMPACTOS TECNOLÓGICOS

Antes de compreender o conceito e os impactos da pós-verdade e da desinformação, necessária uma abordagem da evolução histórica da sociedade desde as primeiras revoluções até a revolução digital (terceira revolução industrial) e mais recentemente, a da *internet* (quarta revolução). Importante ainda uma incursão nas revoluções da sociedade a partir da visão de alguns autores, mas sempre com o objetivo de compreender a sociedade informacional e os impactos tecnológicos para a vida no planeta.

A partir da evolução da tecnologia da informação (TIC), emergiu no fim do segundo milênio a revolução tecnológica, modificando a forma de vida ao dar novo sentido às relações humanas a partir de profundas transformações sociais, econômicas e políticas. Ainda, trouxe a interdependência global das economias, a organização das empresas conectadas em rede, possibilitando uma grande flexibilidade de gerenciamento, o que trouxe a necessidade de reestruturação do próprio modelo capitalista (Castells, 2002, p. 39).

Para Castells (2002, p. 67), a história é uma série de situações de estabilidade, com intervalos de eventos raros e importantes, contribuindo para a estabilidade subsequente. Na visão do Autor, no final do século XX, viveu-se um desses raros momentos pelo novo paradigma tecnológico, considerando a

microeletrônica, a computação, a telecomunicação, a optoeletrônica e a engenharia genética, incluídas dentre as tecnologias da informação. E os impactos para a vida no planeta são muitos. Apenas para contextualizar as transformações colossais causadas pelas novas tecnologias, como até bem pouco tempo não se tinha noção das consequências advindas do desenvolvimento da *internet*, outras revoluções distantes do radar político vêm ganhando forma atualmente, a exemplo daquelas causadas pelo avanço da inteligência artificial e da biotecnologia. O problema é que as estruturas democráticas atuais não são capazes de dar respostas a todas essas questões, a iniciar pelos eleitores que majoritariamente não possuem conhecimentos de biologia e cibernética, por exemplo. Assim, a democracia vai perdendo o controle dos fatos, abalando os pressupostos da confiança e do reconhecimento do eleitor pelo mecanismo democrático (Harari, 2016, p. 382).

Porém, para chegar aos impactos da tecnologia no mundo cotidiano, sobretudo no que se refere ao fluxo de informação, importante fazer um itinerário histórico das grandes revoluções. Schwab (2016, p. 18) aponta que a primeira revolução foi a agrícola, ocorrida há 10.000 anos, onde a busca por comida deu lugar à agricultura, através da combinação entre a domesticação dos animais e o esforço dos seres humanos. O resultado foi a urbanização, o aumento populacional e o surgimento das cidades. Após a revolução agrícola, várias outras surgiram a partir da segunda metade do século XVIII. A primeira revolução industrial (entre 1760 e 1840), foi marcada pela construção das ferrovias e invenção da máquina a vapor, dando início à produção mecânica. Já a segunda, no fim do século XIX e início do século XX), possuiu o seu desenvolvimento através da eletricidade e da linha de montagem, possibilitando a produção em massa.

A terceira revolução industrial, na década de 1960, foi caracterizada pela revolução digital, com o desenvolvimento dos semicondutores, da computação *mainframe* (1960), do computador pessoal (1970-1980) e da internet na década de 1990 (Schwab, 2016, p. 18). Assim, as três primeiras revoluções industriais mudaram o mundo nos últimos 250 anos, transformando a criação de valor pelos seres humanos e em cada um mudando não apenas a indústria, mas a vida e a interação entre as pessoas (Schwab; Davis, 2019). Porém, como uma consequência da terceira revolução, está-se diante da quarta revolução que teve início na virada do século, caracterizada por uma internet onipresente, móvel, pela inteligência artificial e pelo

aprendizado da máquina. E mais, a quarta revolução não está restrita à indústria, posto que as novas descobertas, tais como o sequenciamento genético, a nanotecnologia, as energias renováveis e a computação quântica, refletem em várias áreas (Schwab, 2016, p. 19).

O mundo vivencia o impacto dessas transformações tecnológicas, destacando-se que a partir da Segunda Guerra Mundial foram reveladas as principais descobertas no campo da eletrônica, quais sejam, o computador e o transistor, esse último, fonte da microeletrônica e ponto nodal da revolução da tecnologia da informação do século XX. O transistor foi inventado em 1947 pelos físicos Bardeen, Brattain e Schockley, permitindo a codificação da lógica e da comunicação das máquinas. Com a produção em escala de transistores, através da exploração do silício, sobreveio em 1957 o impulsionamento da microeletrônica com a criação do circuito integrado, evoluindo até o surgimento do microprocessador (1971), o que pode ser definido como o computador em um único *chip,* gestando assim a grande revolução (Castells, 2002, p. 76-77).

Em 1975, Ed Roberts criou a primeira caixa de computação, inspirando o *Apple* I e, posteriormente o *Apple* II, primeiro microcomputador com sucesso comercial, desenvolvido pelos jovens Steve Wozniak e Steve Jobs. Para não perder campo no mercado, em 1981, a IBM criou o seu microcomputador dando-lhe um nome de forte apelo comercial, o PC (Computador Pessoal). Porém, de fácil clonagem, a novidade da IBM permitiu a difusão de um padrão comum em todo o mundo, apesar da reconhecida superioridade da *Apple*. Ainda, catalisando a propagação dos microcomputadores, foram desenvolvidos os *softwares* de sistemas operacionais para PCs, idealizados por Bill Gates e Paul Allen na década de 1970, com a fundação da *Microsoft* (Castells, 2002, p. 79-80).

E foi justamente a partir da criação do computador pessoal que a máquina não ficaria mais restrita ao processamento de dados voltados às grandes empresas e à produção, tornando-se um instrumento de criação de textos, imagens, músicas, organização, simulação e até mesmo de diversão, ocorrendo uma grande fusão entre a informática e as telecomunicações, com o cinema e com a televisão (Lévy, 1999, p. 30). Por sua vez, o desenvolvimento das redes se tornou possível graças ao avanço das telecomunicações e da tecnologia de integração de computadores em rede, a

partir da utilização da fibra ótica, do laser e das redes de banda larga integradas, na década de 1990 (Castells, 2002, p. 81).

Já a *internet*, que anunciou a Era da Informação, foi criada e desenvolvida nas três últimas décadas do século XX, inicialmente como uma estratégia militar, pelo Departamento de Defesa dos Estados Unidos da América, a partir da ideia de criar uma rede independente de centros de controle, inabalável até mesmo por ataques nucleares. A Primeira rede foi a *Arpanet* (1969), precursora da *Internet* (década de 1980) e que foi privatizada apenas em 1995. Contudo, a revolução ganha corpo com a ascensão da telefonia móvel em 1997, possibilitando a utilização da *internet* para transmitir voz e dados, revolucionando as telecomunicações e sua respectiva indústria (Castells, 2002, p. 82-90).

Por todos esses motivos, deve-se concordar com Schwab (2016, p. 15), pois de fato vivencia-se uma quarta revolução que não se limita a uma fase da terceira. Ainda, ao contrário das revoluções industriais precedentes, a atual evolui de forma veloz e exponencial, com base na revolução digital. A quarta revolução, portanto, é a revolução da internet onipresente, móvel, acessível, da inteligência artificial e do aprendizado da máquina, com reflexos em várias áreas. A globalização e a velocidade da evolução das tecnologias digitais, tais como a Internet das Coisas (IoT), a inteligência artificial (IA) e a robótica, estão modificando significativamente a sociedade, tornando o meio ambiente e os valores das pessoas cada vez mais diversificados e complexos (Fukuyama, 2018, p. 47).

Como dito no início do presente tópico, é importante destacar que outros autores apresentam enfoques distintos em relação aos momentos históricos, tangenciando os conceitos anteriores. A visão do trajeto pode ser distinta, mas é ponto de convergência que estamos diante de profundas transformações, caracterizadas pela sociedade informacional. A depender das circunstâncias, a Quarta Revolução Industrial poderá oferecer oportunidades não só para aqueles que detiveram os benefícios e já desfrutam das revoluções industriais anteriores, mas também para aqueles que não desfrutam, como forma de buscar o desenvolvimento humano. A combinação das tecnologias deve permitir que as pessoas tenham a oportunidade de alcançar mais liberdade e índices adequados nas áreas da saúde, da educação e amenizar a insegurança causada pelas incertezas (Schwab; Davis, 2019).

Como exemplo de enfoque distinto, mas bastante similar no conteúdo, Harayama (2017, p. 10) faz um itinerário a partir do termo “sociedade”, guardando alguma sintonia com a visão de Schwab (2016), sendo que a sociedade 1.0 compreendia o ser humano como caçador e coletor, enquanto a sociedade 2.0 foi caracterizada pelo cultivo da agricultura. A Sociedade 3.0, de sua vez, compreendeu a industrialização através da Revolução Industrial. A Sociedade 4.0, mais recente, é caracterizada pela informação e a vida em redes de informação. É nesse contexto que surge a Sociedade 5.0, construída a partir e sobre a Sociedade 4.0, mas com foco em uma sociedade próspera e centrada no ser humano.

A ideia da Sociedade 5.0 é de uma construção planejada de uma sociedade em que as necessidades são atendidas fornecendo de forma adequada produtos e serviços necessários às pessoas, considerando-os de alta qualidade por meio das oportunidades geradas pelas novas tecnologias (Harayama, 2017, p. 10). Surgido pela primeira vez em 2016 no Japão como uma política de governo, o termo “Sociedade 5.0” pode ser definido como uma sociedade da inteligência, com a forte integração do espaço físico e do ciberespaço. Embora focado na humanidade, refere-se a um novo tipo de sociedade onde a inovação na ciência e na tecnologia ocupa um lugar de destaque, com o objetivo de resolver as questões sociais e garantir o desenvolvimento (Salgues, 2018, p. 1). Apesar do enfoque distinto, percebe-se a simetria com os autores que trabalham o termo revolução e a ideia de aplicar a tecnologia a serviço dos seres humanos. A questão, abordada também por Schwab e Davis (2019), para quem a Quarta Revolução Industrial deve ser vista não apenas sob a ótica de que a tecnologia é apenas uma simples ferramenta inevitável, mas sim de encontrar nela a oportunidade de “oferecer ao maior número de pessoas a capacidade de impactar positivamente a sua família, organização e comunidade, influenciando e orientando os sistemas que nos rodeiam e moldam nossa vida”.

Aliás, Schwab e Davis (2019) passam uma visão otimista ao assinalar que perspectivas lançadas no sentido de que as tecnologias estão fora do controle e possuem valor neutro, são enganosas. Propõem assim uma visão mais construtiva e focada numa abordagem mais centrada nos seres humanos, na medida em que as tecnologias são políticas e personificam os desejos e compromissos de sua criação, assim como as tecnologias e a sociedade vão se moldando umas às outras. As

tecnologias devem ser encaradas como soluções desenvolvidas através de processos sociais, carregadas de valores e prioridades.

Necessário na abordagem desse início de capítulo um recorte para destacar a contribuição de Floridi (2014, p. 87) que num contexto bastante distinto sobre a evolução da sociedade, apresenta as revoluções sob o enfoque da ciência, como modificadora da compreensão do homem de duas maneiras, uma sobre o mundo (extrovertida) e outra sobre ele mesmo (introvertida). A primeira revolução científica veio com a ruptura da ideia de que o homem seria o centro do universo, quando Nicolau Copérnico publicou em 1543 o seu tratado sobre os movimentos dos planetas ao redor do sol. A primeira revolução científica, portanto, demonstrou que o homem não é o centro do universo. Posteriormente à revolução copernicana, o homem ainda manteve a concepção de que era o centro do planeta, até que Darwin, em 1858, publicou a "Origem das Espécies por Meio da Seleção Natural", demonstrando que as espécies de vida evoluíram ao longo dos anos de ancestrais comuns por meio de uma seleção natural, inaugurando a segunda revolução científica (Floridi, 2014, p. 88). O homem já não era mais o centro do planeta.

Mesmo após Copérnico e Darwin, o homem mantinha a ideia de que a espécie era completamente responsável pelos seus próprios pensamentos. Porém, Sigmund Freud (1856-1939) quebrou essa ilusão através de seu trabalho psicanalítico, caracterizando a terceira revolução científica, com o argumento de que a mente também é inconsciente e sujeita a mecanismos de defesa como a repressão. Assim, hoje o homem é obrigado a reconhecer que não é o centro do universo, que faz parte do reino animal e que não possui mentes cartesianas totalmente transparentes (Floridi, 2014, p. 90).

Após as três revoluções, adveio a sugestão de Blaise Pascal (1623-1662) de que a dignidade do homem reside no pensamento. Contudo, o mesmo filósofo construiu a máquina capaz de realizar as quatro operações aritméticas, a calculadora pascalina. Posteriormente, Thomas Hobbes, em sua obra Leviatã, ao relatar a ideia inovadora, afirmou que pensar seria raciocinar, raciocinar seria calcular e calcular já poderia ser feito por um Pascalina. Porém, Pascal não havia considerado a possibilidade de projetar máquinas autônomas que poderiam superar a raça humana no processamento de informações lógicas, até que Alan Turing, pai da quarta revolução, demonstrou que o homem não é mais o mestre indiscutível da infosfera.

Turing publicou seu artigo clássico intitulado “Máquinas de Computação e Inteligência”, a partir do qual a ciência da computação e as TIC exerceram as mudanças sob o enfoque extrovertido (sobre o mundo), quanto introvertido (sobre o próprio homem) (Floridi, 2104, p. 91).

Portanto, não importa se no momento está-se vivenciando a quarta revolução trazida por Schwab (2016), ou a Sociedade 5.0, construída a partir da perspectiva do Governo Japonês ou ainda, a quarta revolução científica apresentada por Floridi (2014; 2015). O fato é que o momento é transformador e disruptivo. A partir desse momento, Castells (2002, p. 268) cunha o termo sociedade informacional numa alusão à organização do atual sistema de produção é baseada na potencialização da produtividade calçada no conhecimento, a partir da tecnologia da informação. Essa sociedade informacional é oriunda do avanço da tecnologia da informação e comunicação acima destacado, que como uma de suas consequências, aumentou exponencialmente o fluxo de informação.

A expressão sociedade informacional ou sociedade da informação foi utilizada para substituir o conceito de sociedade pós-industrial, referindo-se às transformações técnicas, organizacionais e administrativas com foco não mais nos insumos baratos de energia e sim nos insumos oportunizados pelos avanços tecnológicos. Esses avanços tecnológicos, aliás, possuem grande parcela da sua evolução como resultado da ação e das iniciativas do Estado, que visam ao desenvolvimento da tecnologia da informação (Werthein, 2000, p. 71-73).

Outro conceito importante nesse contexto é o de “ciberespaço”, palavra inventada em 1984 por Wiliam Gibson no romance de ficção científica Neuromante. O ciberespaço é definindo com o “o espaço de comunicação aberto pela interconexão mundial dos computadores e das memórias dos computadores” (Lévy, 1999, p. 92). Nesse contexto, considera-se que a perspectiva da digitalização das informações tornará o ciberespaço o principal canal de comunicação da humanidade. Ainda, considerando a primeira grande transformação na ecologia das mídias, que foi a passagem das culturas orais para a escrita, o crescimento do ciberespaço terá, ou já tem, um efeito tão radical quanto a invenção da própria escrita (Lévy, 1999, p. 113).

Ainda no campo da informação, Floridi (2014, p. 40) faz a relação com a mudança que as Tecnologias de Informação e Comunicação trazem para a própria natureza, trabalhando assim o termo infosfera. Cunhado na década de setenta, o

termo “infosfera” baseia-se no termo “biosfera” e diz respeito a todo o ambiente informacional, comparável com o ciberespaço, porém diferente dele pois abarca também os espaços de informação *offline* e analógicos. Contudo, o maior impacto das TICs é justamente a transição do analógico para o digital, com o aumento dos espaços informacionais dentro dos quais passa-se cada vez mais tempo (Floridi, 2014, p. 41).

As novas tecnologias vêm apagando o que Floridi (2014, p. 41) chama de “fricção informacional”, que seria a dificuldade de fluir a informação do remetente para o receptor. O autor exemplifica com a dificuldade de comunicação num ambiente barulhento como um *pub*, onde para se conseguir pedir uma cerveja, talvez sejam necessários alguns gestos. De outro lado, no caso do mundo virtual, a “supercondutividade de dados” facilita o fluxo de informação na infosfera, trazendo como consequências a dificuldade de ignorar a mensagem, o aumento exponencial no conhecimento comum, o aumento da responsabilidade (pois quanto mais informações estiverem a apenas um clique de distância, menos o usuário será perdoado por não as verificar) e a falta de privacidade informacional.

Aieta (2020, p. 216) assinala a existência de uma sociedade algorítmica, ou seja, uma sociedade em que as decisões sociais e econômicas são tomadas por algoritmos, robôs e agentes de inteligência artificial. De acordo com a autora, algoritmos são “sequencias finitas de ações executáveis que buscam obter uma solução para um determinado tipo de problema, podendo seu uso ser entendido como a aplicação de uma fórmula matemática para resolver um problema”. Os algoritmos formam as bolhas que são o conjunto de dados usados para fazer uma edição da informação, de forma invisível e que se destina a personalizar a navegação, dividindo as pessoas em grupos polarizados.

Feita a necessária contextualização, a revolução digital trouxe o encolhimento e a fragmentação dos movimentos sociais, com as pessoas tendendo a se agrupar em identidades primárias, tais como grupos religiosos, étnicos, nacionalistas e territoriais, onde a busca pela da identidade coletiva ou individual, do que as pessoas são, e não o que as pessoas fazem, constitui-se na principal fonte de significado. Há uma esquizofrenia estrutural que interfere na comunicação, não havendo sequer o debate social e político, com os grupos apenas enxergando o diferente como uma ameaça (Castells, 2002, p. 41). E isso ocorre, na relação que Lévy (1999, p. 130) faz entre ciberespaço (espaço de comunicação trazido pela

conexão dos computadores) com a “cibercultura", empregada como a aspiração de construção de um laço social, sem fronteiras territoriais, sem relações institucionais e de poder, pautada em pontos de interesses compartilhados.

Por outro lado, esse homem digital, numa visão trazida por Salgues (2018, p. 78), tem fácil acesso à informação, trazendo com isso algumas fraquezas como a sobrecarga (infobesidade) e dedicação de tempo cada vez mais destinado às mídias, assim como com o desaparecimento e a redução da influência do Estado-nação. Como ameaças, a manipulação da propaganda, o desapontamento diante da diferença entre o mundo real e o mundo virtual, realidade que por outro lado abre caminho para o novo, como a facilidade para o desenvolvimento da educação.

A interação e os vestígios nas mais variadas plataformas definem a presença digital do homem. Como consequências, a crescente polarização e a divulgação de informações imprecisas, falta de transparência em relação aos algoritmos de informações, pegadas digitais, publicidade direcionada, informações e notícias personalizadas, bem como a facilidade para criar movimentos sociais *on-line* (Schwab, 2016, p. 121). Aliás, os dados pessoais, que são literalmente minerados pelas plataformas de mídias sociais, podem ser considerados o petróleo da economia do século XXI (Jones, 2019, p. 8).

Prevendo a história que seria desenhada no futuro próximo, Harari (2016, p. 346) demonstrou que estudo recente apontava que o algoritmo do *Facebook* é mais assertivo do que pessoas próximas, amigos e familiares, na tarefa de definir a personalidade de um determinado ser humano, através do monitoramento dos *likes* nas páginas da web, imagens e clipes. Previu assim que, na próxima eleição presidencial norte americana, o *Facebook* poderia reconhecer as opiniões políticas, assim como identificar os votos que poderiam fazer a diferença no resultado e até mesmo, identificar a forma de mudá-los. E isso, como será visto, aconteceu. Os dados que deixamos na *internet*, portanto, são, de fato, recursos valiosos.

Assim, o impacto das novas tecnologias tem o poder de reverter a revolução humanista, repassando o poder aos algoritmos, até mesmo porque através das ciências biológicas se chegou à conclusão de que os organismos são algoritmos, derrubando o muro entre o orgânico e o inorgânico. Em estudos recentes, cientistas das biociências demonstraram que a emoção não é um fenômeno desconhecido, mas sim algoritmos bioquímicos. Algoritmo é método, ou seja, um conjunto de passos que

pode ser usado para fazer cálculos, resolver problemas e tomar decisões (Harari, 2016, p. 88-351). E continua o autor: “Os algoritmos do Google e do Facebook sabem não apenas como você se sente, como sabem 1 milhão de outras coisas a seu respeito das quais você mal suspeita” (Harari, 2016, p. 399).

Desta forma, uma das consequências do avanço tecnológico é que estamos diante de um novo sistema de comunicação pautado numa linguagem universal digital que promove uma integração global a partir das palavras, sons e imagens. E isso se dá com a personalização da informação ao gosto de cada indivíduo, através dos dados e vestígios na *internet*. Assim, num próximo item, buscar-se-á apresentar um fenômeno potencializado pela ascensão das novas tecnologias e pela sociedade da informação: a pós-verdade.

2.2 (DES) INFORMAÇÃO NA ERA DA PÓS-VERDADE

Como foi demonstrado no item acima, a ascensão da tecnologia da informação, sobretudo com o surgimento da *internet* que conectou o ser humano em uma rede mundial de informação trouxe mudança significativas para a vida no planeta. A velocidade da interação da comunicação fez surgir muitas oportunidades e ao mesmo tempo, muitos problemas, dentre eles o agravamento de um fenômeno antigo, qual seja, a relativização da verdade.

Em 2016 os Dicionários Oxford nomearam “pós-verdade” como a palavra do ano, como consequência dos eventos que marcaram a votação do *Brexit* e a eleição presidencial dos Estados Unidos, caracterizados pela ofuscação dos fatos, desapego aos padrões probatórios e a mentira nua e crua. A propósito, “pós-verdade” indica o momento em que fatos objetivos possuem importância menor do que os apelos emocionais e as crenças pessoais, com a ressalva de que o prefixo pós não remete necessariamente ao passado (sentido temporal), mas é trazido como significativo da ideia de que a verdade passa a ser irrelevante (McIntyre, 2018, p. 1-5). Estamos diante de um cenário que fez surgir o populismo ameaçador, com a razão suplantada pela emoção, com a diversidade perdendo para o nativismo e a liberdade para a autocracia (D’Ancona, 2018, p. 19).

O termo *pós-verdade* pode ser descrito, ainda, como um dado momento histórico em que, a expressiva velocidade da comunicação que multiplica as

informações que são repassadas, trazendo como consequência um reforço no posicionamento pessoal ao analisar se determinada informação é verdadeira ou falsa, diante do processo ideológico que caracteriza o processo de interpretação. Significa “uma sociedade que se importa mais com seu bem-estar diante das informações do que com a qualidade delas ou sua ligação com o real” (Siebert; Pereira, 2020, p. 243).

Alguns autores trazem o pós-modernismo como um elemento determinante para o colapso das narrativas oficiais, a partir de teóricos franceses como Foucault e Derrida, cujas ideias foram disseminadas nas universidades americanas na segunda metade do século XX. De um modo geral, as ideias pós-modernistas trazem a negativa de uma realidade objetiva que não pode ser revista pela percepção do homem, consagrando o princípio da subjetividade (Kakutani, 2018). Para Dunker (2017, p. 5) a pós-verdade é uma reação negativa à pós-modernidade, um falso contrário, uma segunda onda desta. Ainda, elemento preponderante para caracterizar a pós-verdade foi o advento da “década do eu”, impulsionada pelo desenvolvimento econômico do pós-guerra na década de 1970, fazendo com que as classes média e operária pudessem se envolver, com tempo e renda disponíveis, em atividades fúteis antes reservadas à aristocracia. Foi a inauguração da autogratificação e do desejo de atenção a qualquer custo (Kakutani, 2018).

Dunker (2017, p. 10) defende que apesar do batismo ter ocorrido em 2016, o marco inicial da pós-verdade tem relação com o ataque às torres gêmeas de Nova Iorque, oportunidade em que a relativização da verdade passou a ser tolerada em função da pauta dos costumes e pela guerra ao terror calçado na intolerância religiosa na perseguição aos muçulmanos. Ainda, em 2011 a verdade foi relativizada em relação à ficção sobre as armas químicas que justificaram o ataque ao Iraque. A Guerra do Iraque, que incendiou a geopolítica na região, dando origem ao Estado Islâmico, é uma verdadeira lição sobre as tragédias que podem ocorrer quando decisões de grande magnitude são tomadas sem o uso da razão e de ponderações de especialistas, mas sim inflamadas por uma convicção ideológica (Kakutani, 2018).

Quanto ao ano de 2016, foi marcado pelo discurso vencedor em campanhas políticas que trouxeram uma nova face conservadora ao mundo. Surge uma nova expressão cognitiva que faz surgir um novo irracionalismo trazendo temas como a objeção do criacionismo contra o darwinismo, os questionamentos à veracidade do aquecimento global, dentre outros (Dunker, 2017, p. 10). O *Brexit*, de

2016, justifica o marco dessa nova era. Trata-se da consulta popular convocada pelo Primeiro Ministro da Inglaterra que buscava rediscutir a permanência do Reino Unido na União Europeia (UE). O grande mote para os que defendiam a saída era justamente a imigração, destacadamente dos refugiados sírios, enquanto os que defendiam a permanência levantaram um discurso técnico pautado nos avanços advindos com o livre trânsito, o comércio e a moeda única. Aparentemente o cenário apresentava-se favorável à permanência, porém, a decisão pela saída surpreendeu. A utilização das mídias sociais, sobretudo o *Twitter,* foi decisiva, com a utilização de robôs e contas falsas (Ruediger; Grassi, 2018, p. 11)

Diante da ideia de se ligar emocionalmente as pessoas, Dominic Cummings, diretor da campanha favorável à saída do Reino Unido na União Europeia, definiu que a mensagem deveria ser clara e apegada aos ressentimentos, ao contrário dos que defendiam a permanência, que preferiram derramar números ao eleitorado. Cummings definiu o discurso propagado pelas mídias sociais, com foco, além da questão da imigração, na possibilidade de adoção do Euro pela Grã-Bretanha, o custo semanal de permanecer na União Europeia e a possibilidade, como uma ameaça, da Turquia ser integrada à comunidade. Dias após vencer o referendo, Daniel Hannan, membro do parlamento europeu pelo Partido Conservador Inglês concedeu entrevista em que relativizou o discurso contra a imigração, ao dizer que não foi proposta a redução, mas apenas o controle da imigração (D'Ancona, 2018, p. 28).

Outro marco da pós verdade, a eleição de Donald Trump à Presidência dos Estados Unidos, é relacionada ao mesmo fenômeno. Donaldo Trump sagrou-se vencedor a partir de uma campanha massissamente tecnológica, utilizando técnicas de estudo da personalidade dos usuários, com a compilação e a análise dos rastros digitais deixados na rede mundial de computadores (Rais *et al.*, 2018, p. 76). Como dito, a empresa inicialmente coletava, armazenava e tratava os dados dos usuários para, num segundo momento, minerar os dados (*data mining)* na busca incessante por indecisos. Um último trabalho consistia em criar uma polarização artificial através de métodos, com destaque para a difusão de notícias falsas (Fornasier; Beck, 2019, p. 189). Agrega-se a todos esses elementos, a participação de agentes russos na campanha presidencial de 2016, com o objetivo de minar a crença dos eleitores na democracia e no próprio sistema eleitoral americano como uma estratégia do Kremlin (Kakutani, 2018).

Ao se constituir em figura icônica e representativa do termo “pós-verdade”, mesmo após tomar posse em 2017, Donald Trump simplesmente despejou diversas informações sem qualquer base nos fatos, como por exemplo, que ele foi o maior vencedor das eleições norte americanas desde Ronald Reagan, que a sua posse foi a mais prestigiada em número de pessoas da história dos Estados Unidos e que a taxa de homicídios nos país era a maior em 47 anos, afirmações facilmente desmentidas por dados, imagens e relatórios oficiais (McIntyre, 2018, p. 2). Apesar de Turcilo e Obrenovic (2020, p. 5) entenderem que seria um exagero sugerir que ambos os eventos, o *Brexit* e eleição norte-americana de 2016, foram diretamente afetados pelas *fake news* e pela influência de potências estrangeiras, inegável a existência e a influência da desinformação antes e depois desses eventos.

Em sua obra, Empoli (2019) cita personalidades como Dominic Cummings, diretor de campanha do *Brexit*, Steve Bannon, o homem forte do *marketing* de Donald Trump, Milo Yiannopoulos, o blogueiro inglês que fez a transgressão virar do campo da esquerda para a direita e Arthur Finkelstein, conselheiro de Viktor Orban, que por sua vez é o porta-estandarte da Europa reacionária. O autor menciona que os engenheiros do caos reinventaram a propaganda política para adaptá-la à era dos *selfies* e das redes sociais. A ação dos engenheiros do caos é eminentemente populista, vez que as redes sociais garantem todos no mesmo plano, sem intermediários. O que conta é o número de curtidas. Como assinalam Siebert e Pereira (2020), a materialização da *pós-verdade* se dá através de memes, piadas, manchetes, livros e boatos, fundamentando determinada posições, de forma que a verdade vem antes do enunciado.

Um outro traço característico da pós-verdade é a negação da ciência, seja por motivos econômicos ou políticos, como foi o caso da indústria tabagista norte americana que por anos patrocinou estudos com o objetivo de negar o potencial cancerígeno de seus produtos. D’Ancona (2018, p. 19), menciona que a ciência é tratada com suspeição, praticamente desprezada. Já no contexto político, em maior medida há um campo aberto para a desinformação, por ser um ambiente em que se pode escolher livremente uma corrente ao invés de confirmar os fatos (McIntyre, 2018, p. 33).

O avanço tecnológico, inegavelmente, potencializa o fenômeno da pós-verdade. Nesse contexto, há preocupação com a desidratação e até mesmo o

desaparecimento da democracia, pois com o aumento na velocidade e volume de proliferação dos dados e da informação, as instituições, como partidos, eleições e parlamentos, tendem a ficar obsoletos. A causa não estaria sequer relacionada à questões éticas, mas sim no fato de que as instituições não teriam a capacidade de processar os dados de forma suficiente, pois evoluíram num período em que a movimentação política era mais rápida do que a tecnologia. Hoje o cenário é inverso. As mudanças são profundas e há uma dependência da vida cotidiana no ciberespaço e muito embora os projetos da *web* não tenham passado pelo crivo do processo democrático, eles influenciam em questões de soberania, fronteiras, privacidade e segurança. E mais, a internet é hoje uma terra sem lei, desgastando a soberania do próprio Estado, representando um risco para a segurança global (Harari, 2016, p. 380-381).

Bom ainda mencionar que uma das raízes mais profundas da pós-verdade, o viés cognitivo, está presente no ser humano há bastante tempo. O esforço para evitar o desconforto psíquico, faz com que o ser humano busque preservar o seu senso de autovalor. E há ainda outro aspecto da dissonância cognitiva que explica muita coisa, o fato de que as tendências irracionais para preservar o autovalor são reforçadas quando o ser humano está cercado de semelhantes que acreditam nas mesmas coisas (McIntyre, 2018, p. 35). Considerando que os algoritmos das mídias são programados para atrair o usuário com mais frequência e por mais tempo à plataforma, os engenheiros do caos de Empoli (2019) trabalham com as aspirações e sobretudo os medos dos eleitores. Assim, a nova propaganda política se alimenta de emoções negativas, garantindo mais participação.

Há um outro aspecto que explica a imposição da emoção sobre a razão. Em grande parte da história humana, as lendas e a mitologia eram compartilhadas para justificar o comportamento humano, sem qualquer apego às evidências e à verificação. Ocorre que a partir da Revolução Científica e do Iluminismo, as narrativas coletivas passaram a concorrer com a razão, com a racionalidade. Porém, com a revolução digital e a sociedade da informação, a emoção vem recuperando campo (D'Ancona, 2018, p. 38). Portanto, a *internet* que permitiu a democratização da informação e facilitou a transparência governamental, possibilitou também a utilização da rede para espalhar desinformação, preconceito e discurso de ódio, amplificando

muitas das dinâmicas em curso, desde a cultura do eu, do *selfie* e da reunião de pessoas em bolhas ideológicas (Kakutani, 2018).

O fluxo de informações, como visto no item anterior, foi potencializado pelos avanços das novas tecnologias até a difusão dos *smartphones*, resultando no ciclo contínuo de notícias individuais com a democratização das mídias que utilizam algoritmos para impulsionar as bolhas. Esses componentes levam os usuários a diminuir o tempo que eles gastam em uma notícia específica, consumindo muitas vezes apenas o título e o apelo visual, sem muito apego à fonte e o texto completo. É inegável o impacto na sociedade, produzindo um ambiente em que as teorias conspiratórias sem suporte em fatos comprovados se sobreponham às ideias dos especialistas (Starts, 2021, p. 24). A democratização da informação e, por consequência, do discurso público, trouxe grandes desafios e consequências, pois a vida em rede potencializou os conteúdos nocivos e indesejados, verdadeira ameaça, como a desinformação, discurso de ódio e a polarização extremista (Osório, 2022, p. 97).

Conforme já foi também abordado, o engajamento dos usuários nas redes é feito através dos algoritmos, sendo que a ideia é mantê-los o máximo de tempo conectados, vendendo mais anúncios. A forma como se mantem o usuário conectado é feito a partir da coleta e análise dos dados, dos vestígios. O procedimento, não só faz as pessoas permanecerem reunidas em bolhas ideológicas, mas facilita mensagens simplistas e sensacionalistas, com base na emoção, nas crenças e, sobretudo, na desinformação, pois as pessoas tendem a se conectarem emocionalmente com o conteúdo (Kakutani, 2018). As redes ou mídias sociais, portanto, reduzem a tolerância com relação às visões alternativas, amplificam a polarização, aumentam a probabilidade de aceitação de notícias ideologicamente compatíveis com os usuários, bloqueando-os para novas informações e, como consequência, criam um contexto para que as notícias falsas atraiam um público de massa. Pesquisas demonstram que pessoas preferem informações que confirmem suas atitudes preexistentes (exposição seletiva), preferem visualizar informações relacionadas às suas crenças (confirmação preconceito) e inclinam-se a aceitar informações que os agrada (Lazer *et al.*, 2018, p. 1095).

Outra característica do estágio atual, é o declínio das mídias tradicionais. Assim, antes de falar da ascensão das mídias sociais, deve-se tratar desse fato. No

apogeu dos principais veículos de imprensa eram responsáveis pela veiculação quase que total da informação (McIntyre, 2018, p. 64). Hoje, os guardiões dos valores jornalísticos estão numa situação difícil enquanto assistem à participação no mercado diminuir diante da crescente difusão do conteúdo baseado em opinião, além de serem acusados de possuírem tendência, mesmo nos esforços de defender a verdade. Donald Trump, por exemplo, passou a chamar as matérias jornalísticas que o contrariam de “notícias falsas”. O público das notícias, formado por muitos partidários, tornou tênue a linha entre a mídia tradicional e a alternativa, uma vez que a ascensão das mídias sociais facilitou o acesso gratuito à informação (McIntyre, 2018, p. 87). Os sites corporativos e as mídias tratam com desdém os jornais impressos, a grande mídia, relacionando-os à uma teoria da conspiração de uma ordem globalista e liberal (D’Ancona, 2018, p. 20).

Evidentemente, o declínio da mídia tradicional tem como grande motivo a *internet*. Antes os jornais perderam espaço para os canais de TV a cabo, mas foi com difusão da *internet* na década de 1990 que os jornais impressos, por exemplo, passaram a encolher (McIntyre, 2018, p. 89). Quanto às mídias, o *Facebook* foi criado em 2004, permitindo a conexão entre amigos existentes e outros, com a possibilidade de compartilhar os pensamentos e de participar de uma comunidade sobre um assunto de preferência. Com o crescimento, o *Facebook* ganhou força como um agregador de notícias. Com o mesmo objetivo, mas com suas particularidades, o *YouTube* foi fundado em 2005 e o *Twitter* em 2006. Portanto, a ascensão da mídia social como fonte de informação trouxe ainda mais confusão entre o que é notícia e o que é opinião, já que as pessoas tendem a compartilhar histórias de *blogs* e *sites* alternativos, como se fossem todas verdadeiras (McIntyre, 2018, p. 89).

O fato é que a obtenção das informações se dá cada vez mais através das mídias sociais do que das fontes tradicionais de notícias e considerando todo esse cenário, tem-se que o grande impacto que as informações falsas causam cotidianamente, criou um dos maiores perigos para a sociedade atual, de acordo com o Fórum Econômico Mundial (Kumar; Shah, 2018, p. 1). A mentira, aliás, é integrante da política desde que os primeiros homens se organizaram em tribos (D’Ancona, 2018, p. 32). As notícias falsas ou *fake* news, conforme alguns sustentam, foram inventadas com o próprio conceito de “notícia”, desde que Johannes Gutenberg inventou a imprensa em 1439. Notícias falsas foram divulgadas através dos tempos,

durante a revolução científica e o Iluminismo, continuaram nos Estados Unidos e em outros lugares muito depois disso, mas finalmente um padrão de “objetividade” começou a surgir, com a criação da *Associated Press* em 1948. Por se constituir numa organização a partir da associação de jornais de várias correntes, a *Associated Press* mantinha a neutralidade em seus relatos editoriais. Isso não significa que não haviam notícias falsas, uma vez que haviam vários jornais de linhas editoriais diferentes e o sensacionalismo foi a grande mola propulsora do desenvolvimento do jornalismo (McIntyre, 2018, p. 97-100).

O cenário favoreceu a proliferação da mentira, estabelecida por McIntyre (2018, p. 9) como intencional ou não, ressaltando que o ápice da pós-verdade é justamente quando o engano e a ilusão atingem alguém que de fato acredita em uma inverdade, mesmo que todas as fontes confiáveis a estejam contestando. Nesse contexto, o avanço da tecnologia da informação propulsiona esse fenômeno, pois a distribuição de textos, imagens, vídeos curtos, animações e *memes* de forma instantânea nas plataformas de mídias sociais instaurou a supremacia da imagem e do instantâneo sobre as palavras. Assim, no momento em que a quantidade é mais importante do que a qualidade, tem-se um campo fértil para a proliferação das chamadas *Fake News*, conhecidas como narrativas alternativas aos fatos (Passarelli; Gomes, 2020, p. 255).

As informações falsas não teriam tanto impacto se os leitores pudessem identificá-las facilmente, mas como visto a lógica por trás do engano bem-sucedido é evidente em várias pesquisas que demonstram de forma muito clara que os humanos são maus juízes em relação à mentira, seja por preconceito, falta de educação ou baixo consumo de informação (Kumar; Shah, 2018, p. 2). Ainda, é característico da pós-verdade o fato de que é muito mais importante do que interpretar determinado acontecimento, criar uma versão ao fato para moldar ao que o indivíduo prefere interpretar, mesmo que sem base científica, o que é facilitado pela circulação desenfreada da informação, como forma de dar estabilidade aos sentidos (Siebert; Pereira, 2020, p. 244).

Como se verifica, a proliferação da mentira, a relativização dos fatos e da ciência, que revelam o fenômeno da pós-verdade, é um problema a ser enfrentado, sobretudo pela ascensão das mídias sociais e das novas tecnologias. Porém, a definição do conceito de desinformação é uma tarefa árdua e que vem ganhando a

atenção do mundo acadêmico, assim como o seu combate, que pressupõe a análise do direito humano à liberdade de expressão e opinião, assuntos que serão abordados no item seguinte.

2.3 DESINFORMAÇÃO: HISTÓRICO-CONCEITUAL E O TENSIONAMENTO COM O DIREITO HUMANO À LIBERDADE DE EXPRESSÃO

Como abordado nos tópicos anteriores, uma das consequências da sociedade da informação e da era da pós-verdade, ambas impulsionadas pelo avanço das TICs e das plataformas de mídias sociais, foi o agravamento da disseminação da desinformação. O agravamento está relacionado também com o desenvolvimento de uma estratégia muito bem definida, com a mineração e tratamento dos dados dos usuários, utilização de contas falsas, robôs e sobretudo, com intenção de obter um ganho, seja econômico ou político.

A sociedade atual tem como uma de suas principais características a influência do uso da tecnologia digital surgindo, como consequência, a preocupação com a proteção dos direitos humanos (Gonzáles-Iza, 2021, p. 1). O crescente número de usuários gerou transformações latentes na comunicação e na interação entre as pessoas, passando de um modelo público de comunicação de “poucos para muitos”, para uma estrutura de “muitos para muitos” (Sander, 2021, p. 163). Ainda, o aumento do fluxo de informação ocasionou a redução do tempo dedicado à própria informação, fazendo com que as pessoas prestem mais atenção no título das matérias, no apelo visual, com pouca dedicação à fonte e ao conteúdo completo (Starts, 2021, p. 25).

Apesar das notícias falsas não serem uma novidade, presenciamos hoje uma difusão em escala alarmante, uma verdadeira poluição informacional global, com várias formas de criar, divulgar e consumir as mensagens, com a transmissão nos mais diversos formatos, através de técnicas que amplificam o seu alcance (Wardle; Derakhshan, 2017, p. 11). Deve ser considerado ainda que o modelo de negócio das mídias, baseado nos algoritmos e na publicidade, com o direcionamento da atenção dos usuários para os conteúdos extremistas e sensacionalistas, além de permitir a rentabilização do conteúdo através de anúncios, demonstra que a própria *internet* oferece os meios para a disseminação em larga escala da desinformação, com destaque para as plataformas de mídias sociais (Lazer *et al.*, 2018, p. 1096). Assim,

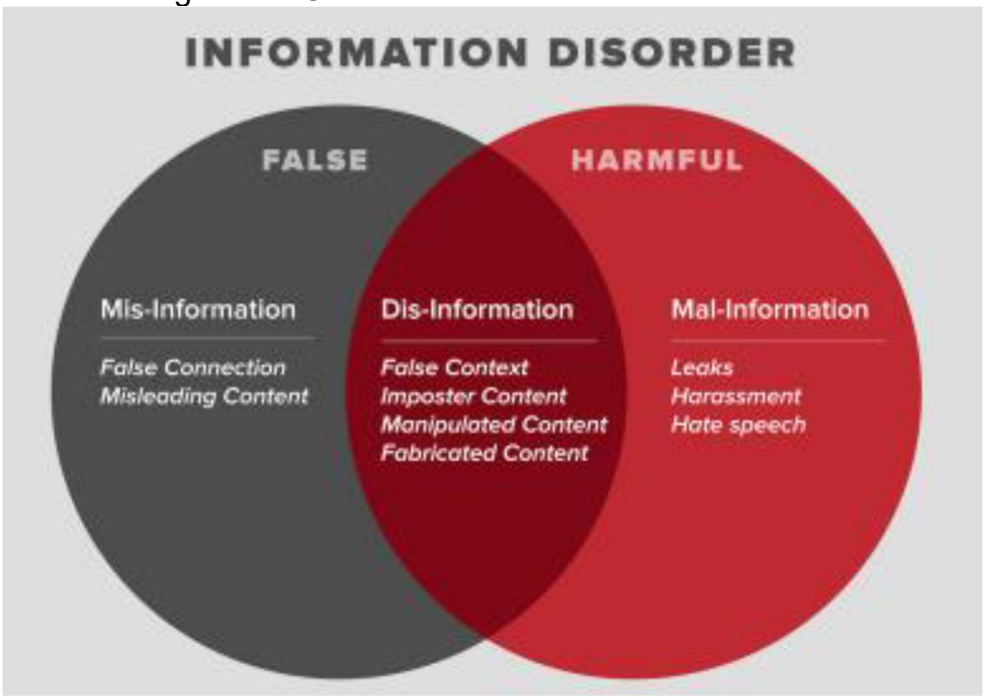
o modelo de negócios das plataformas é um campo fértil para propagar a desinformação, uma vez que se baseia na receita advinda da publicidade e, através de seus algoritmos, tem como intenção viciar os “usuários”, prendendo sua atenção e mantendo-os o máximo de tempo conectados (Jones, 2019, p. 6).

Com a crescente a adesão dos seres humanos às mídias sociais e com o aumento exponencial do fluxo de informação, o campo se tornou fértil para a proliferação de informações distorcidas ou inverídicas, seja por interesses políticos ou financeiros, fenômeno esse designado inicialmente como *fake news* (Chirwa; Manyana, 2021, p. 60). Contudo, o conceito de desinformação é uma questão tormentosa, muito embora existam alguns parâmetros definidos pelos grandes estudiosos do tema.

Inicialmente foi popularizado o termo *Fake News* (notícias falsas), como forma de utilizar a saliência do termo para a discussão de um importante tema. *Fake News* seriam informações fabricadas que imitam as mídias de notícias na forma. No entanto carecendo de normas e processos editoriais como forma de garantir a veracidade e credibilidade (Lazer *et al.*, 2018, p. 1094). Mais recentemente, autores e organismos internacionais têm optado pelo termo desinformação, mais completo e não restrito ao termo *fake* news, que ficaria limitado à produção de notícias. O termo, portanto, seria inadequado para abordar um tema tão complexo (Wardle; Derakhshan, 2017, p. 5).

Turcilo e Obrenovic (2020, p. 5) trazem como origem da palavra desinformação o Dicionário da Língua Russa de 1949, com o significado de o “ato de enganar com a ajuda de informações falsas”, supostamente utilizado para descrever a propaganda política do nazismo. O fato é que diante da complexidade e da nova realidade, Wardle e Derakhshan (2017, p. 5) por entenderem que o termo *fake news* não consegue alcançar as diversas facetas da desordem informacional, introduzem três novos conceitos que caracterizariam a desinformação. O primeiro seria o conceito de *misinformation,* que compreende as informações falsas compartilhadas sem a intenção de causar dano. O segundo, de *disinformation*, que refere-se às informações falsas compartilhadas com a intenção de causar dano. Por último, a *malinformation*, termo empregado para definir as informações verdadeiras ou baseadas na realidade, mas indevidamente tornadas públicas, com a intenção de causar dano. A gravura abaixo demonstra os conceitos:

Figura 1 – Conceitos da desordem informacional



Fonte: Wardle e Derakhshan (2017, p. 5).

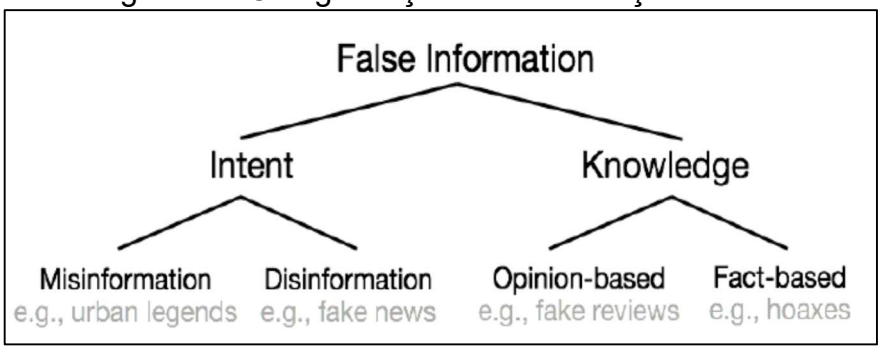
A desinformação foi categorizada em informações falsas com base na intenção e conhecimento do conteúdo. Sem trazer a *malinformation,* Kumar e Shah (2018, p. 2), classificou que, no que se refere à intenção, as informações podem ser intituladas com *misinformation,* criada sem a intenção de enganar e *disinformation*, o oposto. Já no que se refere ao conhecimento, as informações falsas serão classificadas como “baseadas em opinião”, onde não há uma verdade única ou “baseada em fatos”, que revelam mentiras sobre questões que possuem um único valor de verdade.

As causas da *misinformation* estão relacionadas com a deturpação ou a distorção de uma informação original que seria verdadeira, seja devido à falta de compreensão, atenção ou até mesmo por viés cognitivo, fazendo com que os atores distribuam a desinformação de forma involuntária. Já a *disinformation*, espalhada com a intenção de enganar, se concentra em influenciar a opinião pública ou em direcionar o tráfego *on-line* para ganhar dinheiro com anúncios (Kumar; Shah, 2018, p. 4).

As informações falsas categorizadas pelo conhecimento do agente, como visto, são baseadas em opiniões, expressando opiniões individuais, honestas ou não e descrevem casos em que não há uma verdade absoluta, como por exemplo as avaliações de produtos em sites de comércio eletrônico. As informações falsas

baseadas em fatos, de sua vez, envolvem a contradição de uma verdade de valor único. Para dificultar a análise do leitor entre informação verdadeira ou falsa, inclui notícias falsas, rumores e fraudes fabricadas (Kumar; Shah, 2018, p. 4).

Figura 2 – Categorização das informações falsas



Fonte: Kumar e Shah (2018, p. 4).

Ang, Anwar e Jayakuma (2021, p. 7) classificam a desinformação como: a) falsidades e rumores conscientemente distribuídos; b) falsidades e rumores propagados sem um objetivo político amplo e c) falsidades distribuídas para fins de ganhos financeiros. No primeiro grupo, essas falsidades conscientemente distribuídas são propagadas como parte de uma agenda política com base em viés ideológico ou como parte de uma campanha de desinformação patrocinada pelo Estado, estando aqui inserida a *disinformation*. Quanto ao segundo grupo, mesmo sem um objetivo político amplo, mesmo não sendo um produto malicioso ou coordenado caracteriza a *misinformation*, porém, pode atingir um status viral e causar danos.

Corroborando o pensamento, um dos grupos que se utiliza da *disinformation* busca o ganho financeiro, com a reação do público e dos cliques, vendendo os anúncios de propaganda. O outro grupo, que de certa forma beneficia-se do modelo de negócios das mídias, tem como intento a vantagem política, usando a zona cinzenta de desinformação para alavancar suas ideias, atacar adversários ou promover uma causa específica. Mas há ainda os atores estatais, que utilizam a *disinformation* para afetar a sociedade, seja por meios de massiva desinformação de forma coordenada. Um exemplo foi a ocupação da Crimeia pela Rússia, com o domínio de longa data da mídia pelo Kremlin nesta parte da Ucrânia, ajudando a moldar as percepções do público local, que desenvolveu o hábito de assistir as mídias russas como principal fonte de informação. Essas táticas são usadas pela Rússia nos

estados bálticos e em vários outros países, ou seja, há o estabelecimento de pré-condições para o sucesso da campanha de influência (Starts, 2021, p. 26).

Aliado ao estabelecimento de pré-condições, os atores estatais também podem se utilizar da exploração de vulnerabilidades, ou seja, das fraquezas sociais como meio para a ruptura da coesão social, buscando relacionamentos com atores locais que podem se beneficiar desse discurso. E por fim, a abordagem abrangente facilitada pelas fraquezas sociais, com o alcance dos efeitos desejados através da utilização da mídia tradicional, dos *trolls,* redes robóticas e vazamentos de documentos através do hackeamento, utilizando as mídias sociais com o objetivo final de desenvolver uma história consistente através da ilusão de que muitas vozes ativas estão engajadas (Starts, 2021, p. 26).

Há ainda que se definir as fases do transtorno de informação, quais sejam, a criação, a produção e a distribuição. De acordo com Wardle e Derakhshan (2017, p. 23) essa subdivisão é importante pois o agente que cria a mensagem pode ser diferente do agente que produz. Ademais, a partir do momento em que ela é distribuída, pode ser reproduzida e novamente distribuída por uma infinidade de agentes e pelas mais diversas motivações. Aliás, juntamente com a motivação e o interprete, a agente é um dos elementos do transtorno da informação e faz parte de todas as três fases.

Quanto às mensagens, podem ser comunicadas pelos agentes pessoalmente, através de fofocas, discursos, através de texto (artigos de jornais, por exemplo) ou através de material audiovisual. Por fim, o intérprete está relacionamento ao público, que não pode ser considerado um receptor passivo, pois cada um dos indivíduos irá interpretar as mensagens de acordo com o seu contexto, pois o ser humano tem dificuldade em aceitar informações que desafiam o senso de identidade (Wardle; Derakhshan, 2017, p. 26).

A democratização da informação trouxe alguns desafios, vez que as plataformas de mídias trouxeram ao mesmo tempo as condições necessárias para a propagação massiva de conteúdos indesejados, tais como incitação à violência, discurso de ódio, ataques ao processo democrático, moldando a opinião pública de acordo com as condições e métodos do agente. A partir dessa constatação, surgem questões a serem enfrentadas, como a governança e a regulação, que serão abordadas no próximo capítulo. O que importa aqui nessas primeiras linhas, é que o

enfrentamento da desinformação trará repercussões sobre a privacidade, liberdade, sobre a capacidade e até mesmo legitimidade dos mecanismos de moderação de conteúdo, com riscos de censura tanto por parte dos agentes estatais, como pelas próprias mídias, que atuam no plano privado (Osório, 2022, p. 97). Especificamente em relação à desinformação, a sua potencialização a partir da atuação das mídias sociais é um problema a ser enfrentado. Contudo, deve ser considerado na discussão o direito humano à liberdade de expressão.

O sistema de liberdade de expressão estabelecido na Constituição da República Federativa do Brasil possui três liberdades: a) liberdade de expressão *stricto senso*, como sendo o direito de externar opiniões, ideias, criações, sentimentos e toda as formas de expressão; b) liberdade de informação, correspondente ao direito de comunicação e transmissão de fatos, ao direito de acesso à informação confiável sobre fatos, ao direito de informar e ao direito de ser informado; c) liberdade de imprensa que corresponde ao direito de qualquer meio de comunicação de exteriorizar ideias e opiniões (Osório, 2022, p. 40).

Ainda, o direito à liberdade de expressão deve ser analisado com vistas a respeitar o sistema universal, como por exemplo o Pacto Internacional dos Direitos Civis e Políticos da Convenção Americana de Direitos Humanos, a Declaração Universal dos Direitos do Homem, a Declaração Americana dos Direitos e Deveres do Homem, a Carta Democrática Interamericana da Organização dos Estados Americanos e a Convenção Americana sobre Direitos Humanos (Osório, 2022, p. 42).

Deve ser reconhecido, outrossim que o direito humano à liberdade de expressão não é absoluto, sendo que até mesmo nos Estados Unidos da América que, em razão da força da Primeira Emenda, trata tal direito de forma cautelosa, a Suprema Corte admite restrições, ainda que em caráter excepcional. No Brasil, apesar do texto constitucional não estabelecer limites expressos à liberdade de expressão, é certo que as limitações decorrem de outros preceitos constitucionais, tais como a honra, a vida, a personalidade, a dignidade da pessoa humana, a infância e a juventude. Contudo, mesmo tais limitações devem observar o princípio da reserva legal, a proteção de outros interesses com elevado valor axiológico e o princípio da proporcionalidade em sua tríplice dimensão, assim entendido como respeitando a adequação, necessidade e proporcionalidade em sentido estrito (Osório, 2022, p. 90).

A Convenção de Direito Humanos, por sua vez, prevê de forma taxativa os limites da liberdade de expressão. São elas, “a proteção dos direitos e reputação das demais pessoas, e a proteção da segurança nacional, da ordem pública ou da saúde ou moral públicas” (Osório, 2022, p. 92).

E mesmo nos casos de restrição deve-se garantir a adoção de remédios a *posteriori*, tais como o direito de resposta, retratação ou mesmo a indenização, evitando-se a censura prévia e sempre buscando conferir à liberdade de expressão uma margem maior de tolerância (Osório, 2022, p. 94). Assim, quando dois direitos entram em conflito, deve haver o sopesamento dos bens jurídicos de modo a verificar se o agente ultrapassa as regras, com uma ponderação baseada no critério da proporcionalidade, sempre tendo em conta que as restrições à liberdade de expressão devem ser interpretadas de forma restritiva e adequadas ao objetivo legítimo a ser perseguido. Por exemplo, numa abordagem civilista, a liberdade de expressão possui três tipos de limitação: a) necessidade de tutela de outros bens jurídicos; b) limitações dos poderes públicos e c) limitações pela via contratual, tais como o direito de exclusividade de um artista ou regras de confidencialidade (Barbosa, 2021, p. 738).

Evidentemente, que há muitos desafios, tensionando o direito humano à liberdade de expressão no ambiente *online* com a divulgação maciça de desinformação, além da proliferação de diversos conteúdos impróprios e lesivos, tais como o discurso de ódio, da incitação à violência e do racismo. De outro lado, a atuação das mídias pode caracterizar um contexto de clara disparidade entre os autores envolvidos, o que necessita que estudos sejam pautados pela garantia dos direitos fundamentais para determinar o grau de autonomia das mídias (Sarlet; Hartmann, 2019, p. 98). Assim, o caminho para estabelecer esses limites tem sido objeto de muitos estudos. Num primeiro aspecto, deve-se partir da constatação de que as principais mídias sociais reúnem bilhões de usuários e atuam em centenas de países, sendo que nesse contexto os direitos humanos podem ser tomados como linhas de convergência trazidas pelo direito internacional.

Jones (2019, p. 31) enumera quatro direitos fundamentais como base para o combate à desinformação: a) o direito à liberdade de pensamento e de ter opiniões sem interferência; b) o direito à privacidade; c) o direito à liberdade de expressão; d) o direito de votar nas eleições. Portanto, o direito à liberdade de expressão deve ser garantido.

Os direitos à liberdade de pensamento, opinião e expressão devem ser melhor esmiuçados, pois interessam ao presente estudo. O direito de liberdade de pensamento e opinião está intrinsicamente ligado ao direito humano à liberdade de expressão, encontrando sua matriz no art. 19 da Declaração Universal dos Direitos do Homem (DUDH), que estatui que “Toda pessoa tem direito à liberdade de opinião e expressão; este direito inclui a liberdade de opinião sem interferência e buscar, receber e difundir informações e ideias por qualquer meio e independentemente das fronteiras” (ONU, 1948). Ainda não há clareza entre os limites da liberdade de pensamento, de consciência e de religião, e a liberdade de ter opiniões, associadas à liberdade de expressão, sendo que o Comitê de Direitos Humanos da ONU declara a liberdade de opinião e expressão como indispensáveis para o pleno desenvolvimento da pessoa, essenciais para qualquer sociedade, constituindo-se em pedra fundamental de todas as liberdades e da democracia (Jones, 2019, p. 33).

Muito embora a natureza declarada e absoluta do direito, somos receptores de frequentes tentativas de influência aos nossos pensamentos e opiniões, como por exemplo ocorre com a mídia e a publicidade. Isso porque deve ser considerado que a liberdade de pensamento inclui tanto o direito de não ter uma opinião inconscientemente manipulada, além do direito de não revelar pensamentos ou opiniões, bem como não ser penalizado por seus pensamentos. Já sobre a liberdade de expressão, a questão mais destacada é estabelecer os parâmetros da fala adequadamente protegida pela liberdade de expressão, sobretudo no ambiente *online*, com o potencial para grande quantidade de informação com alcance transnacional, em escala de alta velocidade e sem qualquer filtro editorial. O direito à liberdade de expressão e o combate às informações indesejadas é ainda mais difícil diante das lacunas na orientação internacional e nacional, cenário que é agravado pela falta de transparência das mídias em relação à moderação de conteúdo (Jones, 2019, p. 33-43).

Num sentido semelhante, importante destacar o Relatório da Relatora Especial sobre a promoção e proteção do direito à liberdade de opinião e expressão da ONU, Irene Khan, que apresenta duas dimensões da liberdade de opinião, sendo uma interna, relacionada ao direito à vida privada e à liberdade de pensamento, e uma externa, que vem relacionada com liberdade de expressão. O Relatório considera o direito à liberdade de opinião absoluto, ao contrário do direito à liberdade de expressão

que muito embora deva ser amplo e inclusivo, elenca as possibilidades de restrição nas balizas no artigo 19, parágrafo 3, do Pacto Internacional sobre Direitos Civis e Políticos: a) as restrições devem estar prescritas em lei; b) a prescrição legal deve ser necessária para o propósito de respeitar os direitos humanos, a reputação de terceiros, a proteção da segurança nacional, a ordem pública, a saúde ou a moral pública (ONU, 1966, p. 7).

Harari (2016, p. 390) vai mais além e aduz que não se deve confundir liberdade de informação com o velho ideal liberal da liberdade de expressão, pois a segunda (liberdade de expressão), foi dada aos humanos e a primeira (liberdade de informação), foi dada à informação. Ainda, que a liberdade de informação pode chocar-se com a liberdade de expressão, pois a primeira circula livremente em detrimento justamente dos direitos humanos de manterem os dados para si. O autor exemplifica ao lembrar que em 11 de janeiro de 2013 o *hacker* Aaron Swartz cometeu suicídio aos 26 anos, pois ficou aborrecido pelo fato da biblioteca digital JSTOR pretender cobrar de seus clientes, quando entendia que a informação deveria ser livre, gratuita e ilimitada. Ele usou a biblioteca do *Massachusetts Institute of Technology* para acessar a JSTOR e baixar milhares de trabalhos científicos para liberar gratuitamente pela internet e, na iminência de ser condenado, enforcou-se.

Osório (2022, p. 48) aponta que a livre circulação e contraposição de ideias, num primeiro momento, seria um caminho para a busca da verdade real, contudo, aponta alguns entraves no caminho, tais como a inexistência de uma verdade universal, as desigualdades econômicas com reflexo no acesso à informação de qualidade e a questão da era da *internet* e das mídias sociais, onde a produção de conteúdo é superabundante, as fontes de informações são difusas e os algoritmos direcionam o conteúdo, com a formação de bolhas.

De toda forma, mesmo estabelecendo as premissas acima, o tratamento dado ao combate à desinformação deve ser lastreado no direito humano à liberdade de expressão. Portanto, é necessário reconhecer a eficácia dos direitos fundamentais, sobretudo “da liberdade de expressão, nas relações entre plataformas digitais privadas e seus usuários, o que jogaria luzes peculiares sobre o problema da autorregulação praticada pelas empresas” (Sarlet; Hartmann, 2019, p. 98).

A *internet*, que acelerou o processo da sociedade em rede, que gestou a sociedade informacional, que trouxe tantas oportunidades e desenvolvimento, modificou a forma de vida em vários aspectos, seja no campo econômico, no social e no político, evidentemente trouxe também grandes desafios. A criação e ascensão das plataformas de mídias sociais é um deles, inegavelmente. Por todos os motivos já elencados, apesar da democratização da informação e da facilitação da comunicação entre os seres humanos, sabe-se que a internet se tornou um cenário fértil para a proliferação de informação indesejada, onde se encaixa a desinformação. E os direitos humanos, universais e que oferecem um padrão de conduta para as empresas de mídias que operam em centenas de países, são apontados pela maioria dos autores como o principal caminho a seguir.

A forma como as mídias aplicam os direitos humanos é através da moderação de conteúdo, um dos meios de combate à desinformação. Compreende-se a moderação como a tarefa assumida pelas mídias sociais com o intuito de verificar o conteúdo veiculado nas plataformas, tendo como objetivo filtrar a postagem indesejada, que é realizada através de ação humana ou da utilização de algoritmos (Rubin, 2022, p. 48). Apesar das mídias não serem responsáveis pela divulgação do conteúdo, elas tomam decisões importantes a respeito do que pode ou não ser veiculado (*content gatekeeping*), e quais informações serão classificadas *(content organizing)*. Na moderação de conteúdo, as empresas de mídia social contam com uma arquitetura que usa os padrões das comunidades, algoritmos e revisão humana, essa última em escala menor. Ao aderir à plataforma, os usuários recebem das empresas de mídias sociais as políticas de uso e os padrões das comunidades, ressaltando que se o conteúdo confrontar com as normas internas, serão excluídos, removidos ou sinalizados (Sander, 2021, p. 163).

Os padrões das comunidades são inspirados nos *standards* de direitos humanos, fornecendo proteção especial para a opinião e a expressão em assuntos de interesse público. A escolha dos direitos humanos, como dito, possui como principal motivo a orientação no direito internacional, como forma de garantir a padronização em vários países (Jørgensen, 2021). Para exemplificar, o *Meta*, que é detentor do *Facebook*, *Instagram* e *WhatsApp*, revela no seu sitio eletrônico que além das suas regras internas, possui a sua política de Direitos Humanos (Política Corporativa de Direitos Humanos) (Meta, 2021). Do mesmo modo, o *Youtube*, de

propriedade da empresa *Google*, apresenta os padrões de Direitos Humanos da *Google* (Google, 2020). O *Twitter,* atual X*,* também informa a *sua* Política Pública para Defesa e Respeito aos direitos das pessoas que utilizam seus serviços, ressaltando o compromisso com a liberdade de expressão e a privacidade (X, 2023).

Verifica-se de plano que esses entes privados atuam sobre o exercício de direitos fundamentais do cidadão, especialmente a liberdade de expressão, como se fossem entes estatais, sendo que as plataformas acabam por definir as regras sobre o conteúdo postado (Sarlet; Hartmann, 2019, p. 96). Na concepção de Jones (2019, p. 30), o direito internacional dos direitos humanos serve como marco legal para orientar uma atividade empresarial em escala global, com impactos na vida das pessoas, com poderes semelhantes aos de um governo local. Bom observar ainda que a prevalência dos princípios de direitos humanos garante a manutenção dos pontos de vista diferentes, porém, prevendo restrições ao conteúdo violento, odioso e que desinforma (Jørgensen, 2021).

As mídias sociais se tornaram palco para a expressão da opinião e da veiculação da informação e não a toa, pode-se afirmar que permitiram a cada indivíduo a facilidade de participar do debate público e político. Osório (2022, p. 97) assinala que o espaço cibernético “tornou-se o principal fórum de exercício da liberdade de expressão, a nova praça pública da democracia”. E assim, a necessidade de proteção a esse princípio necessita que seja aplicado o princípio da proporcionalidade e, quando outros direitos fundamentais se chocarem com o direito humano de expressão e opinião, há que se fazer a ponderação sempre com o norte voltado para a sua preservação máxima.

Já foi dito que os direitos humanos devem nortear a atuação das mídias sociais e especificamente no que se refere à moderação de conteúdo e a regulação, é inevitável que haverá uma linha sensível entre o bem jurídico que se pretende tutelar (como o combate à desinformação, ao discurso de ódio e a proteção da democracia), há um específico que já nasce com a criação a desenvolvimento dessa realidade e que foi conquistado a duras penas, que é justamente a liberdade de expressão e de opinião. No entanto, apesar do parâmetro definido, as mídias são cobradas pela falta de transparência, o que vem prejudicando a proteção dos direitos humanos, considerando que o desenvolvimento das plataformas tem sido "uma caixa preta",

porque se desconhece os processos de tomada de decisão (Gonzáles-Iza, 2021, p. 2).

Também já foi dito que um dos problemas é que além da utilização de algoritmos nos seus modelos de negócios, os mesmos são utilizados para a moderação de conteúdo através da utilização da inteligência artificial. A situação apresenta um grande paradoxo, pois ao analisar os números das principais plataformas de mídias sociais constata-se que são bilhares de usuários conectados, algumas delas do tamanho de países de proporções continentais. O volume de conteúdos postados é grande e ao mesmo tempo é impensável que tais plataformas consigam empregar seres humanos em número suficiente para a missão de moderar o conteúdo. Assim, estamos diante de um dilema em que a revisão dos conteúdos é feita em grande escala por um formato não humano. Nesse contexto, exige-se que as regras sobre moderação de conteúdo devem ser claras, objetivas e transparentes, de modo que o usuário possa prever a atitude a tomar (Jørgensen, 2021).

Assim, analisar as principais iniciativas globais a respeito do combate à desinformação a serem aplicadas pelas mídias sociais é extremamente importante para aprimorar a moderação de conteúdo, considerando o direito humano à liberdade de expressão.

Diante do cenário descrito, o foco do próximo capítulo será abordar a evolução das mídias sociais e aprofundar o impacto que a desinformação tem causado para a sociedade, bem como as estratégias de governança e de combate à desinformação, além de abordar a moderação de conteúdo pelas mídias sociais.

3 MÍDIAS SOCIAIS, DESINFORMAÇÃO E O SEU TRATAMENTO JURÍDICO

No presente capítulo buscar-se-á trazer uma dimensão do impacto oriundo da burocratização da *internet*, do avanço das telecomunicações e da integração entre os computadores, fatos que modificaram de forma destacada a vida em sociedade, numa revolução na forma de comunicação entre as pessoas ao redor do mundo.

Diante desse cenário, considerando o fluxo de informação e todas as peculiaridades dessa nova forma de comunicação em que os usuários interagem através de compartilhamento de fotos, vídeos, participação em comunidades e *microbloggings*, a desinformação surge como um grande problema a ser enfrentado pela humanidade, dado a sua potencial influência. Assim, de forma paralela à emergência das mídias sociais, serão apontados os impactos da desinformação diante dessa nova realidade.

Ainda, será abordada a questão da governança e das estratégias de combate à desinformação, oportunidade em que se fará um itinerário histórico do debate, a partir do surgimento e evolução das grandes plataformas de mídias sociais. No mesmo tópico serão trazidas algumas experiências governamentais no tratamento da regulação da moderação de conteúdo.

Por fim, ainda nesse capítulo, será apresentado um panorama da moderação de conteúdo pelas mídias sociais, ancoradas nos padrões de cada comunidade e nas *guidelines*, além de se trazer as contribuições das cartas de princípios (iniciativas da sociedade civil organizada, por vezes com participação governamental).

3.1 EMERGÊNCIA DAS MÍDIAS SOCIAIS E OS IMPACTOS DA DESINFORMAÇÃO

Como reportado no primeiro capítulo, a partir da década de 1990, o avanço das telecomunicações e da tecnologia de integração de computadores fez desenvolver as redes a partir da utilização da fibra ótica, do laser e das redes de banda larga integradas (Castells, 2002, p. 81). Nesse contexto, não demorou muito para surgirem as atuais e grandes plataformas de mídia social nos anos 2000, trazendo diversas modificações no campo da difusão da informação e da comunicação de massa.

A *internet* propiciou a sociabilização através de ferramentas de comunicação, onde os atores podem interagir e se comunicar com outros, deixando na rede os chamados rastros. A tecnologia digital, impulsionada pela *internet* propiciou novas práticas e modelos de negócios, substituindo o conceito de usuário pelo de navegador (Passarelli; Gomes, 2020, p. 254). É a comunicação mediada pelo computador, formando as redes. A rede, de sua vez, é utilizada metaforicamente para a observação das conexões estabelecidas entre esses atores. Portanto, atores e conexões são os elementos de uma rede social na *internet* (Recuero, 2009, p. 24).

Nos primórdios da difusão da *internet* a rede foi introduzida como fonte de consulta e pesquisa para a obtenção de dados, dando origem à *web* 1.0, sendo o usuário um mero expectador, com pouca interação, contudo, foi sentida a ampliação da fonte de conhecimento e informação. A partir daí, surge uma nova internet, a *web* 2.0, intitulada *web* participativa, trazendo a ideia de plataforma, com a difusão dos *blogs, chats,* mídias e redes sociais, com a ideia de que os usuários passam a ser criadores de conteúdo (Passarelli; Gomes, 2020, p. 257-259).

A *web* 3.0, chamada de *web* semântica, que trouxe a melhor utilização dos dados a partir da inteligência artificial (IA) e do aprendizado da máquina (*learning machine*), compreendendo melhor as necessidades individuais e a personalização do conteúdo (Passarelli; Gomes, 2020, p. 265), mas foi já no contexto da *web* 2.0 que surgiu a ideia de mídia social. Para Telles (2010) há distinção entre os conceitos de redes sociais e mídias sociais. As redes sociais são espécies do gênero mídias sociais, havendo redes com características de *microblogging*, de compartilhamento de vídeos e de fotos. As mídias sociais, portanto, incluem a infinidades de redes e sites de relacionamentos, ou seja, os meios ou canais onde a informação vem a ser transmitida. O objetivo das mídias sociais é oferecer serviços de *internet,* com o intuito de propiciar aos usuários a comunicação online, o compartilhamento de conteúdo, oferecendo diversos meios de conexão, dentre os quais as redes sociais (Chaouch, 2022, p. 22).

A opção pelo termo mídia social provém da palavra inglesa *media*, que significa meios. Assim, as mídias sociais são os meios que reúnem os sítios da *web* e *softwares* da *web* e que permitem a troca de informações, não apenas opiniões e comentários, gerando o desenvolvimento das plataformas de redes sociais onde interagem os usuários. Em sentido amplo, o termo rede social tem como característica

toda correlação e interação pessoal em uma comunidade, seja *online* ou *offline*. Desta forma é necessário especificar o termo rede social *online*, como sendo os aplicativos ou *sites* que permite aos usuários trocar informações e opiniões na *Internet* pelos mais diversos tipos de dados, sejam mensagens, vídeos, arquivos e comentários (Pierre, 2018).

Pode-se afirmar, a título ilustrativo, que apesar de ter sido superada rapidamente pelos grandes sites de mídias sociais, o *Orkut* foi o primeiro a obter popularidade em 2004 com a aquisição e lançamento pelo Google. Foi criado por Orkut Buyukkokten em 2001. Atualmente, pode-se destacar o *Facebook*, sistema criado pelo norte americano Mark Zuckerberg, também lançado em 2004, funcionando através de perfis e comunidades. O *Twitter*, por sua vez, é um site de *microblogging* fundado por Jack Dorsey, Biz Stone e Evan Williams em 2006, num projeto da empresa Odeo (Recuero, 2009, p. 165-173).

O *LinkedIn* foi criado na Califórnia (EUA) em 2002 por Reid Hoffman, Konstantin Guericke, Jean Luc Vaillant, Allen Blue e Eric Ly com a finalidade de recrutamento pessoal para as empresas, sendo que muitos o utilizam como uma rede profissional. O *YouTube* atua com o compartilhamento de vídeos na *internet* e também surgiu na Califórnia (EUA), em 2005, tendo como criadores Steve Chen, Chade Hurley e Jawed Karim. O *Instagram* permite que os usuários possam compartilhar fotos e vídeos através de assistentes digitais pessoais e foi fundado em 2010 por Kevin Systrom e Mike Krieger (Madakam; Tripathi, 2021, p. 9).

O vertiginoso crescimento desses sites comunitários trouxe um novo paradigma de comunicação diante da grande integração de usuários, num grande impacto social e econômico. A sociedade contemporânea foi transformada, fomentando de forma acentuada a difusão da informação e da comunicação, permitindo o exercício pleno da liberdade de expressão. As mídias sociais são parte importante e avançada da vida das pessoas, sendo que nas últimas duas décadas testemunhou-se a sua grande popularidade (Madakam; Tripathi, 2021, p. 7).

Ainda, o surgimento da Internet representou a possibilidade de realização plena do direito de expressão livre, instantânea, a baixo custo, com impacto considerável no jornalismo e na forma de circulação e acesso à informação. Propiciou, a internet, que todos possam ser jornalistas, formadores de opinião e editores de conteúdo (Bento, 2016, p. 102). E a evolução, a partir do surgimento da *web* 2.0 foi

muito veloz, fazendo com que as mídias sociais fossem ganhando um protagonismo como fonte de informação e divulgação de conteúdo.

Para que se tenha uma dimensão do tamanho de algumas plataformas, em 2016 Schwab (2016, p. 122) já afirmava que se os três *sites* mais populares de mídia social fossem países, eles teriam 1 bilhão de pessoas a mais do que a China. E os dados mostram que em quase todos os países o uso das mídias sociais continua aumentando, sendo que no início de 2023 o *Facebook* possuía 2,989 bilhões de usuários. O *Youtube* possuía 2,527 bilhões de usuário, o *Instagram* 1,628 bilhões, o *TikTok* 1 bilhão, o *LinkedIn* 922,3 milhões e o *Twitter* 372,9 milhões, de acordo com o DataReportal (2023). Os dados impressionam, sendo que o número de usuários que formam as comunidades das principais plataformas de mídias sociais de fato supera a população de vários países.

Trazendo esses dados para o Brasil, estima-se que no início de 2023 haviam 181,8 milhões de usuários de *internet*, numa penetração de 84,3%, sendo que, em relação às mídias sociais, havia 152,4 milhões de usuários de mídia social, equivalente a 70,6% da população total. Quanto às plataformas, em números de usuários no início de 2023, o *Youtube* possuía 142,0 milhões, o *Instagram* 113,5 milhões, o *Facebook* 109,1 milhões, o *TikTok* 82,21 milhões, o *LinkedIn* 59,00 milhões e o *Twitter* 24,30 milhões (DataReportal, 2023).

Evidentemente, a ascensão das mídias sociais impactou de forma considerável o tecido social, com consequências para transformações culturais e sociais profundas. Com a invenção do alfabeto em 700 a.C. na Grécia, a cultura alfabetizada separou a comunicação escrita do sistema audiovisual de símbolos e percepções, relegando o mundo dos sons e imagens às artes. No século XX com o filme, o rádio e a televisão, a comunicação audiovisual superou a influência da comunicação escrita. 2.700 anos após a invenção do alfabeto, a transformação tecnológica integra as modalidades de comunicação oral, escrita e audiovisual na comunicação humana através de uma rede interativa (Castells, 2002, p. 413). Uma inundação de textos, imagens são gerados e distribuídos instantaneamente, gerando o que Passarelli e Gomes (2020, p. 255) intitulam de supremacia da imagem sobre as palavras, criando uma revolução na forma de comunicação e ao mesmo tempo, trazendo preocupações com a proteção de dados e com a abundância de informação de má-qualidade.

A ascensão das mídias sociais, o declínio das mídias tradicionais e o modelo de negócios das plataformas que lucra com conteúdo sensacionalista e emocional, trouxe consigo grandes desafios relacionados à proliferação desenfreada de conteúdos nocivos, dentre eles a desinformação, sobretudo aquela divulgada com o objetivo de causar dano, de influenciar a opinião pública ou para obter ganhos financeiros. O impacto das informações falsas, aliás, é um dos perigos modernos para a sociedade, conforme assinalado pelo Fórum Econômico Mundial (Kumar; Shah, 2018, p. 1). O modelo de negócios é pernicioso, com a utilização dos dados pessoais, considerado o petróleo da economia do século XXI, modo pelo qual se financia o uso digital gratuito das plataformas (Jones, 2019, p. 8).

Deve ser considerado que o avanço da tecnologia tornou o processo de criação e disseminação complexo, com o uso de atores agindo de forma independente ou em grande escala, através da utilização de redes de *bots*, o que é facilitado pelo fato dos seres humanos possuírem suscetibilidade para receber e espalhar notícia falsa. Cria-se assim a ilusão do consenso em relação à informação, pois o leitor ao observar comentários semelhantes, refletindo o mesmo ponto de vista, convencendo-se ao não perceber que a discussão tem origem em uma única fonte. A larga escala, com a utilização de comportamento não autêntico, gera a mesma informação para um grande público, inflando a percepção de que uma informação falsa seja legítima (Kumar; Shah, 2018, p. 6). A reiteração de uma mensagem, ainda que chocante e errada, a torna aceitável, pois no dizer de Jones (2019, p. 9), a repetição normaliza.

O conteúdo da desinformação, que como dito é criado e disseminado de forma complexa, utiliza-se de palavras, fotos e vídeos, influenciando as pessoas de forma a provocar a divisão. Inclui a criação dos chamados memes (imagens com poucas palavras capazes de impactar a consciência humana), o discurso de ódio, a ‘trollagem’ (postagens com o objetivo de irritar, dividir ou assediar), os *deep fakes* (áudio e vídeo cujo processo tecnológico de fabricação dificulta a detecção da fraude), a utilização de falsos sites de notícias e de identidades falsas, enfim, transformando a campanha política numa guerra, suprimindo o debate racional (Jones, 2019, p. 11).

Para que se tenha ideia do impacto que a desinformação pode causar para a democracia, estudos em campanhas eleitorais constataram que as contas de *bots* são formadas por aproximadamente um quinto das conversas sobre política do *Twitter*, sendo que esse meio é provavelmente o maior veiculador de notícias falsas

(Kumar; Shah, 2018, p. 6). A desinformação no discurso público não traz apenas problemas no debate eleitoral, mas influencia a opinião pública de forma mais ampla, com a polarização de pontos de vistas sobre uma série que questões, tais como a vacinação de crianças, questões de gênero e de transgênero, alimentando, por conseguinte a preferência eleitoral e definindo o debate político (Jones, 2019, p. 8).

Ainda de acordo com Kumar e Shah (2018, p. 7-9) várias pesquisas buscaram medir a capacidade dos seres humanos na detecção das informações falsas, chegando a uma conclusão negativa, sobretudo quando a desinformação é criada de maneira inteligente para dar uma conotação de credibilidade e legitimidade. A questão é potencializada com a utilização de algoritmos que recomendam conteúdos personalizados e que criam o efeito de uma câmara de eco que se caracteriza pela polarização do conteúdo voltado às suas crenças.

A complexidade é tamanha, que levou Jones (2019, p. 9) a afirmar que “a veracidade de informação é apenas a ponta do desafio”. De acordo com a autora, experimentos demonstraram que as emoções podem ser manipuladas através da emoção dos outros expostas nas mídias, desempenhando a comunicação uma função ritualística impulsionando as conexões e reforçando o sentimento do “nós contra eles”. É a sensação fascinante de ter a nossa visão de mundo reforçada por notícias que a confirmam, pouco importa se verdadeiras ou falsas, encorajando o populismo.

A desinformação possui impacto nos mais diversos campos, como a política, o mercado de ações, resposta a desastres naturais e ações terroristas, vez que que informações falsas se espalham amplamente nas mídias sociais com um atraso médio de doze horas entre o seu início e a sua confrontação. Ainda, num estudo comparativo entre histórias falsas e reais relacionadas às eleições norte americanas de 2016, constatou-se que as notícias falsas tiveram um total de 8.711.000 engajamentos acima dos 7.367.000 engajamentos das notícias reais (Kumar; Shah, 2018, p. 10). Bom sempre ressaltar que o conteúdo que mais possui sucesso é o que desperta a indignação e a emoção, despertando sentimentos como raiva, medo e desconfiança, de forma que a captura da emoção é a chave para a influência do comportamento (Jones, 2019, p. 9).

Outro estudo dentre outras questões, mostrou que dentre artigos falsos publicados no Wikipedia, embora 90% dos artigos sejam identificados imediatamente em uma hora após serem aprovados, cerca de 1% dos boatos falsos, bem escritos,

sobrevivem por mais de um ano sem serem detectados. Ainda recentemente, foi realizado o maior estudo sobre a disseminação de mais de 126.000 rumores no Twitter durante um período de 11 anos e, em comparação com informações verdadeiras, chegou-se à conclusão que os tweets contendo informações falsas se espalham significativamente mais longe, mais rápido e de forma mais ampla (Kumar; Shah, 2018, p. 11).

Nesse sentido, os atores políticos se utilizam dos dados pessoais, o petróleo do século XXI, para direcionar a mensagem para potenciais eleitores específicos, gerando o máximo de impacto na informação e desinformação, numa escala assustadoramente maior do que nos tempos em que o material era apenas impresso, o que por óbvio potencializa, distorce e forma a opinião pública (Jones, 2019, p. 16). O fato é que a sociedade acompanha toda essa evolução tecnológica, com as implicações relacionadas não só ao grande fluxo de desinformação, mas de uma forma geral, com as preocupações com as questões éticas relacionadas a esse novo paradigma trazido pelas novas tecnologias (Werthein, 2000, p. 75). Conforme explica Tandoc Junior (2021, p. 44), o aumento da desinformação exige, para seu combate, soluções ponderadas e sustentáveis, através de uma abordagem multidisciplinar, envolvendo intervenções tecnológicas, econômicas, jurídicas e sociais.

3.2 GOVERNANÇA, ESTRATÉGIAS DE COMBATE À DESINFORMAÇÃO E AS INICIATIVAS GOVERNAMENTAIS

No primeiro capítulo foi abordado o fenômeno da desinformação, suas facetas e consequências para a democracia. Essa nova realidade implementada pelas redes interativas e pelo fluxo de informação traz como consequências o impacto dos conteúdos indesejados, sobretudo quando coordenados com a intenção de causar danos. Especificamente no que se refere à desinformação, é favorecida por um ambiente em que o processamento emocional de notícias aumenta significativamente (Starts, 2021, p. 24). Ainda, a desinformação é um dos possíveis conteúdos indesejados, podendo ser destacado o discurso de ódio, o terrorismo, incitação à violência, exploração sexual, dentre outros que levaram ao estudo da regulação das

plataformas, tendo em vista o alcance social da atuação das mídias, apesar de serem empresas privadas.

Deve ser considerando que a propaganda tem sido usada há muito tempo como uma ferramenta de política externa e de busca de poder, sendo que o desafio de combater as notícias falsas não é novo. No período da Revolução Francesa, as notícias falsas eram tidas como subversivas, a exemplo da condenação de franceses em 1792, assim como o combate aos ataques subversivos de propaganda da Rússia Soviética a países estrangeiros desde 1920. Depois de 1945, as preocupações relativas à comunicação de massa como um canal potencial de propaganda, especialmente através do rádio, foram vistas como um prenúncio de oportunidades para trazer informação e propaganda para as casas das pessoas a nível internacional de uma forma sem precedentes. A propaganda se distinguia da educação e da informação não pelo seu conteúdo, como a desinformação, mas enquanto fossem deliberadamente compartilhadas para causar danos. Como resultado destas preocupações do pós-guerra, durante a década de 1940 houve extensos debates sobre a futura regulamentação da propaganda. A Assembleia Geral da ONU de 1947 adotou duas resoluções sobre este assunto, uma condenando a propaganda de guerra e outra, a respeito de informações falsas e distorcidas (Jones, 2019, p. 61).

De um modo geral, o controle da internet vem sendo debatido desde a privatização total da *internet* em abril de 1995. Vale recapitular, incialmente a *Arpanet* surgiu como uma estratégia militar americana, passando em 1990 a se chamar *Internet* e ser operada pela *Nacional Science Foundation*, até ser totalmente privatizada. Ocorre que a privatização não veio acompanhada da previsão de uma autoridade de supervisão, surgindo diversos mecanismos improvisados e, mesmo com a criação de um órgão regulador em 1998 nos EUA (IANA/ICANN), em 1999 não havia clareza quanto à autoridade naquele país e no mundo (Castells, 2002, p. 83).

Feito o registro, o surgimento e ascensão das mídias sociais, que nada mais são do que produtos de empresas privadas, porém com gigantesco impacto social, trouxe à tona a necessidade de regulação. É importante compreender a gênese da discussão acerca da responsabilidade das redes interativas sobre o conteúdo gerado por seus usuários.

Castells (2002, p. 442) faz menção às comunidades virtuais, como sendo redes de comunicação interativas, sendo que a partir de 1990 foram criadas dezenas

de milhares em todo o mundo, com a maioria delas com base nos EUA. As maiores mídias sociais que conhecemos hoje, como por exemplos o *Twitter* e o *Facebook*, foram criadas nos anos 2000 nos EUA, porém, o debate sobre a reponsabilidade pelo conteúdo postado pelos usuários tem sua gênese em dois casos emblemáticos julgados anteriormente, nos anos 90 (Estarque; Archegas, 2021, p. 19). Por motivos óbvios, o desenvolvimento e massificação do uso das mídias sociais, bem como o desenvolvimento do modelo de negócios tiveram como influência esses casos paradigmáticos.

O primeiro caso judicial ocorreu no ano de 1991, quando o Tribunal Distrital dos EUA para o Distrito Sul de Nova York decidiu o caso *Cubby v CompuServe.* Nesse caso, a plataforma *CompuServe* foi acionada pela empresa Cubby visto que a parte Ré desenvolvia e fornecia produtos e serviços relacionados à computadores, incluindo o *CompuServe Information Service* ("CIS"), espécie de "biblioteca eletrônica" em que os assinantes podem acessar de um computador pessoal ou de um terminal. O fundamento da ação foi pautado pelo ataque direto a um produto da *Cubby,* chamado *Skuttlebut* (um banco de dados para publicar e distribuir eletronicamente notícias e fofocas nos noticiários de televisão e indústrias de rádio), por uma concorrente assinante da *CompuServe* através de declarações falsas e difamatórias. Os ataques ocorreram através de um fórum da *CompuServe,* portanto, a empresa era mera intermediária do conteúdo (Nova Iorque, 1991).

A *CompuServe* não contestou o fato de que as declarações de seus assinantes (concorrentes da *Cubby)* eram difamatórias, apenas argumentou que na condição de mero distribuidor, sem responsabilidade pela edição, não poderiam ser responsabilizados pelas postagens (Nova Iorque, 1991). O Tribunal Distrital entendeu que a empresa *CompuServe* não possuía relação direta com o conteúdo, não promovendo o controle prévio sobre a veiculação, sendo impossível ter conhecimento do material antes de publicado (Estarque; Archegas, 2021, p. 17). Para o Tribunal, foi incontroverso que o usuário carrega o texto nos bancos de dados da *CompuServe*, disponibilizando-o de forma instantânea para os demais assinantes, ou seja, a empresa não tem “mais controle editorial sobre tal publicação do que uma biblioteca pública, livraria ou banca de jornal” (Nova Iorque, 1991, tradução própria).

Noutro importante caso, o *Stratton Oakmont v Prodigy Services*, de 1995, a Suprema Corte de Nova York entendeu que a empresa que optar por moderar o

conteúdo, possui responsabilidade. A empresa *Stratton Oakmmont, Inc.*, do ramo de investimentos em valores mobiliários teria praticado atos fraudulentos na oferta pública de ações de uma determinada empresa. A *Prodigy Services* seria uma desenvolvedora da plataforma digital em que foram divulgados os atos fraudulentos. O julgado, inclusive, reporta ao caso *Cubby x CompuServe,* fazendo uma distinção com o caso *Stratton Oakmont v Prodigy Services*, pois a *Prodigy* teria se apresentando ao público controlando o conteúdo dos seus boletins, utilizando um programa de triagem automática e de diretrizes que os líderes do Conselho deveriam aplicar. Assim, *a Prodigy* estaria tomando decisões quanto ao conteúdo, caracterizando o controle editoral (Nova Iorque, 1995).

Pelos dois casos emblemáticos, e seguindo a tradição da jurisprudência do *common law* do direito norte americano, no que se refere às plataformas de mídias sociais, foi estabelecida essa distinção entre um editor (que opta por moderar conteúdo) e um mero distribuidor. No primeiro caso, como visto, a plataforma passa a ser responsabilizada pelo conteúdo. Desta forma, como controlar o conteúdo poderia gerar a responsabilidade civil, as mídias sociais foram naturalmente desenvolvendo seu modelo de negócios como simples distribuidoras, prática que trouxe uma série de consequências para a coletividade (Estarque; Archegas, 2021, p. 18).

Para harmonizar os precedentes balizadores foi editada a Seção 230, da *Communications Decency Act*, de 1996 que traz na redação do artigo que “nenhum provedor ou usuário de serviço interativo de computador será tratado como editor de qualquer informação fornecida por terceiros” (Estarque; Archegas, 2021, p. 18). Partiu-se do pressuposto de que as plataformas não devem ser caracterizadas como editoras, além de conter a cláusula chamada de “cláusula do bom samaritano”, que trouxe a extensão da imunidade às mídias que optem por moderar conteúdo. Ainda, a imunidade se estende às decisões das plataformas acerca da retirada do conteúdo, inclusive de conteúdo legal, estando sujeita a algumas exceções em relação à responsabilidade criminal federal, bem como de reivindicações de propriedade intelectual. Foi ainda alterada em 2018 para limitar imunidades em casos de tráfico sexual (Jones, 2019, p. 6).

Importante considerar ainda, nesse contexto, que a liberdade de expressão nos Estados Unidos possui um valor superlativo, inspirada na Primeira Emenda à Constituição. Em síntese, a Primeira Emenda estipula que o Congresso não poderá

limitar a liberdade de expressão, assim com a liberdade religiosa, liberdade de reunião e direito de petição aos órgãos públicos[1]. Porém, mesmo tendo as empresas sede nos Estados Unidos, por pressão dos usuários e por questão de sobrevivência no mercado, as plataformas optaram por moderar o conteúdo e garantir a segurança do usuário, ainda que de forma tímida. O mercado exigiu, em contraponto à liberdade ilimitada de expressão, que o espaço não fosse transformado numa "terra de ninguém", onde os usuários não se sentissem acolhidos e seguros. Contudo, considerando os conflitos que foram surgindo com a proliferação e democratização das redes sociais, tanto as entidades da sociedade civil e, sobretudo, as entidades governamentais, não se conformaram com a autorregulação, passando a exigir políticas ainda mais rígidas das plataformas. Surgiram diversas iniciativas, como as cartas de princípios.

Pela lógica do modelo de negócios, as mídias não são responsáveis diretamente pela geração das postagens, contudo, elas possuem um papel fundamental pela tomada de importantes decisões a respeito da permissão da veiculação, através da moderação de conteúdo, que pode ser entendida como os mecanismos implementados para determinar quais informações são permitidas e proibidas em suas plataformas (*content gatekeeping*), bem como de que forma com que serão classificadas *(content organizing)*. Como se pode imaginar pelos números expressivos de usuários, são bilhões de postagens por segundo, de forma que as empresas, naturalmente, não possuem número de colaboradores suficientes para o trabalho, fazendo com que na moderação de todo esse conteúdo necessitem de uma arquitetura que utiliza os padrões das comunidades, algoritmos e uma pequena revisão humana (Sander, 2021, p. 163).

Outra questão que dificulta a regulamentação está relacionada à soberania e à territorialidade. Na concepção de Lévy (1999, p. 206) o ciberespaço é desterritorializante enquanto o Estado moderno tem sua base, sobretudo, no território. Serviços, dados e informações transitam instantaneamente no planeta digital sem filtros, ao passo que as legislações nacionais só valem para as fronteiras dos Estados. Como vem ocorrendo em muitos casos, um servidor pode se instalar em qualquer país

[1] "O Congresso não fará nenhuma lei a respeito do estabelecimento de uma religião, ou proibindo seu livre exercício; ou cerceando a liberdade de expressão ou de imprensa; ou o direito do povo de se reunir pacificamente e de solicitar ao governo a reparação de queixas" (Estados Unidos, 1791).

e organizar as comunicações indesejadas e proibidas, tornando a legislação pátria inaplicável.

Há ainda outra questão relacionada às dificuldades da resposta estatal, na medida em que apesar dos esforços de governos e ONGs nos debates sobre a reestruturação da internet, é muito mais difícil modificar um sistema já acabado do que um que está sendo projetado, além é claro, que a política estatal, não acompanha a metamorfose frequente da *internet* (Harari, 2016, p. 381). Enfim, as discussões a respeito da governança das plataformas, entendida como "o conjunto de relações jurídicas, políticas e econômicas que estruturam interações entre usuários, empresas de tecnologia, governos e outras partes interessadas importantes no ecossistema da plataforma" (Gorwa, 2019, tradução própria), vêm sendo objeto de estudos ao redor do mundo.

As próprias plataformas estabelecem regras de interação e permanência em determinada comunidade, tomando suas decisões com base nelas. Contudo, há outros atores que podem influir nesse poder conferido às plataformas. As normas provenientes dos Estados podem regular o setor, mas também definir o que seja ilegal e que deve ser perseguido e responsabilizado. Ainda, destaca-se o trabalho da sociedade civil e da comunidade científica na formulação de recomendações acerca da moderação. Destaca-se de igual modo o surgimento de atores privados independentes de revisão, como o *Oversight Board*, da Meta (Silva; Gertrudes, 2023, p. 6).

Há uma tendência, como resposta às preocupações com a liberdade de expressão e defesa dos direitos digitais surgidas a partir de leis sobre o conteúdo digital editadas na Alemanha, Cingapura e Austrália, além de formulação de compromissos voluntários, uma série de princípios e reorganização da supervisão institucional. No entanto, múltiplos desafios, tais como o modelo de negócios, a liberdade de expressão, a falta de experiência política, preocupações com o sufocamento da inovação tecnológica, além da caixa preta existente em cada plataforma (Gorwa, 2019).

Gorwa (2019) ensina que existem três tipos de atores principais na regulação: a empresa, as Organizações Não Governamentais, sociedade civil, bem como o Estado, incluindo governos e grupos supranacionais, como a União Europeia e as Nações Unidas. O fato é que, além das empresas de mídias, os demais atores

são importantes para o combate ao conteúdo indesejado, vez que embora a autorregulação, embora contenha vantagens para resolver de forma mais ágil determinadas questões, as plataformas não vêm apresentando procedimentos transparentes para permitir uma fiscalização efetiva pela sociedade civil (Silva; Gertrudes, 2023, p. 10).

São três modelos de regulação, a primeira é a autorregulação pelas próprias empresas de mídias sociais, através da moderação de conteúdo pautada nos termos de uso, *guidelines* e padrões de cada comunidade. Há ainda a corregulação. No caso do Brasil, governo ausente à regulação específica, a atuação estatal é bastante influenciada pelo Marco Civil da Internet. Esse regime de moderação, com a participação do Estado é denominado corregulação, com eventual monitoramento da autorregulação, muito embora a tarefa de criar uma norma estatal específica não seja fácil, a exemplo da experiência alemã, que incentivou demasiadamente a moderação excessiva pelas plataformas, além de delegar ao setor privado a tarefa de definir o que seria ilegal. A tarefa governamental é interessante, porém, para a fixação de limites para atividade e para buscar garantias aos usuários. Por derradeiro, há ainda a possibilidade da heterorregulação, em que as normas são editadas pelo Estado, sem participação alguma das plataformas, o que esbarra na falta de *expertise* do setor público (Silva; Gertrudes, 2023, p. 12).

Enfoque interessante é a compreensão de que o Direito e o mercado não são as únicas formas de restrição da conduta humana. Há ainda as leis da natureza e, cotidianamente, as leis do mundo virtual, como por exemplo, se uma plataforma de mídia social, através de seus códigos, definir que ninguém irá publicar nada após determinado horário não haverá chance de descumprimento. Nesse contexto Sarlet e Hartmann (2019, p. 97), posicionam-se por um modelo intitulado autorregulação regulada, aceitando a autonomia na gestão das plataformas sem descartar um controle externo, pois a remoção de postagem pelas empresas sem muitos esclarecimentos oferece maior risco de silenciamento, do que o proposto pelas autoridades públicas.

Alguns Estados passaram a regular as mídias sociais, com destaque para a edição da Lei Alemã NezDG, em 2017, exigindo que as plataformas com mais de 2 milhões de usuários registrados na República Federal da Alemanha, bloqueiem ou excluam os conteúdos ilegais ou “manifestamente ilegais” sob pena de multa no

importe de U$ 50 milhões de euros (Alemanha, 2017). Além da obrigação de fornecimento de relatórios detalhados das reclamações, a lei alemã prevê que o provedor facilite os canais de reclamações e que remova ou bloqueie o acesso, em 24 horas, do conteúdo “manifestamente ilegal”, após o recebimento da reclamação. Ainda, através do mesmo procedimento, que seja removido ou bloqueado o conteúdo “ilegal” em até 7 dias após a notificação. Muito embora o bom propósito de combater o discurso de ódio e a desinformação, a NetzDG trouxe um conceito abstrato de ilegalidade, levantando questões de liberdade de expressão, uma vez que delega uma enorme quantidade de decisões a empresas privadas, com penalidades e prazos muito curtos. Vislumbra-se, portanto, que o excesso de regulamentação pode ser nocivo e até mesmo ineficaz (Jørgensen, 2021, p. 2).

A França, por sua vez apresentou em 2019 um relatório intitulado *Creating a French framework to make social media platforms more accountable: Acting in France with a European vision*. Conforme o relatório francês, a regulação deve ser baseada em uma política pública que garanta as liberdades individuais e a liberdade empresarial das plataformas. Ainda, sustenta a regulamentação com foco na responsabilização das redes sociais, implementado por uma autoridade administrativa independente, com base em três pressupostos: a) transparência da função de ordenação de conteúdo; b) transparência da função que implementa os Termos de Serviço e a moderação de conteúdo; c) Defesa da integridade dos usuários (França, 2019).

Pelo que se denota, o modelo francês prevê o diálogo entre os operadores, o governo, o parlamento e a sociedade civil, bem como uma cooperação europeia, reforçando a capacidade dos Estados de agir contra plataformas globais. Possui como foco, ainda, o dever de transparência das regras do jogo (França, 2019). Na mesma linha, em 2020, o Reino Unido publicou o White Paper “*Online Harms White Paper*: *Full government response to the consultation*”, que prevê o dever de transparências das redes, assim como um órgão independente e público, nos moldes do que constou do relatório francês (Reino Unido, 2020).

No Brasil, a discussão ainda é insipiente, ressaltando-se a existência do Comitê Gestor da Internet no Brasil – CGI.br, criado pela Portaria Interministerial nº 147/1995, posteriormente alterada pelo Decreto nº 4.829/2003. O Comitê é formado por órgãos estatais, representante privados, do terceiro setor e da sociedade civil

organizada, tendo como atribuições, dentre outras, estabelecer diretrizes estratégica relacionadas ao uso e desenvolvimento da Internet no Brasil. A Resolução nº 2009/003, fixou os princípios para a governança e o uso da *internet* no Brasil, destacando-se dentre outros aspectos a liberdade, privacidade e os direitos humanos (Comitê Gestor da Internet no Brasil, 2009).

Importante mencionar, ainda que não disponha especificamente sobre a atuação das mídias sociais, no âmbito nacional tem-se o Marco Civil da *Internet*, trazido pela Lei 12.965, de 23 de abril de 2014, que estabelece princípios, garantias, direitos e deveres para o uso da internet no Brasil. Referido diploma legal tem como fundamento o direito à liberdade de expressão, bem como os direitos humanos, a livre iniciativa, a livre concorrência, a finalidade social da rede, dentre outros não menos importantes (Brasil, 2014).

Guarda relevância para o tema da moderação de conteúdo, a Seção III do Marco Civil da *Internet*, que trata da responsabilidade pelos conteúdos gerados por terceiros, estabelecendo-se no seu artigo 18 que o provedor de conexão à internet não será responsabilizado civilmente. Muito embora o conceito de mídias sociais seja menos amplo do que o de provedor de conexão à internet, fica claro, ainda que por analogia, que o legislador pátrio fez a opção pela não responsabilização, a não ser que o provedor descumpra uma ordem judicial, tal como referido no artigo 19. Como se infere, o marco normativo brasileiro não foi concebido com o objetivo de relacionar a liberdade de expressão aos conteúdos restringidos pela moderação dos provedores de aplicação de internet, um termo mais amplo, que abrange tanto sites, blogs e as plataformas. O que é estimulado com o dispositivo é justamente o oposto, pois evita a ação contra conteúdos criados pelos usuários, como forma de evitar a responsabilização (Monteiro *et al.*, 2021, p. 8).

Merece atenção, igualmente, o artigo 21 do Marco Civil da *Internet*, que prevê a responsabilidade subsidiária do "provedor de aplicações de internet" pela violação da intimidade, sem autorização de seus participantes, por conteúdo contendo "cenas de nudez ou de atos sexuais de caráter privado", após a notificação do interessado. O mesmo artigo aponta uma outra figura, qual seja, "provedor de aplicações de internet" (Brasil, 2014), o que se aproxima do conceito de mídias sociais.

Porém, como se verifica, a preocupação originária do legislador nacional foi centrada em apenas um dos temas que merecer a atenção das redes, da sociedade

e do governo, qual seja, as cenas de nudez e atos sexuais privados. Com toda a discussão e a crescente pressão da sociedade civil e do governo, as plataformas optaram por definir medidas de autorregulação, a partir de uma perspectiva global e necessária para definir uma linearidade de atuação, considerando o fato de que operam em diversos países.

O fato é que o tema da desinformação ganhou relevo e atenção da sociedade civil, de estudiosos e das próprias mídias, de forma que ao longo dos últimos anos, foram propostas diversas medidas e estratégias para a regulação da moderação de conteúdo de uma forma geral e, mais especificamente, para o combate à desinformação, o que será tratado na seqüência.

3.3 MODERAÇÃO DE CONTEÚDO PELAS MÍDIAS SOCIAIS: PADRÕES DAS COMUNIDADES, GUIDELINES E AS CARTAS DE PRINCÍPIOS

Em vista da crescente adesão dos usuários nas plataformas, ganha corpo a atenção quanto à atuação das mídias sociais que, apesar do seu caráter privado, gera uma série de repercussões na esfera coletiva. Nesse contexto, a autorregulação das plataformas pressupõe a implantação de mecanismos de moderação de conteúdo, com base nos seus padrões das comunidades e suas orientações (*guidelines)*. A moderação de conteúdo pode ser definida como o conjunto de regras e procedimentos “usados pelas plataformas para remover, limitar alcance, rotular conteúdo como desinformação, assim como suspender ou remover contas” (Monteiro *et al.*, 2021, p. 7).

Moderação de conteúdo, além de serem regras e procedimentos para remover, limitar ou rotular o conteúdo, bem como remover ou suspender contas, pressupõe a violação dos termos de uso, como forma de impactar a sua disponibilidade, visibilidade e credibilidade (Oliva; Tavares; Valente, 2020, p. 11). Para Barbosa (2021, p. 761), a exclusão de conteúdos por parte dos administradores da rede é permitida, mas apenas quando a publicação viole direitos alheios ou se constitua numa prática abusiva de direito, em situações como publicação de fatos manifestamente falsos, de forma consciente e com o fim de obter vantagem.

Uma questão importante diz respeito aos critérios para definição dessas regras e padrões que nortearão as plataformas e seus usuários, pois deve-se

trabalhar com o fato de que as maiores plataformas de mídias sociais operam em centenas de países, trazendo a necessidade de definição das regras de convivência se dê a partir das políticas da comunidade de cada plataforma, inspirada num direito universal, reconhecido pela maioria dos países. Natural, portanto, que os *standards* de direitos humanos fossem efeitos para inspirar a autorregulação. A escolha possui vários motivos, dentre eles a orientação no direito internacional, garantindo um padrão em vários países. Ao invés de discutir num primeiro plano a responsabilidade pelo conteúdo, considera-se inicialmente a proteção dos direitos e liberdades individuais. A partir desses pontos de convergências, as normas de freio aos Estados e empresas possui condições de seguir um mesmo padrão, uma linearidade (Jørgensen, 2021, p. 2)

Assim, para estudar o funcionamento e as implicações das mídias sociais no Direito e na sociedade, exige-se a análise de questões como a responsabilidade das plataformas pelo conteúdo, a moderação, os padrões da comunidade e as políticas corporativas de direitos humanos. O ponto central é que as regras devem ser bastante claras e específica para prever a conduta dos usuários, além de garantir a transparência quanto ao conteúdo que será descartado, com as restrições delimitadas nos princípios de direitos humanos. A ideia de visão social dos princípios de direitos humanos, garante a inclusão e os pontos de vista diferentes, prevê restrições ao conteúdo à violência, ódio e desinformação, garantindo como consequência uma proteção individual (Jørgensen, 2021, p. 3).

Retomando ao tema em escala universal, deve ser considerado que o modelo de negócio das mídias sociais é bem diferente daquele utilizado pela indústria tradicional. No caso, os usuários das plataformas de mídias sociais não compram o produto, eles pagam pelo acesso com seus dados, que são transformados em receita. Os dados, aliás, constituem-se em verdadeira fonte de receita para as plataformas, pois a partir deles conseguem direcionar a publicidade. Assim sendo, a necessidade de proteção ao usuário deve ser o ponto central da moderação de conteúdo, com o objetivo de evitar danos a eles e à sociedade (Buhmann; Olivera, 2020, p. 138).

O esforço centrado na autorregulação fez o grupo *Meta*, detentor do *Facebook*, *Instagram* e *WhatsApp*, além das suas políticas de privacidade e termos de uso, editar a sua política de Direitos Humanos, disponibilizando em seu *site* oficial, com objetivo de informar seus usuários, a sua Política Corporativa de Direitos

Humanos (Meta, 2021). De igual forma, o *Youtube*, que pertence à empresa *Google*, além de suas políticas internas, segue os padrões de Direitos Humanos da *Google* reconhecidos internacionalmente, disponíveis na sua página oficial (Google, 2020).

O *Twitter*, atual X, também possui e disponibiliza em seu *site* sua Política Pública para Defesa e Respeito aos direitos das pessoas que utilizam seus serviços. Na política, a rede social enfatiza seu compromisso com à liberdade de expressão e a privacidade. A Política de Direitos Humanos do *Twitter* tem como base a Declaração de Direitos dos Estados Unidos e a Convenção Europeia de Direitos Humanos, que objetiva a Proteção dos Direitos Humanos de Liberdade e, também, os Princípios Orientadores sobre Empresas e Direitos Humanos das Nações Unidas (X, 2023).

Para exemplificar a complexidade e o resultado dessa abordagem calçada nos direitos humanos, o *Facebook*, maior mídia social do mundo, possui seus *standards* da comunidade inspirados nos preceitos universais, com a ressalva de que diante de milhares de postagens por segundo, é inegável que só existe uma possibilidade para o primeiro filtro: a inteligência artificial. Sander (2021, p. 163) assinala que para cumprir a tarefa de moderar o conteúdo, as mídias sociais contam com a inteligência artificial a partir dos algoritmos e da revisão humana, a depender do tamanho, dos recursos e da cultura da plataforma.

Os padrões da comunidade do *Facebook*, ou *standards,* pretendem garantir a autenticidade, segurança, privacidade e dignidade. No item integridade/autenticidade, se encontra o tópico desinformação. No tópico desinformação, no qual vem mencionada a dificuldade de definir a notícia falsa, o esforço da rede para reduzir a disseminação de boatos e a desinformação viral, além de direcionar usuários para informações oficiais, informar a proibição de contas falsas, a fraude e o comportamento inautêntico coordenado (Meta, 2021).

No mesmo campo, o *Facebook* informa as desinformações que serão removidas: a) desinformação com risco de agressão física ou violência; b) desinformação prejudicial sobre a saúde; c) desinformação que pode contribuir ou contribua diretamente para o risco de interferência na capacidade das pessoas participarem do pleito eleitoral; d) mídia manipulada, que pode se tornar viral rapidamente (Meta, 2021).

O *Facebook* reconhece, portanto, a dificuldade de definir o que seria desinformação, sobretudo para que não se invada o direito à liberdade de expressão.

Contudo, elegeu meios para o combate, através da disseminação da informação oficial e confiável e do combate à fraude e contas falas. Ainda, de acordo com seus princípios, informa que removerá desinformação na área da saúde, pleito eleitoral, que traz risco à agressão ou violência, assim como a que viraliza rapidamente. Muito embora as iniciativas das plataformas, estudos indicam que há muito que avançar, pois atuando em contextos culturais diferentes, em escala global, somado à preocupação com a confiança na inteligência artificial utilizada na moderação, a elaboração de políticas e diretrizes tem falhado, tanto no cerceamento à liberdade de expressão, como em outras vezes permitindo o conteúdo ofensivo e violento (Monteiro *et al.*, 2021, p. 13).

Os esforços das redes, portanto, devem ainda considerar na moderação de conteúdo que uma concepção estrutural dos direitos humanos a partir de uma visão holística, assim como uma abordagem cultural de cada localidade daria mais ênfase na obrigação de proteger a liberdade de expressão, como base para exigir que os Estados assegurem mecanismos efetivos de transparência, processo legal, responsabilidade e supervisão da plataforma, numa relação público-privada (Sander, 2021, p. 192). Assim, além do combate à desinformação, deve ser garantida a transparência das redes, a facilitação dos meios de revisão das decisões, regras claras quanto ao que é considerado desinformação, quanto ao uso de inteligência artificial, assim como quanto ao respeito às diferenças culturais e línguas locais.

De importância para a análise do tema, não se pode deixar de mencionar as cartas de princípios, que são iniciativas da sociedade civil organizada, por vezes com participação governamental, com o intuito de implementar diretrizes para a moderação de conteúdo, com foco na proteção do usuário, nos *standards* de direitos humanos e na liberdade de expressão (Estarque; Archegas, 2021, p. 29). Mesmo não vinculando as plataformas, as recomendações podem auxiliar para a promoção de uma relação sadia, respeitando os direitos humanos, espaço virtual mais saudável e em respeito aos direitos humanos. A constatação, aliás, é de que após a edição dos Princípios de Santa Clara, as plataformas avançaram na garantia de formas de revisão das decisões de moderação de conteúdo, permitindo aos usuários mecanismos de contestação (Silva; Gertrudes, 2023, p.11).

Destaca-se nesse cenário a *Manila Principles*, de 2015, que foi editada com o objetivo propor medidas que garantam o direito à liberdade de expressão, a partir

de recomendações baseadas em instrumentos de direitos humanos internacionais, assim como em outros marcos internacionais legais (Electronic Frontier Foundation *et al.*, 2015). Em resumo, a carta considera que os intermediários devem ser protegidos por lei da responsabilização, com regras claras, objetivas, excluindo a reponsabilidade quando não tenham realizado qualquer modificação no conteúdo, sendo que a remoção deve vir acompanhada de uma ordem judicial. A carta prestigia também o devido processo legal, assim como estabelece que as leis, ordens e as práticas de restrição de conteúdo, devem levar em conta a necessidade e proporcionalidade. Por fim, ela menciona os deveres de transparência e de prestação de contas. Aliás, uma primeira crítica quanto aos desafios trazidos pelas novas tecnologias em defesa dos direitos humanos é a falta de transparência, considerando que o desenvolvimento das midias tem sido "uma caixa preta", porque os usuários desconhecem os processos de tomada de decisão (Gonzáles-Iza, 2021, p. 3).

Anos após foram apresentados os princípios de Santa Clara (*Santa Clara Principles)*, originados em 2018, na *Content Moderation at Scale* nos Estados Unidos. Na oportunidade, um grupo de estudiosos em direitos humanos e especialistas desenvolveu um conjunto de princípios sobre a melhor forma de obter transparência e responsabilidade quanto à moderação de conteúdo nas redes sociais. Doze grandes empresas aderiram aos Princípios de Santa Clara, com destaque para a *Apple*, *Facebook* (*Meta*), *Google*, *Reddit*, *Twitter* e *Github* (Electronic Frontier Foundation *et al.,* 2018).

De início foi apresentada a primeira versão, os Princípios de Santa Clara 1.0, sendo que com o passar do tempo e da necessidade de aprofundamento, surgiram os Princípios de Santa Clara 2.0. Quanto ao Princípios de Santa Clara 1.0, pautada sobretudo na transparência, preveem a necessidade de veiculação do número total de postagens removidas e de contas suspensas, o dever de avisar cada usuário cujo conteúdo for removido ou suspenso, e por fim, que as empresas devem garantir oportunidades para recursos e revisões (Electronic Frontier Foundation *et al.*, 2018). A transparência, o aviso e a possibilidade de recurso são importantes, na medida em que as decisões das plataformas são tomadas com desconfiança, mais destacadamente em função da limitação de acesso às informações inerentes à moderação de conteúdo (Monteiro *et al.*, 2021, p. 7).

Com relação aos Princípios de Santa Clara 2.0, que remonta os anos de 2020 e 2021, foram divididos em princípios fundamentais e operacionais. Os princípios fundamentais constituem-se em cinco: a) Direitos Humanos e devido processo legal; b) Regras e políticas compreensíveis; c) Competência Cultural; d) Envolvimento do Estado na Moderação de Conteúdo; e) Integridade e Explicação. Quanto aos princípios operacionais, foram mantidos os princípios de transparência, da notificação e do recurso (Electronic Frontier Foundation *et al.*, 2018).

Esses princípios operacionais revelam-se importantes na medida em que, conforme já mencionado, considerando o volume de publicações por minuto, foram desenvolvidas ferramentas de inteligência artificial que nem sempre atuarão com precisão na moderação de conteúdo, muitas vezes tolhendo a liberdade de expressão do usuário e o pior, muitas vezes podem, por equívoco, permitir o conteúdo nocivo e proibir o ativismo. Uma pesquisa do InternetLab apontou que tecnologias desenvolvidas e utilizadas por plataformas para aferir uma publicação podem não ser "capazes de diferenciar conteúdo de ódio dirigido a LGBTQs do conteúdo publicado pelos próprios membros da comunidade LGBTQ" (Monteiro *et al.*, 2021, p. 13).

Evidentemente, o fato de serem utilizados algoritmos no controle de conteúdo, faz com que a análise seja padronizada trazendo diversos problemas, razão pela qual deve-se garantir um percentual de revisão através de análise humana. Nesse contexto, iniciativas como o *oversight board*, projeto do *Facebook* e *Instagram*, que é um tribunal administrativo de revisão das decisões das máquinas (algoritmos) formado por personalidades do mundo todo (Barbosa, 2021, p. 762).

Por fim, a versão 2.0 dos Princípios de Santa Clara, estabeleceu dois princípios em relação ao Poder Público, considerando os compromissos internacionais, como por exemplo o respeito à liberdade de expressão, consagrado no art. 19 da Declaração Universal dos Direitos Humanos. São eles: 1) Removendo Barreiras à Transparência da Empresa, 2) Promovendo a Transparência do Governo (Electronic Frontier Foundation *et al.*, 2018).

Ainda, temos outras cartas de tamanha importância, como a *Change the Terms*, recomendações corporativas para impedir o crescimento das atividades extremistas e odiosas (Center for American Progress *et al*., 2021) e, no combate ao terrorismo, a *Christchurch Call*, elaborada após o ataque terrorista de 2019, realizado por um supremacista branco na cidade de *Christchurch*, na Nova Zelândia, fato de

grande repercussão pela audácia do atirador em transmitir sua ação ao vivo pelo *Facebook* (França; Nova Zelândia, 2019).

Bom referir ainda que, numa iniciativa individual, em 2020, o *Facebook* publicou a *Charting a Way Forward: Online Content Regulation*. Os principais pontos abordados são: a) Como moderar conteúdo prejudicial e ao mesmo tempo preservar a liberdade de expressão?; b) Como as regulamentações devem aumentar a responsabilidade das plataformas da Internet?; c) A regulamentação deve exigir que as empresas de Internet cumpram certas metas de desempenho?; e) A regulamentação deve definir qual "conteúdo prejudicial" deve ser proibido na internet? (Bickert, 2020).

A moderação de conteúdo tem como objeto diversos outros temas, como por exemplo o discurso de ódio, o racismo, a homofobia, a violência e o terrorismo, porém, as cartas de princípios são também direcionadas ao combate à desinformação, com a pretensão de tornar o ambiente virtual mais saudável. Especificamente no que se refere à desinformação, ganha destaque a Chamada de Paris, que a seguir será exposta. A *Paris Call* é uma declaração não vinculativa com o objetivo de combater a desinformação. Ela pressupõe a importância do ciberespaço em todos os aspectos da vida, entendendo que esse deve ser aberto, seguro, estável, acessível e pacífico. A *Paris Call* menciona ainda que o direito internacional, incluindo a Carta das Nações Unidas, o direito internacional humanitário e o direito internacional consuetudinário devem ser aplicados ao uso de tecnologias de informação e comunicação (TIC) pelos Estados (Paris, 2018).

A Chamada de Paris estabelece como premissa que os mesmos direitos que as pessoas têm enquanto estão *off-line* também devem ser resguardados enquanto estão *online*, reafirmando a aplicabilidade das leis internacionais de direitos humanos no ciberespaço. A chamada parte de nove princípios: a) Proteção dos indivíduos e da infraestrutura; b) Proteção da Internet; c) Defesa dos processos eleitorais; d) Defesa da propriedade intelectual; e) Não proliferação; f) Segurança do ciclo de vida; g) Higiene cibernética; h) Proteção contra o *hack* privado; i) Normas internacionais (Paris, 2018).

O terceiro princípio, ainda que afeto diretamente ao processo eleitoral, trouxe elementos para o combate à desinformação no ciberespaço, sendo que a partir da chamada, foi criada a Comissão Transatlântica de Integridade Eleitoral (TCEI),

formada por pessoas de diferentes origens com um objetivo comum: “garantir que as pessoas decidam livremente, com base em informações independentes, quem deve representá-las” (Paris, 2018, tradução própria). A TCEI é uma iniciativa da *Alliance of Democracies Foundation,* fundada por Anders Fogh Rasmussen em 2017, numa parceria entre a Microsoft e o Governo do Canadá (Paris, 2018).

Como parte dessas iniciativas, o Governo do Canadá realizou seis workshops, com o objetivo de evitar a interferência no processo eleitoral (Paris, 2018). Foram trabalhados temas como a melhora do compartilhamento de informações, interferência estrangeira *versus* influência aceitável do Estado-nação e interferência eleitoral em um ambiente de pandemia. Ainda, trouxe a corresponsabilidade do combate à desinformação entre o governo, a mídia de notícias, as plataformas de mídia social, a academia e a sociedade civil, com atenção para a ameaça de interferência na infraestrutura eleitoral. Ressaltou ainda necessidade de capacitação dos cidadãos, de forma a treinar a comunidade para combater as ameaças de interferência eleitoral (Canadá, 2020).

Em suma, a Carta de Paris e a TCEI oferecem diversas diretrizes para o combate à desinformação, pois ainda que direcionadas para o sistema eleitoral podem ser aproveitadas, em boa medida, para uma salutar moderação de conteúdo no tema da desinformação geral. A Comissão Transatlântica de Integridade Eleitoral (TCEI) tem o objetivo de partilhar práticas entre os eleitores e instituições democráticas. Ainda, busca a TCEI sensibilizar o público sobre efeitos adversos da interferência sobre a democracia, com foco na capacitação da sociedade civil e dos governos para defender a democracia (Paris, 2018; Canadá, 2020).

A realização dos *workshops* realizados pelo Governo do Canadá partiu da intenção de fomentar estudos no sentido de se evitar a interferência no processo eleitoral, reagir à interferência eleitoral durante uma pandemia e ainda, capacitar os cidadãos. Os objetivos dessas oficinas foram: a) saber mais sobre as melhores práticas de todo o mundo; b) destacar as principais observações de especialistas na área; d) identificar os próximos passos concretos; e) saber como o governo do Canadá pode combater melhor a interferência eleitoral (Canadá, 2020).

O primeiro *workshop* trabalhou a ideia de que o compartilhamento eficaz de informações é importante para o combate da interferência da desinformação no processo eleitoral, com as premissas de que todas as partes do ciclo eleitoral são

vulneráveis a interferências, de que se deve reconhecer que atividades isoladas podem, de forma cumulativa, agravar a interferência. Ainda, devem ser identificados os pontos de contato entre o governo, a sociedade civil e a indústria, com a comunicação simples e interativa, realizando o planejamento de cenários e de resposta rápida, com o compartilhamento das melhores práticas ao redor do mundo. O projeto passa por uma coordenação intragovernamental, envolvimento dos partidos políticos e a criação de um grupo de especialistas apartidários (Canadá, 2020).

O segundo *workshop* colheu recomendações de diversos especialistas sobre a “interferência estrangeira”, com a noção de que a coerção mina as liberdades da sociedade democrática. Ainda, reconheceu que o conceito de engano ou falta de transparência corrói a integridade institucional, objetivo fundamental dos atores autoritários. De se mencionar ainda a análise da intenção da atividade, pois o ator estrangeiro buscará perturbar, manipular, danificar ou corroer a confiança nas organizações, instituições e processos democráticos (Canadá, 2020).

De sua vez, o terceiro *workshop* tratou do combate à interferência eleitoral em um ambiente de pandemia, considerando o aumento da desinformação com fundamento no medo sobre a segurança no voto. Como recomendações, a garantia de que a segurança cibernética não seja deixada de lado, o trabalho em conjunto para conscientizar sobre os processos de votação e possíveis atrasos. Mencionou ainda o treinamento de parceiros locais na conscientização e o aproveitamento das tecnologias de votação confiáveis, adaptadas às suas comunidades, com combinações de votação pessoal, por correio e portátil (Canadá, 2020).

O quarto *workshop* se voltou para a interferência estrangeira no ambiente de informação, através do esforço e compartilhamento de responsabilidades entre o Governo, as redes de notícias, as plataformas de mídia social, a academia e a sociedade civil. As redes de notícias devem oferecer informações úteis, manter-se como “árbitro jornalístico da verdade”. Já as plataformas de mídia social devem reconhecer o valor das parcerias com especialistas, ser transparentes com seus esforços para lidar com a desinformação. O Governo deve evitar a intervenção durante as eleições, fazendo com que as pessoas certas respondam à interferência, bem como garantir o acesso a informações em tempo real (Canadá, 2020).

Já o quinto *workshop* tratou da defesa, detecção e recuperação, com práticas de recomendações ligadas à proteção dos sistemas de recenseamento

eleitoral, sistemas de votação e implementação de auditorias. Por fim, o sexto *workshop* traz recomendações relacionadas à capacitação dos cidadãos, com o envolvimento de palestrantes e organizações comunitárias confiáveis, líderes reconhecidos para comunicar informações precisas com eficácia. Ainda, que o combate à desinformação deve ser pro ativo, com a inundação do espaço com informações precisas, educação das pessoas sobre como votar e os mecanismos da democracia e a construção do discurso coeso na sociedade em torno de valores da democracia – fornecer mensagens positivas. Mencionou ainda a necessidade da construção de uma alfabetização digital e a utilização de um sistema de verificação de fatos eficaz (Canadá, 2020).

Como se percebe, além do controle na moderação de conteúdo, da legislação eleitoral, uma série de medidas devem ser tomadas para minimizar o impacto da desinformação em massa. Em resumo, a carta e as iniciativas dele advindas, ressaltam a importância da informação de qualidade, do envolvimento de todos os atores, da necessária defesa das instituições democráticas - alvo frequente da estratégia de desinformar, e da necessidade de promover a capacitação e a alfabetização digital.

4 DIRETRIZES INTERNACIONAIS NO COMBATE À DESINFORMAÇÃO NAS MÍDIAS SOCIAIS COM FOCO NA MODERAÇÃO DE CONTEÚDO

A desinformação através das mídias sociais, como visto, tem levado diversas entidades governamentais e privadas a se debruçar sobre o tema como forma de estabelecer diretrizes e mecanismos de mitigação do impacto negativo. Também como foi verificado, tendo em vista o tensionamento entre o combate à desinformação e o direito à liberdade de expressão, é inegável o papel dos direitos humanos para servir de baliza em relação à atuação das empresas detentoras das plataformas.

Mais recentemente, podemos destacar documentos importantes, que serão a seguir analisados, quais sejam: o Relatório da ONU sobre Desinformação, Liberdade de Opinião e Expressão de 2021 e o Código de Conduta Reforçado da União Europeia de 2022.

Com base nessas informações, buscar-se-á estabelecer mecanismos de combate à desinformação com foco na moderação de conteúdo.

4.1 *STANDARDS* GLOBAIS DE DIREITOS HUMANOS

Foi referido no capítulo anterior que as principais plataformas de mídias sociais possuem usuários em diversos países e que, apesar de serem empresas privadas, sua atuação possui forte impacto social. Desta forma, muitos estudiosos da matéria apontam a necessidade de direcionar a atuação das redes através de um direito que seja universal, capaz de trazer uma solução ao menos linear para responder aos enormes desafios impostos por essa nova realidade. Os *standards* de Direitos Humanos, assim, são apontados como um norte para a atuação das mídias sociais.

Em princípio, não há uma obrigação direta das empresas em aplicar os Direitos Humanos, contudo, há uma vinculação clara da necessidade de respeitá-los. No ano de 2011, a Organização das Nações Unidas adotou os Princípios Orientadores sobre Negócios e Direitos, conhecidos como Princípios *Ruggie*, norteando os estados a coibir e proteger os cidadãos contra abusos praticados por empresas nos seus territórios, no que se encaixa, evidentemente, às plataformas digitais. Portanto, trata-se de um mecanismo internacional palpável e de cumprimento ao menos esperado

para nortear a atividade de empresas cuja atuação é impactante na vida das pessoas e da comunidade (Jones, 2019, p. 31).

Dentre a ampla gama de Direitos Humanos, estabelecendo uma relação entre Declaração Universal dos Direitos Humanos – DUDH e o Pacto Internacional de Direitos Civis e Políticos - PIDCP, Jones (2019, p. 32) enumera cinco direitos fundamentais considerando dois deles em conjunto: a) direito à liberdade de pensamento e o de ter opiniões sem interferência; b) direito à privacidade; c) direito à liberdade de expressão; d) o direito de votar nas eleições.

O artigo 18 da DUDH prevê que “todos têm direito à liberdade de pensamento, consciência e religião”, ao passo que o art. 19 dispõe de forma clara que “toda pessoa tem direito à liberdade de opinião e expressão; este direito inclui a liberdade de opinião sem interferência e buscar, receber e difundir informações e ideias por qualquer meio e independentemente das fronteiras” (ONU, 1948). Contudo, muito embora a positivação da garantia, há pouca interpretação sobre essa liberdade que os Direitos Humanos classificam como absoluta, percebendo-se tentativas deliberadas de influenciar os pensamentos e opiniões, através da mídia. Ocorre que a liberdade de pensamento possui também como dimensão justamente o direito de não ter a opinião inconscientemente manipulada, num cenário em que se constata o abuso da utilização das plataformas digitais para manipular as opiniões através da desinformação, seja para fins políticos, econômicos e outros (Jones, 2019, p. 37).

O direito de ter opiniões sem interferência é absoluto, reforçando a ideia de que se proíba o controle do santuário interno da mente do indivíduo. De igual modo, não se pode forçar a divulgação de uma opinião, o que minaria, a *contrario senso*, o direito de opinião sem interferência. Em resumo, o direito de ter opiniões, diferentemente do direito à liberdade de expressão, não estará sujeito às restrições governamentais. Assim, os esforços deliberados para influenciar opiniões através de meios não autênticos constituir-se-ia numa violação ou manipulação da autonomia mental (Aswad, 2020, p. 323).

Quanto ao direito à privacidade, o artigo 12 da DUDH menciona que ninguém estará sujeito “a intromissões arbitrárias na sua vida privada, na sua família, no seu domicílio ou na sua correspondência, nem ataques à sua honra e reputação. Toda pessoa tem direito à proteção da lei contra tais interferência ou ataques” (ONU, 1948).

Assim, o direito à privacidade está ligado ao direito de opção do cidadão em não divulgar seus dados pessoais, o que ainda merece grande reflexão. Em 2018 o Gabinete do Alto Comissariado das Nações Unidas para os Direitos Humanos (ACNUDH) elaborou relatório sobre o direito à privacidade na era digital destacando a necessidade de padrões mínimos que devem regular o tratamento de dados pessoais. No relatório, considerou-se que o tratamento de dados deve ser justo, lícito e transparente. Ainda, que os usuários devem ser informados e o uso dos dados através do consentimento livre e inequívoco e ainda, que o tratamento dos dados pessoais deve ser proporcional e dentro de um propósito legítimo, com segurança da sua manutenção. Porém, o que se verifica é uma ampla utilização desses dados em processos algorítmicos e políticos que permitem a realização de campanhas em escala assustadora, sem contar nas enormes receitas geradas para as plataformas digitais, numa ligação clara entre o direito à liberdade de pensamento e o direito à privacidade, refletindo diretamente na tomada de decisões (Jones, 2019, p. 38-39).

A preocupação com o uso de dados, aliás, não surgiu apenas pela forma com que as empresas adquirem esses dados, mas sobretudo pela forma com que elas os utilizam, com publicidade direcionada e com o fim de maximizar o envolvimento e o tempo do usuário na rede. Nesse sentido, Ellen Weintraub, Comissária da Comissão Eleitoral Federal do EUA, defendeu que a "microssegmentação" deveria ser severamente limitada para anúncios eleitorais, pois aumenta o risco de danos causados pela desinformação. Ainda, surgem questões relacionadas ao fato de que o uso de algoritmos promove conteúdos abusivos, odiosos e discriminatórios, tendo em vista a tendência do ser humano em prestar mais atenção ao conteúdo sensacionalista (Aswad, 2020, p. 320).

De outro ângulo o conceito de má-informação (*malinformation)* parece guardar relação com a violação à privacidade de uma pessoa sem qualquer interesse ou justificativa pública. Lembrando, a má-informação é baseada na realidade, mas usada para causar danos a uma pessoa. Como exemplo trazido por Wardle e Derakhshan (2019, p. 48), seria um relatório revelador da orientação sexual de uma pessoa sem qualquer justificativa plausível ou interesse público.

O artigo 19 da DUDH garante a liberdade de opinião e expressão, incluindo a liberdade de opinião sem interferência e de "buscar, receber e difundir informações e ideias por qualquer meio e independentemente das fronteiras" (ONU, 1948). De sua

vez, o PIDCP prevê no artigo 19 que o direito inclui “a liberdade de procurar, receber e difundir informações e ideias de todos os tipos, independentemente de fronteiras, oralmente, por escrito ou por escrito, em a forma de arte, ou através de qualquer outro meio de sua escolha” (ONU, 1966). Como restrições, dispõe o respeito aos direitos ou a reputação de outras pessoas, a proteção da segurança nacional, da ordem, da saúde ou da moral públicas, assim como estatui no seu artigo 20 a proibição de propaganda de guerra, da defesa do ódio nacional, racial, religioso, que constitua incitação à discriminação, hostilidade ou violência (Jones, 2019, p. 41).

Bento (2016, p. 96) ensina que a liberdade de pensamento e de expressão é ponto nodal do arcabouço institucional das sociedades democráticas, apresentando sua tripla função. Trata-se, em primeiro plano, de um direito individual que define a característica única dos seres humanos, a capacidade de pensar e de se comunicar para construir de forma coletiva a sua representação da realidade, seja na arte, na ciência, na tecnologia ou na política. Num segundo plano, possui relação com a democracia, permitindo aos cidadãos manifestar sua expressão, seus questionamentos e contestar livremente os desígnios da vida política. Numa terceira face, a liberdade de expressão é instrumento de defesa de outros direitos, tais como o direito de reunião, associação, liberdade religiosa, participação política, enfim, jamais poderá ser compreendida apenas no sentido individual, mas também no sentido difuso.

O direito à liberdade de expressão compreende a expressão de ideais e o seu recebimento, sendo uma importante medida para o combate às tentativas governamentais de minar as dissidências e controlar o fluxo de informações. Nesse contexto, as restrições devem ser interpretadas taxativamente, sendo evidente, no entanto, que os limites se encontram na preservação dos direitos de terceiros e em questões como o discurso de ódio que geram discriminação, violência, bem como a garantia da segurança nacional. Jones (2019, p. 43) menciona que na segunda grande guerra viu-se uma disseminação massiva de propaganda desinformação através do rádio, trazendo temores e incitação à guerra e à violência. Outro exemplo claro de aplicação dessas restrições vem da decisão do Tribunal Constitucional Federal da Alemanha (1 BvR 673/18), de 22 de junho de 2018, que decidiu pela compatibilidade com a lei fundamental e a liberdade de expressão nela insculpida, a criminalização pela negação do genocídio, obtido através de análise de uma reclamação

constitucional (Verfassungsbeschwerde). Na decisão, consagrou-se o entendimento de que a desinformação prejudica a formação da opinião e a negativa do genocídio ultrapassa os limites do debate pacífico, afetando a paz social, não se encontrando o discurso de sua negação protegido pela liberdade de expressão (Sarlet, 2019, p. 1214).

A *internet* e as plataformas de mídias sociais apresentam inúmeras possibilidade para o exercício da liberdade de expressão, incluindo as minorias e discursos alternativos, contudo, a definição de parâmetros para evitar os discursos estabelecidos nas restrições é algo desafiador. O debate é desafiador, pois de um lado há os que defendem que a liberdade de expressão é ilimitada, não permitindo qualquer controle sobre o que é dito, ao tempo em que muitos entendem que o fenômeno da desinformação deve ser expurgado radicalmente do mundo digital, contudo, sendo que a radicalidade dos extremos não parece encontrar uma resposta (Barbosa, 2021, p. 734).

O mais indicado, parece, é compreender que as restrições à liberdade de expressão devem estar previstas em lei e na medida necessária para um fim legítimo inerente à proteção dos direitos de terceiros, da segurança nacional, da ordem pública ou da saúde e moral públicas. Assim, as restrições devem ser razoáveis e não podem servir de pretexto para permitir, aos Estados, a restrição desmedida como por exemplo, restringir o debate político em nome do direito de voto. Em linhas gerais, a invocação de uma restrição à liberdade de expressão deve vir acompanhada de uma justificativa legítima, específica e individualizada, assim como pela observância da proporcionalidade da medida que deve possuir relação direta entre a expressão e a ameaça (ONU, 2021a, p. 19).

Contudo, da liberdade de expressão não pode resultar, por exemplo, uma ofensa gratuita e sem justificativa em qualquer fim legítimo, tanto a uma pessoa como a uma comunidade e, se assim o for, estar-se-á diante de um abuso da liberdade. E nesse contexto parece claro que a liberdade de expressão pode e deve ser sindicado como qualquer outro direito (Barbosa, 2021, p. 744). A liberdade de expressão não pode servir de escudo para o incitamento ao ódio, para a manipulação digital e não autêntica da informação, para a discriminação e incitação à violência.

No entanto, nunca deve se perder de vista que, muito embora a liberdade de expressão não seja um direito absoluto, as restrições devem ser aplicadas de forma

posterior e jamais através de censura prévia. Essas mesmas restrições devem ser previstas em lei (ato do legislativo), de forma adequada, com um fim legítimo e de uma necessidade social premente e indispensável. Em relação à proporcionalidade, o direito de resposta ou retratação vai ser sempre preferível à reparação civil e o direito penal somente deve ser utilizado em última razão, tais como em casos de danos individuais ou à ordem pública (Bento, 2016, p. 104).

Sobre a desinformação, bom ressaltar ainda que as principais plataformas de mídias sociais foram desenvolvidas no Vale do Silício, sendo que os Estados Unidos possuem uma abordagem expansiva do conceito de liberdade de expressão, focando o combate através da verificação de fatos e na alfabetização digital, medidas importantes, mas insuficientes, pois os desafios são muito maiores. Ainda, se verifica na prática é que a moderação de conteúdo está, em grande escala, nas mãos das plataformas digitais, além do que há a necessidade de mais transparência na utilização dos algoritmos como forma de compreender melhor o seu impacto (Jones, 2019, p. 44).

Especificamente em relação à desinformação, Jones (2019, p. 46) entende que se de um lado o direito à liberdade de expressão deve ser implantado com o cuidado de não permitir o uso indevido do setor público para evitar a censura, tal não é motivo para permitir que importantes interesses públicos fiquem nas mãos do setor privado. Assim como não se deve impor uma legislação que simplesmente proíba ou censure a desinformação, sem foco, não há como não observar o direito internacional para, de forma proporcional ao dano, impor as restrições de forma cuidadosa de forma a evitar, com a velocidade e o fluxo da informação *online*, que se utilize as campanhas de desinformação como arma de guerra.

Por fim, o artigo 21 da DUDH menciona o direito humano de participar nos assuntos públicos e de votar, no sentido de que toda a pessoa tem direito de exercer o poder, seja diretamente ou por meio dos seus representantes, sendo a vontade do povo a representação da autoridade do governo, garantida a livre expressão, norma essa que encontra simetria com o artigo 25 do PIDCP (ONU, 1948). Assim, o direito de voto garante o direito de participar em eleições livres e justas, num sistema eleitoral livre, o que envolve a liberdade de pensamento, opinião e expressão, incluindo a liberdade de exercer atividade política individualmente ou através de partidos políticos,

livre debate, de crítica, de se opor, de publicar material político, de fazer campanha eleitoral e de divulgar ideias.

A noção de eleições democráticas, de expressão da vontade política, é relacionada com o princípio fundamental da autodeterminação dos povos, sendo caracterizada por três eixos fundamentais de acordo com os padrões internacionais: o direito de participar na direção dos assuntos públicos; o direito de votar e de ser eleito e o direito de ter acesso às funções públicas. As limitações ao direito de participação devem estar previstas em lei sem conteúdo discriminatório, com critérios objetivos e razoáveis. O direito à participação nos desígnios públicos somente é possível em conexão com outros direitos humanos e liberdade fundamentais, tais como a liberdade de expressão, o acesso à informação, a liberdade de associação e reunião, igualdade, educação, segurança e julgamento justo (ONU, 2021a, p. 9-14).

Assim, o direito de expressão que já possui enorme relevo, deve ter a proteção maximizada durante os processos eleitorais. No que se refere à liberdade de expressão na *internet*, deve de igual modo, estarem as proibições previstas em lei, de forma necessária, razoável e proporcional. Contudo, sabe-se que o espaço digital modificou substancialmente as comunicações, sobretudo em razão do enorme fluxo de informação e do fato de que cada usuário possui totais condições de ser o promotor ou editor da informação, ao invés de mero expectador/receptor. Há um dilema que inclui a proliferação de meios digitais para censurar a informação e de outro lado, a divulgação de desinformação nas redes, o que representa um grande desafio em termos de qualidade da informação. Esses componentes refletem no processo de escolha, afetando a capacidade dos usuários/eleitores de formar opinião livre e independente (ONU, 2021a, p. 14).

A desordem da informação e a utilização de mecanismos de manipulação de grandes volumes de dados é um grande desafio, pois deve prever a proteção do direito humano de liberdade de expressão ainda que as informações não sejam corretas e que possam chocar e perturbar, porém, devendo ser considerado que a desinformação generalizada representa uma grande ameaça à democracia. De tudo o que foi visto, sabe-se que as plataformas de mídias sociais podem ser utilizadas para influenciar o resultado das eleições, através do uso de algoritmos e tratamento de dados, de forma que desinformação pode resultar na violação justamente no direito de eleições livres e democráticas, pois a liberdade de expressão e o acesso à

informação podem ser afetadas pelo acesso desenfreado e manipulado de informações falsas (ONU, 2021a, p. 14).

Reforçando o posicionamento, Jones (2019, p. 49), também expressa a preocupação no sentido de que em tais circunstâncias, o uso de algoritmos que priorizam a desinformação, que manipulam essa vontade, por toda a experiência e estudos recentes, podem também reduzir a capacidade de influência do debate público e influenciar no direito ao voto livre. Como se vê, os principais *standards* de direitos humanos relacionados à desinformação possuem entrelaçamentos muito claros, além de se relacionarem com outros direitos humanos e fundamentais para a convivência pacífica no planeta.

A desinformação, portanto, é um problema que precisa de tratamento adequado e lida com componentes sensíveis. Nesse contexto, para melhor compreender o estágio da discussão internacional sobre o combate à desinformação, será abordado no próximo tópico dois documentos recentes e de extrema relevância, quais sejam, o Relatório da ONU sobre Desinformação, Liberdade de Opinião e Expressão de 2021 e o Código de Conduta Reforçado da União Europeia de 2022.

4.2 RELATÓRIO DA ONU SOBRE DESINFORMAÇÃO, LIBERDADE DE OPINIÃO E EXPRESSÃO DE 2021 E O CÓDIGO DE CONDUTA REFORÇADO DA UNIÃO EUROPEIA DE 2022

O mundo todo busca respostas e medidas para enfrentar o fenômeno, tendo em vista a crescente difusão da desinformação através das mídias sociais, alimentadas por atores estatais e não estatais, e os vários danos experimentados. Porém, antes de ingressar nos dois relatórios, o da ONU de 2021 e o da União Européia de 2022, bom mencionar que em 2107 diversas organizações elaboraram uma declaração conjunta. Pelos seus órgãos, a Organização das Nações Unidas (ONU), a Organização para a Segurança e Cooperação na Europa (OSCE), a Organização dos Estados Americanos (OEA) e a Comissão Africana dos Direitos Humanos e dos Povos (CADHP) editaram a declaração conjunta sobre liberdade de expressão e “fake news”, desinformação e propaganda (OEA, 2017).

O documento foi dividido em tópicos, destacando-se os princípios gerais no sentido de que iniciativas que cuidaram de estabelecer as restrições à liberdade de

expressão devem estar previstas em lei, para um interesse legítimo reconhecido pelo direito internacional e de forma proporcional, para proibir a defesa do ódio que constituam incitação à violência, discriminação ou hostilidade. Ressaltou que os intermediários nunca devem ser responsáveis por qualquer conteúdo de terceiros, a não ser que intervenham no conteúdo ou se recusem a obedecer a uma ordem adotada por um órgão supervisor independente (como um Tribunal). Foi expressa a necessidade de proteção aos indivíduos por responsabilidade de meramente redistribuir ou promover, por meio de intermediários, conteúdos de que não são autores e que não tenham modificado. Ainda, que os bloqueios de um IP (*internet protocol*) ou de um site inteiro só se justifica desde que previsto em lei e necessário para a proteção de um direito humano ou interesse público legítimo. O sistema de filtragem de conteúdo por um governo e que não são controlados pelo usuário final não são justificáveis como restrição à liberdade de expressão (OEA, 2017).

No tópico seguinte (Padrões sobre Desinformação e Propaganda), foi declarado que não deve haver proibição geral baseada em ideias vagas e ambíguas. As leis de difamação criminal são indevidamente restritivas, devendo ser abolidas. Já as leis civis são legítimas apenas se os réus tiverem plena oportunidade e não provarem a veracidade. Os atores estatais não devem promover ou patrocinar declarações sabidamente falsas ou que deveriam saber, assim como devem divulgar informações confiáveis e fidedignas (OEA, 2017).

Prosseguindo, no terceiro tópico (Ambiente favorável à liberdade de expressão), é ressaltada a obrigação positiva do Estado de promover um ambiente de comunicação livre, independente e diversificado, de estabelecer uma estrutura regulatória clara para as emissoras que seja supervisionada por um órgão protegido contra interferência ou pressão política e comercial e que promova um setor de radiodifusão livre, independente e diversificado. Ainda, devem garantir a presença de meios de comunicação de serviço público fortes, independentes e com recursos adequados, que operem sob um mandato claro para servir o interesse público geral e estabelecer e manter altos padrões de jornalismo (OEA, 2017).

Os Estados devem também implementar medidas para promover a diversidade da mídia, com subsídios ou outras formas de apoio financeiro ou técnico para a produção de conteúdo midiático diversificado e de qualidade, proibição da concentração indevida de propriedade de mídia, bem como a exigência de

transparências dos meios de comunicação. Devem ser implementadas medidas para promover a alfabetização midiática e digital, além de promover a igualdade, a não discriminação, a compreensão intercultural e outros valores democráticos (OEA, 2017).

Importante para com os objetivos do presente estudo, em relação aos intermediários foi estabelecido que, na medida em que pretendem tomar ações para restringir o conteúdo de terceiros, as políticas devem ser claras e pré-determinadas, calçadas em critérios objetivamente justificáveis e jamais em objetivos ideológicos ou políticos. De preferência, sejam adotadas as políticas após consulta aos seus usuários. Aos usuários deve ser proporcionada a compreensão facilitada das políticas e práticas e a moderação deve respeitar as garantias mínimas do devido processo legal, com a devida notificação. A moderação de conteúdo por processos automatizados (algoritmos ou outros) deve ser razoável, de acordo com necessidades competitivas ou operacionais legítimas. Por fim, os intermediários devem apoiar a pesquisa e o desenvolvimento de soluções tecnológicas apropriadas para o combate à desinformação e propaganda, bem como devem apoiar iniciativas para a verificação de fatos (OEA, 2017)

Contudo, o problema da desinformação foi maximizado nos últimos anos, o que levou a Organização das Nações Unidas em 2021, com a contribuição de 119 organizações da sociedade civil e entidades acadêmicas, 3 organizações internacionais, 3 Estados Membros e 3 empresas, além de consultas às organizações da sociedade civil e reuniões com Estados Membros, empresas de mídia e especialistas, a editar através da Relatora Especial da Organização das Nações Unidas sobre a promoção e proteção do direito à liberdade de opinião e expressão, Irene Khan, o relatório A/HRC/47/25 com diversas conclusões e recomendações para a proteção da liberdade de expressão (ONU, 2021b)

O relatório conclui, inicialmente, que não se deve perder de vista a importância da tecnologia digital para a democracia, de forma que o direito à liberdade de opinião e expressão não faze parte do problema. Ao mesmo tempo, reconhece a complexidade do fenômeno da desinformação e suas graves consequências para a destruição da confiança nas instituições democráticas (ONU, 2021b, p. 17).

A desinformação, de acordo com o relatório, é um fenômeno complexo e multifacetado com graves consequências para a confiança das pessoas nas

instituições democráticas, bem com expõe a facilidade de sua propagação onde a informação pública é deficiente e o jornalismo investigativo é limitado. Enfim, a desinformação é um problema, e são ainda mais problemáticas as repostas pelos Estados e pelas empresas, gerando desconfiança dos usuários na integridade da informação (ONU, 2021b, p. 17). No mesmo sentido, como parte das iniciativas do terceiro princípio da Chamada de Pais, o Governo do Canadá trouxe no primeiro *workshop* a ideia de que o compartilhamento de informações de qualidade é fundamental para o combate da interferência da desinformação no processo eleitoral (Canadá, 2020).

Assim, num primeiro plano, o Relatório demonstra sua preocupação com a liberdade de expressão e opinião, apontando serem necessárias respostas coletivas e multidimensionais, sendo os Estados os principais responsáveis por fazer cumprir as leis internacionais de direitos humanos. Os Estados, portanto, não devem encorajar ou divulgar notícias e postagem que sabem ou deveriam saber que são falsas, assim como não devem autorizar o fechamento da *internet* como meio de combater a desinformação. A liberdade de expressão não deve ser restringida, com as exceções do disposto nos artigos 19 e 20 do Pacto Internacional sobre Direitos Civis e Políticos, já referidos, cujas limitações devem ser interpretadas de forma estrita e rigorosa (ONU, 2021b, 1966).

Quanto às regras das empresas sobre moderação de conteúdo, estas devem ser claras, objetivas e transparentes, de modo que o usuário possa prever a atitude a tomar (Jørgensen, 2021, p.2). E mais, o foco da regulação dos meios de comunicação social não deve ser o conteúdo, mas sim a transparência, o direito ao processo legal e o dever de precaução das empresas relativamente aos direitos humanos, garantindo a independência dos órgãos reguladores. Ainda, deve ser priorizada a reorientação do modelo de negócios das mídias, programado para valorizar o sensacionalismo e obter lucro com isso (ONU, 2021b, p. 18).

Neste passo, o *Facebook*, ao submeter suas contribuições ao relatório A/HRC/47/25, reconhece que a repressão da desinformação pode restringir a liberdade de expressão, alegando que as regras devem ser pautadas na legalidade, necessidade, proporcionalidade para proteger danos a outros direitos, tais como a segurança nacional, a ordem pública e a saúde pública. Ainda, destaca-se das considerações do *Facebook* a revelação de que a plataforma subdivide o combate à

autenticidade em três planos: *misinformation, disinformation e* operações de influências, essas últimas que oferecem maior gravidade. O Facebook propôs diversas medidas para combater as operações de influência, dentre as quais se destacam a transparência nos anúncios de propaganda política; desenvolvimento de relatórios sobre comportamento inautêntico; sanções econômicas, diplomáticas e/ou criminais; apoio à pesquisa técnica e apoio à alfabetização midiática e digital (Facebook, 2021).

Retomando o relatório, menciona ainda que os Estados devem garantir o direito à informação, aumentando a transparência e divulgando dados oficiais na *Internet* e *off-line*, além de garantir a segurança dos jornalistas. Aponta que a mídia, a informação e a alfabetização digital capacitam as pessoas, preparando-as contra a desinformação, devendo fazer parte do currículo nacional (ONU, 2021b, p. 18). Na declaração conjunta de 2017, a ONU já estabelecia a importância da alfabetização midiática e digital.

Acentua novamente o relatório que as mídias devem respeitar os direitos humanos, pois apesar de serem empresas privadas, têm na sua atuação um grande impacto na esfera pública, o que gera uma enorme responsabilidade social. Diante do fato de que são poucas as empresas que dominam o mercado, devem melhorar a moderação de conteúdo e revisar seus modelos de negócios. Devem ainda capacitar os usuários, aumentar a transparência e garantir o devido processo (ONU, 2021b, p. 18). Nesse ponto, o relatório vai ao encontro das iniciativas concluídas a partir da *Paris Call*, posto que o sexto *workshop* promovido pelo Governo do Canadá, erigiu como fundamento que o cidadão, quando bem informado, acaba pensando criticamente, compreendendo os fundamentos da democracia (Canadá, 2020).

Sobre o modelo de negócios, as empresas devem garantir que suas atividades comerciais, a práticas de coleta e processamento de dados estejam em conformidade com os padrões internacionais de direitos humanos. Ainda, as empresas devem revisar seus modelos de publicidade, garantindo que não afetem negativamente a diversidade de opiniões e ideias. Conforme o relatório, as mídias devem adotar políticas de publicidade e conteúdo claras e rigorosamente definidas sobre *Disinformation* e *Misinformation* que estejam de acordo com a lei internacional de direitos humanos. Sobre a sua forma de atuação, as plataformas de mídias sociais devem fornecer informações claras e significativas sobre os parâmetros de seus

algoritmos, permitindo que os usuários recebam múltiplas possibilidade de visualizações, aumentando as suas escolhas (ONU, 2021b, p. 19).

No que tange à transparência, as empresas devem publicar relatórios completos, detalhados e contextualizados, assim como relatórios sobre circunstâncias excepcionais. Os relatórios devem compreender: a) medidas tomadas, b) recursos interpostos; c) número de conteúdos que são partilhados ou consultados, d) o número de utilizadores alcançados e e) o número de reclamações e pedidos de remoção. Quanto aos recursos, as empresas devem estabelecer mecanismos internos de apelação sobre a moderação, com a criação de mecanismos externos de supervisão (ONU, 2021b, p. 19). O ponto reafirma os três princípios operacionais de Santa Clara 1.0.: número, aviso e recurso (Electronic Frontier Foundation *et al.*, 2018).

As empresas devem ainda priorizar o combate à desinformação de gênero online, dedicar mais recursos para entender melhor os contextos locais que promovem a *Disinformation* e *Misinformation,* com foco nas disparidades de linguagem e conhecimento. Por último, o Relatório realça a importância do sistema de direitos humanos das Nações Unidas, considerando o Conselho de Direitos Humanos fundamental para combater a *disinformation* e *misinformation*, com respeito à liberdade de opinião e expressão, consultando regulamente os Estados, empresas, organizações da sociedade civil e atores, para estabelecer iniciativas de proteção e promoção dos direitos humanos no espaço digital (ONU, 2021b, p. 20).

Como iniciativa mais recente e contando com a adesão de 34 signatários, dentre eles a Adobe, Avaaz, Google, Microsoft, Meta, TikTok, Twitter, a União Europeia promoveu uma revisão do Código de Conduta de 2018, com a edição do Código de Conduta Reforçado Sobre Desinformação.

Os signatários reconheceram o seu papel no combate à Desinformação e as definições tecnicamente mais corretas de *misinformation*, como sendo conteúdo falso ou enganoso compartilhado sem intenção prejudicial, mas que podem ser prejudiciais e *disinformation*, como sendo o conteúdo falso ou enganoso divulgado com intenção de enganar e causar danos, buscando ganhos econômicos ou políticos. Reconheceram as operações de influência de informações, que seriam exercícios coordenados por atores estrangeiros para inserir uma série de meios enganosos a um público alvo; e as operações estrangeiras de interferência no espaço da informação, que podem ser entendidas como os esforços coercitivos e enganosos para romper a

livre formação e expressão da vontade política dos indivíduos por um ator estatal estrangeiro (União Europeia, 2022).

Foram assumidos 44 compromissos, divididos em 8 áreas: a) Desmonetização dos fornecedores de desinformação; b) Transparência da propaganda política; c) Garantia da integridade dos serviços; d) Capacitação dos usuários; e) Capacitação de pesquisadores; f) Capacitação da comunidade de verificação de fatos; g) Centro de transparência e força-tarefa; h) Estrutura de monitoramento reforçada.

No primeiro campo, houve o compromisso dos signatários para desmonetizar a disseminação de desinformação, com a melhora das políticas e sistemas que determinam a elegibilidade do conteúdo a ser monetizado, com a criação de um grupo de trabalho para o desenvolvimento de uma metodologia e de relatórios sobre os esforços de desmonetização (União Europeia, 2022). O Relatório da ONU (2021b) já apontava que a moderação de conteúdo, por si, não é suficiente para mudar os comportamentos, caso não haja uma revisão no modelo de negócios.

Ainda, comprometeram-se os signatários em evitar o uso indevido de sistemas de publicidade para disseminar desinformação na forma de mensagens publicitárias, bem como trocar as melhores práticas para fortalecer a cooperação com os atores relevantes. Nesse contexto, já foi mencionado que o modelo de negócios das mídias é baseado na receita de publicidade, sendo que as plataformas são projetadas, inclusive por seus algoritmos, para viciar os “usuários” e fazê-los permanecer o máximo de tempo conectado. Por isso, diminuir os recursos financeiros destinados àqueles que disseminam desinformação é primordial para minimizar os danos.

Sobre a publicidade, ao reconhecer a importância da publicidade política e publicitária na formação das campanhas e nos debates públicos, com o compromisso de adotar uma definição comum de publicidade política, bem como indicar claramente até que ponto tal publicidade é permitida ou proibida, com a emissão de rotulagem nos anúncios para distingui-los como conteúdo pago, para facilitar a compreensão aos usuários. Os signatários se comprometeram, também, a manter repositórios de informações políticas, mantendo o monitoramento e pesquisa contínuos para entender e responder aos riscos relacionados à desinformação na publicidade política (União Europeia, 2022).

Quanto à integridade dos serviços, os signatários se comprometeram a intensificar o combate à *misinformation, disinformation* e aos comportamentos manipuladores inadmissíveis. No primeiro compromisso, vem previsto o dever de limitar os comportamentos e práticas manipulativas inadmissíveis em seus serviços, devendo tais serem revistos periodicamente à luz das últimas evidências sobre as condutas, incluindo a criação e uso de contas falsas, aquisições de contas e amplificação orientada por robôs, operações de *hack-and-leak* (hackeamento e vazamento), condutas destinadas a ampliar artificialmente o alcance, dentre outros artifícios (União Europeia, 2022). Nesse ponto, o Relatório da ONU (2021b), já levantava a que as definições apresentadas pelas mídias geralmente são muito amplas, nem sempre explicando de forma clara o tipo de dano e a probabilidade de causar remoção de conteúdo, rotulagem ou outras sanções.

Ainda, os signatários, comprometeram-se a operar canais de intercâmbio entre suas equipes relevantes para compartilhar informações sobre o comportamento inautêntico entre as plataformas. Estudo da FGV revela a importância desse compromisso ao descrever que nas eleições nacionais brasileiras de 2018 foi demonstrado o surgimento de contas automatizadas como estratégia de manipulação por meio de boatos e difamação, com a utilização de robôs, potencializando a disseminação (Ruediger; Grassi, 2022).

No que se refere à capacitação de usuários, reconheceram os signatários como uma importante medida para detectar e denunciar informações falsas e/ou conteúdo enganoso, reconhecendo, porém, de acordo com o artigo 10 da Convenção Europeia dos Direitos Humanos e os artigos 7, 8, 11, 47 e 52 da Carta dos Direitos Fundamentais da União, as sinalizações pelos usuários deve respeitar a liberdade de expressão, o direito a comunicações privadas, à proteção de dados pessoais e uma solução proporcional. O compromisso diz respeito ao fortalecimento dos esforços na área de alfabetização midiática e do pensamento crítico, com a inclusão de grupos vulneráveis, equipando os usuários para identificar a desinformação (União Europeia, 2022).

Dentro dessas balizas, os signatários se comprometeram ainda, a fornecer aos usuários a necessária funcionalidade para sinalizar informações falsas e/ou enganosas prejudiciais que violem as políticas da comunidade, tomando as medidas necessárias para garantir que tal seja protegida do abuso humano ou do

comportamento não autêntico, evitando a sinalização em massa para silenciar vozes com outra linha de pensamento (União Europeia, 2022).

Nessa área, constou do Código de Conduta o compromisso de informação (aviso) ao usuário quando sua conta tenha sido sinalizada, dando-lhe a possibilidade de recurso. Quanto aos serviços de mensagem, comprometem-se os signatários a manter e implementar recursos e iniciativas para capacitar os usuários a pensar criticamente sobre as informações que recebem, facilitando o acesso dos usuários a informações oficiais. Ainda, comprometem-se a limitar a propagação viral da desinformação em seus serviços, como por exemplo, recursos para limitar a encaminhamento de informações em várias conversas (União Europeia, 2022).

Na área da capacitação da comunidade de pesquisa, reconheceram os signatários a importância de permitir o acesso aos dados da plataforma, bem como dar suporte adequado para suas atividades como parte de uma estratégia de combate à desinformação. Ainda, comprometem-se a realizar pesquisas com base em metodologia e padrões éticos, bem como compartilhar conjuntos de dados, resultados de pesquisas e metodologias com públicos relevantes (União Europeia, 2022). Nesse ponto, o Relatório da ONU (2021b) já sinalizava que a falta de transparência e acesso aos dados continua sendo uma deficiência das mídias, o que impede o controle independente e afeta a responsabilidade e a confiança, ressaltando que faltam mais informações sobre a confiabilidade e precisão dos sistemas de inteligência artificial.

Quanto à capacitação da comunidade de verificação de fatos, comprometem-se a criar um quadro de transparência e uma cooperação estruturada com a comunidade de verificação de fatos da União Europeia, incluindo e usando *fact-checking* nos seus serviços, com a obrigação de operar com base em rigorosas regras éticas e de transparência, com forma de manter a sua independência (União Europeia, 2022).

Quanto ao Centro de Transparência, reconhecem os signatários a importância da transparência na luta contra a desinformação, bem como o interesse legítimo do público em receber informações precisas sobre a implementação do Código, com o estabelecimento de um sítio eletrônico do Centro de Transparência. O Centro de Transparência deve conter informações atualizadas relacionadas à implementação dos compromissos assumidos, com informações claras, de fácil

pesquisa e compreensão, listando cada compromisso e medidas subscritas, os termos de serviço e as políticas de cada comunidade (União Europeia, 2022).

Ademais, os signatários se comprometeram a participar da Força Tarefa presidida pela Comissão Europeia, que pode convidar especialistas como moderatobservadores para apoiar os trabalhos, sendo as decisões tomadas por consenso. Os signatários comprometem-se a dedicar recursos financeiros e humanos para garantir a implementação de seus compromissos, fornecendo à Comissão Europeia, no prazo de 1 mês após o final do período de implementação (6 meses após a assinatura do Código) a linha base dos relatórios. Ainda, comprometem-se a trabalhar no Grupo de Trabalho para desenvolver indicadores e publicá-los no prazo de 9 meses a partir da assinatura. Devem ainda, em situações especiais, como eleições ou crise, a informação adequada de dados (União Europeia, 2022).

Por fim, as maiores empresas de mídia se comprometeram a serem auditadas às suas próprias custas, para avaliar a conformidade com os compromissos assumido, através de organizações independentes, com experiência comprovada na área da desinformação e sem conflito de interesses com o provedor da plataforma.

Os compromissos assumidos, num primeiro momento, ingressaram em temáticas importantes para o tratamento adequado do combate à desinformação, cujo resultado, evidentemente, vai ser conferido na medida do comprometimento efetivo das mídias no cumprimento das metas e no trabalho conjunto com o Grupo de Trabalho presidido pela União Europeia.

4.3 RECOMENDAÇÕES PARA O APRIMORAMENTO DO COMBATE À DESINFORMAÇÃO PELAS PRINCIPAIS MÍDIAS SOCIAIS, COM FOCO NA MODERAÇÃO DE CONTEÚDO

Em relação à moderação de conteúdo podemos catalogar as recomendações em oito grupos, a saber: a) observância dos direitos humanos; b) legislação definindo os limites do direito humano da liberdade de expressão; c) combate ao comportamento não autêntico; d) definição do conceito de desinformação; e) direito à privacidade no processo regulatório; f) desmonetização de fornecedores que ampliam a desinformação; h) a importância da veiculação dos números, avisos e

possibilidade de recursos; i) capacitação do usuário, pesquisadores e comunidade de verificação de fatos,

Ainda que num primeiro momento possa parecer que alguns tópicos não tenham relação direta com a moderação de conteúdo pelas mídias sociais, é inegável que as medidas indiretas são facilitadoras da atividade das plataformas. É fato que a regulação não acompanha a velocidade da propagação da desinformação através das mídias sociais, sendo que a moderação de conteúdo pelas plataformas é um importante mecanismo de combate, contudo, não pode ser o único na medida em que um assunto que impacta diretamente na vida em sociedade.

Em primeiro lugar, por todos os motivos já elencados, os direitos humanos devem estar no centro de qualquer discussão sobre a moderação de conteúdo, bem como sobre a regulamentação, como forma de estabelecer balizas para salvaguardar os indivíduos/usuários da atuação estatal e das mídias, com foco nos direitos à liberdade de pensamento e opinião, à privacidade, à liberdade de expressão e a direito de voto e de participar nos assuntos públicos (Jones, 2019, p. 52).

Qualquer iniciativa dos Estados, em termos de regulação, deve colocar os princípios de proteção dos seus cidadãos contra o abuso dos direitos humanos por parte das empresas, pois as iniciativas voluntárias não podem competir adequadamente com o modelo de negócios que favorecem conteúdos divisivos. A discussão deve estar pautada no PIDCP e na DUDH, bem como nos trabalhos e relatórios da ONU e de organismos internacionais, tais como o Código de Conduta Reforçado Sobre Desinformação da União Europeia. Os princípios de Santa Clara 2.0 trouxeram de forma muito clara que em todas as fases do devido processo de moderação de conteúdo, as plataformas de mídias sociais devem garantir a aplicação dos direitos humanos (Electronic Frontier Foundation *et al.*, 2018). De igual modo, a *Cristchurch Call*, além de prever a transparência na construção dos padrões das comunidades e termos de uso, também menciona que esses devem estar vinculados diretamente com os direitos humanos (França; Nova Zelândia, 2019). Ainda, a Chamada de Paris garante que o direito internacional, incluindo a Carta das Nações Unidas, o direito internacional humanitário e o direito internacional consuetudinário devem ser aplicáveis ao uso de tecnologias de informação e comunicação (TIC) pelos Estados (Paris, 2018).

Os Estados devem editar leis em sentido estrito, que garantam e regulamentem a liberdade de expressão, levando em consideração os *standarts* de direitos humanos, com restrições claras e amparadas no PIDCP e na DUDH, segundo os critérios do interesse legítimo tutelado e da razoabilidade e proporcionalidade. Sobre a liberdade de expressão, a internet deve ser aberta e sem censura. A orientação estatal e da jurisprudência no combate à desinformação deve estar pautada nos 19 e 20 do PIDCP com as restrições sendo aplicadas de forma proporcional para a garantia aos direitos ou a reputação de outras pessoas, a proteção da segurança nacional, da ordem, da saúde ou da moral públicas, a proibição de propaganda de guerra, da defesa do ódio nacional, racial, religioso, que constitua incitação à discriminação, hostilidade ou violência. O princípio mais importante a ser observado aqui é o da “neutralidade da regulação”. Isso significa que um discurso só deve ser proibido quando representar um perigo para outras pessoas e instituições (Bento, 2016, p. 103).

As regras para a moderação de conteúdo devem proibir a remoção do discurso político protegido pela liberdade de expressão e, ao mesmo tempo, exigir a remoção do discurso inconsistente, como o incitamento à violência. As plataformas, de sua vez, devem estabelecer quadros que permitam uma tomada de decisão eficiente, justa e específica ao contexto, refletindo os padrões da legislação em matéria de direitos humanos. As plataformas maiores, devem estabelecer mecanismos de escrutínio imparciais para supervisionar e revisar a sua tomada de decisões, a exemplo do *Facebook* que criou uma instância de revisão. As empresas de mídias sociais devem apresentar toda a transparência possível, com informações claras sobre as suas regras, fornecendo ainda dados sobre o gerenciamento do conteúdo e sobre as decisões de remoção (Jones, 2019, p. 58).

De igual modo, as plataformas, na sua missão subsidiária de moderar o conteúdo, devem ter a responsabilidade de respeitar os *standards* de direitos humanos. Restrições a esses direitos devem possuir como foco a diferenciação entre influência legítima e ilegítima no discurso *online*.

Na linha do Código de Conduta Reforçado sobre Desinformação da União Europeia, Jones (2019, p. 54) defende ainda uma participação dos Estados para buscar mudanças estruturais nas plataformas digitais como forma de proteção contra a manipulação inconsciente ou involuntária do pensamento, com foco na

transparência da utilização de técnicas persuasivas. Importantíssimo o combate às atividades incompatíveis com a democracia e com o direito à liberdade de opinião e pensamento, bem como de votar e participar do debate político, sendo importante medidas de combate ao comportamento não autêntico através de *bots*, *trolls* e algoritmos que priorizam a desinformação e direcionamento com o propósito de manipular o comportamento dos eleitores e usuários.

Quanto às campanhas de desinformação, é urgente que os Estados considerem uma visão sistemática e de técnicas que aumentem a manipulação, tais como o uso de *bots, cyborgs e trolls*. As plataformas digitais precisaram ser mais transparentes quanto aos seus objetivos e quanto ao uso de algoritmos, de forma a ajustá-los para diminuir a amplificação da desinformação. Ainda, é importante a rotulagem do material político, com informação de quem produziu, quem pagou e quanto pagou (Jones, 2019, p. 55). A regulação deve ter como foco principal esse comportamento não autêntico, pois a moderação de conteúdo apenas será efetiva quando as mídias forem capazes, se já não o são, de remover a utilização de expedientes tecnológicos que potencializem não apenas a desinformação, mas todo o discurso de ódio e discriminatório que incite à violência e fomente o ataque à democracia.

Wardle e Derakhshan (2017, p. 80) mencionam que é necessário reprimir a chamada amplificação computacional, tomando medidas eficazes e rápidas em face das contas automatizadas para impulsionar o conteúdo. Trata-se da garantia à integridade dos serviços, com o dever de limitar os comportamentos e práticas manipulativas inadmissíveis em seus serviços, devendo tais comportamentos e práticas serem revistos periodicamente à luz das últimas evidências sobre as condutas, incluindo a criação e uso de contas falsas, aquisições de contas e amplificação orientada por robôs, operações de *hack-and-leak* (hackeamento e vazamento), condutas destinadas a ampliar artificialmente o alcance, dentre outros artifícios (União Europeia, 2022). Os esforços devem ser conjuntos e integrados entre todas as plataformas, pois um robô, uma conta falsa ou anônima não pode invocar o direito à liberdade de expressão.

Ainda, é de vital importância que as mídias sociais assumam compromissos globais como o fizeram com a União Europeia. Nesse contexto, é estratégica a adoção da definição do conceito de desinformação com base nas definições mais consensuais

de *disinformation, misinformation e malinformation*, pois parece muito claro que o ponto de partida para o combate à desordem informacional e a moderação de conteúdo, passa pela aceitação, pela comunidade internacional, de um conceito padrão daquilo que se combate.

No que se refere ao direito à privacidade, a regulação deve levar em conta a coleta e utilização dos dados pessoais, o petróleo da modernidade, devendo tais questões estarem inseridas no processo regulatório, incluindo a utilização dos algoritmos. Aqui também deve ser ressaltada a necessidade de transparência, aliada ao reforço do consentimento, as chamadas leis de proteção de dados que devem garantir, por exemplo, que os usuários possam verificar facilmente, num clique, quais dados de perfil e outras informações sobre eles estão sendo mantidos e compartilhados, por quem e por qual valor. De nada adianta a moderação sem transparência, sem a incursão no modelo de negócio das mídias e sem a proteção de dados (Jones, 2019, p. 56).

A exemplo do Código Reforçado da União Europeia, a desmonetização de fornecedores que ampliam a desinformação irá auxiliar a moderação de conteúdo e o combate à desinformação, mas somente será efetivo com esforço conjunto das mídias nessa tarefa. A moderação, sem a incursão no modelo de negócios das mídias será insuficiente, de forma que uma das medidas seria diminuir os recursos financeiros de quem dissemina a desinformação. Por isso é importante que as empresas de tecnologia, assim com as redes de publicidade busquem mecanismos de evitar que os fornecedores de desinformação obtenham ganhos financeiros (Wardle; Derakhshan, 2017, p. 80).

Especificamente sobre a moderação de conteúdos, o documento que mais diretamente tratou do tema foram os Princípios de Santa Clara, que preveem já na primeira geração (princípios 1.0), a necessidade de transparência, de apresentação do número total de postagens removidas, bem como das contas suspensas. Ainda, deve avisar cada usuário cujo conteúdo for removido ou suspenso, e por fim, as empresas devem garantir oportunidades para recursos e revisões, o trinômio: número, aviso e recurso. Nesse contexto, evidentemente, as empresas devem publicar as suas regras e políticas de forma clara e precisa, orientando e informando os usuários, em local de fácil acesso. Trata-se de importante providência, prevista nos Princípios de

Santa Clara 2.0, que visa garantir uma previsibilidade do comportamento do usuário nas plataformas de mídias sociais (Electronic Frontier Foundation *et al.*, 2018).

O primeiro princípio de Santa Clara, referente ao número, consistente na necessidade de veiculação do número total de postagens removidas e de contas suspensas, de forma permanente ou temporária, em razão às violações de suas diretrizes de conteúdo. Devem constar em relatório o número total de postagens sinalizadas, removidas e contas suspensas. Quanto ao aviso, estatui que as plataformas devem avisar cada usuário cujo conteúdo for removido ou que tiver sua conta suspensa acerca das razões para a remoção ou suspensão. Trata-se de importante mecanismo que viabilizará a manifestação do usuário quanto à sua defesa. Nesse aspecto, as empresas devem fornecer orientações detalhadas à comunidade sobre quais conteúdos são proibidos, com exemplificação. É importante destacar no aviso, as diretrizes usadas pelos revisores, bem como fornecer uma explicação de como eventual detecção automatizada é utilizada (Electronic Frontier Foundation *et al.*, 2018).

Ainda quanto ao aviso, devem ser fornecidas as seguintes informações: URL (*Uniform Resource Locator*), trecho de conteúdo e/ou outras informações suficientes para permitir a identificação do conteúdo removido, a cláusula específica das diretrizes em que incidiu o conteúdo violado e como o conteúdo foi detectado e removido, além da explicação quanto aos procedimentos pelo qual o usuário poderá elaborar o recurso da decisão. Por fim, quanto ao recurso, evidentemente, as empresas devem fornecer ampla possibilidade de recursos acerca de eventual remoção de conteúdo ou suspensão de conta. A possibilidade de recursos deve envolver a revisão humana por pessoas não envolvidas na primeira decisão, oportunidade apresentar informações adicionais e a notificação do resultado da decisão recorrida (Electronic Frontier Foundation *et al.*, 2018).

De vital importância, para que se garanta uma eficiente moderação de conteúdo são os esforços de capacitação do usuário, pesquisadores e comunidade de verificação de fatos, aliados à práticas efetivas de alfabetização digital e do pensamento crítico, facilitando a identificação da desinformação garantindo a funcionalidade para sinalizar informações falsas ou enganosas (União Europeia, 2022). Indo mais além, devem as plataformas de mídias sociais fornecer os dados necessários aos pesquisadores, com o fim de abordar adequadamente a desordem

de informação e avaliar formas de melhorar a integridade dos dados (Wardle; Derakhshan, 2017, p. 80).

Na área da capacitação da comunidade de pesquisa é importante o fornecimento de dados, sendo importante que as empresas de tecnologia se comprometam com a realização de pesquisas, compilando e compartilhando os dados e resultados (União Europeia, 2022). Wardle e Derakhshan (2017, p. 80) vão mais além, ao recomendar que sejam exibidos automaticamente informações contextuais e metadados como forma de auxiliar os usuários a verificar a autenticidade de um conteúdo, citando como exemplo quando um site foi registrado ou demonstrando se a imagem não é antiga em relação ao fato, com marca de verificação (identificação visual) comum para esses indicadores em todas as redes.

Aliado a esse ponto, evidentemente o usuário deve ser avisado (informação) de que sua conta foi sinalizada, garantindo-lhe a possibilidade de recurso. Nesse ponto, importante que os serviços de mensagem promovam medidas similares dentro das suas especificidades, seja com a criação de recursos e iniciativas para capacitar os usuários, assim como busquem a limitação da veiculação viral da desinformação (União Europeia, 2022).

As plataformas de mídias sociais devem fornecer informações claras e significativas sobre os parâmetros de seus algoritmos, permitindo que os usuários recebam múltiplas possibilidade de visualizações, aumentando as suas escolhas (ONU, 2021b, p. 19). E mais, devem fornecer critérios transparentes quando ocorrer qualquer alteração algorítmica que rebaixem a classificação de um conteúdo (Wardle; Derakhshan, 2017, p. 80).

Deve ser reforçado, portanto, que a moderação de conteúdo é uma das armas à disposição do combate à desinformação e isoladamente não será efetiva, contudo, desde que prescrita no direito internacional, nas iniciativas da sociedade civil e tendo como aliadas outras técnicas importantes, poderá minimizar os impactos danosos às pessoas e à democracia.

5 CONCLUSÃO

A revolução da *internet* trouxe inegáveis impactos na vida cotidiana e com a característica de não ser restrita à indústria como as revoluções anteriores, pois as novidades trazidas refletem em muitas áreas do conhecimento. A ascensão da telefonia móvel, no final da década de 1990, popularizou e democratizou ainda mais a *internet*. São várias as visões do itinerário histórico, mas o fato é que os autores convergem para a experimentação de um momento de enormes transformações, o que culminou com a sociedade informacional abordada por Castells (2002) já no início da década de 2000.

Uma das consequências é o aumento vertiginoso do volume de informação circulando no planeta, com o avanço tecnológico revolucionando a comunicação social, fazendo com que não só as empresas, mas também as pessoas estejam conectadas em rede, com uma tendência de agrupamento segundo ideais e aptidões, com a busca incessante pela confirmação. O cenário é favorável à disseminação da desinformação, com o componente de que vários estudos comprovam que as notícias negativas e sensacionalistas ganham mais adeptos, além é claro do fato de que o ser humano sente-se mais confortável com notícias que confirmem o seu posicionamento.

A facilidade de disseminação da desinformação é potencializada pela utilização dos algoritmos, espécie de fórmula matemática que permite com que os usuários permaneçam nas chamadas “bolhas”, que são coletivos de afinidades entre determinados indivíduos. Outro componente é o modelo de negócios das mídias sociais que é o de vender publicidade através da utilização da ação dos algoritmos, que estabelecem um direcionamento da veiculação a partir dos dados dos usuários. Essa, portanto, a fórmula para manter o “usuário” numa relação de dependência, recebendo anúncios e permanecendo o máximo de tempo conectado na rede.

Ainda no cenário da comunicação social, percebeu-se claramente o declínio das mídias tradicionais, pois cada vez mais as plataformas de mídias sociais ganham campo como fornecedores de notícias. O que ocorreu, em verdade, foi o fato de que os usuários, ao mesmo tempo em que recebem a notícia, passara a ser a fonte delas, muitas vezes com a divulgação, com o compartilhamento e com e edição da informação. Assim, apesar de que a mentira não foi inventada nos tempos atuais, cotidianamente a sua difusão é enorme, a ponto de influenciar eleições, campanhas

de saúde e de transformar o termo pós-verdade na palavra do ano em 2016, de acordo com os Dicionários *Oxford.* A verdade foi relativizada e pode-se constatar que o termo *Fake News* (notícias falsas) não conseguiu mais abarcar a infinidade de elementos da desordem informacional, vez que restrita à produção de notícias. Diante da complexidade do tema, autores como Wardle e Derakhshan (2017, p. 5) introduziram três novos conceitos que caracterizariam a desinformação. O primeiro seria a *misinformation,* que compreende as informações falsas compartilhadas sem a intenção de causar dano. O segundo, a *disinformation*, que por sua vez refere-se às informações falsas compartilhadas com a intenção de causar dano. Por último, a *malinformation*, que é o termo empregado para definir as informações verdadeiras ou baseadas na realidade, mas indevidamente tornadas públicas, com a intenção de causar dano.

As plataformas de mídias sociais são empresas privadas, regidas pelo princípio da livre iniciativa. Contudo, por todos os motivos elencados, fica evidente que a sua atuação possui forte impacto social e econômico, de forma que surgiram questionamentos sobre o seu grau de responsabilidade pelas postagens e publicações dos usuários. A discussão remonta a década de 1990, em dois casos emblemáticos julgados nos Estados Unidos, estabelecendo uma distinção entre um editor (plataforma que opta por moderar conteúdo) e um mero distribuidor. No caso em que a plataforma optasse por moderar o conteúdo, poderia ensejar a responsabilidade, de forma que as mídias foram se desenvolvendo como meras distribuidoras. Nesse contexto, a Seção 230, da *Communications Decency Act*, de 1996 pacificou o cenário ao estabelecer a não responsabilização do editor.

As plataformas então passaram a moderar o conteúdo através de suas próprias regras, conhecidas como Termos de Uso, *guidelines* ou padrões da comunidade. A moderação de conteúdo é importante mecanismo de combate à desinformação e outras atividades nocivas dos usuários, como o discurso de ódio, incitação à violência, importunação sexual, dentre outros. No entanto, o mundo todo começou a discutir essas regras, surgindo diversas iniciativas, tanto governamentais, como da sociedade civil organizada, conhecidas como cartas de princípios.

Um componente que é motivo de reflexão é o fato de que as plataformas de mídias sociais possuem atuação em diversos países, cada qual com sua legislação, soberania e cultura, o que traz a necessidade de definição de regras

padronizadas, a partir da escolha de um direito universal capaz de ser seguido linearmente, tanto pelo Estado que opte por regulamentar a atuação das mídias, como pelas empresas do setor. Com o natural tensionamento entre a moderação de conteúdo e o direito à liberdade de expressão, aliado a necessidade de estabelecer regras lineares de convivência, os *standards* de direitos humanos passaram a inspirar a autorregulação das plataformas, sendo a discussão pautada não apenas pela responsabilidade quanto ao conteúdo, mas partindo da necessidade inicial de se proteger os direitos e liberdades individuais.

O direito de ter opiniões sem interferência é absoluto, diferentemente do direito à liberdade de expressão. Assim, a análise deve ser feita sob dois enfoques, tanto no que se refere à liberdade do indivíduo, quanto ao direito de não ter a opinião deliberadamente afetada pelo comportamento não autêntico. De igual modo, há uma preocupação especial no que se refere ao uso dos dados dos usuários sem o consentimento e informação, sobretudo porque é evidente que é através destes que os algoritmos agem, gerando receita de publicidade para as redes. A ideia de que o uso das plataformas é gratuito faz com que ocorram abusos nos usos de dados, que para muitos é considerado o petróleo da atualidade.

Mas o tensionamento maior entre o combate à desinformação é o direito humano à liberdade de expressão, ponto nodal e garantidor das instituições e demais liberdades, tais como o direito de reunião, associação, liberdade religiosa e de participação na política. Assim, as restrições ao direito à liberdade de expressão, pelos *standards* de direitos humanos, devem estar previstas em lei em sentido estrito, de forma necessária para combater um fim legítimo e proporcional ao mal supostamente causado. As restrições podem ocorrer apenas para garantir direitos de terceiros, a segurança nacional, a ordem pública, a saúde e moral públicas, o combate à propaganda de guerra, da defesa do ódio nacional, racial, religioso, que constitua incitação à discriminação, hostilidade ou violência.

Dentre as iniciativas, podemos destacar Relatório da ONU de 2021 e o da União Européia de 2022, vez que através deles houve uma linha bem definida a respeito do combate à desinformação, sobretudo pelos compromissos assumidos pelas empresas de mídias sociais com a União Europeia em relação à transparência e à atuação em relação ao seu modelo de negócios. Assim, tanto em relação aos dois relatórios, como tendo por base as demais iniciativas, em especial as cartas de

princípios, se concluiu pelo grupo de recomendações a seguir descritos: a) observância dos direitos humanos; b) legislação definindo os limites do direito humano da liberdade de expressão; c) combate ao comportamento não autêntico; d) definição do conceito de desinformação; e) direito à privacidade no processo regulatório; f) desmonetização de fornecedores que ampliam a desinformação; h) a importância da veiculação dos números, avisos e possibilidade de recursos; i) capacitação do usuário, pesquisadores e comunidade de verificação de fatos.

Como dito acima, a moderação de conteúdo é importante medida para o combate à desinformação, mas isoladamente pouco poderá fazer, sendo que as recomendações acima estão elencadas justamente para subsidiar, algumas direta ou indiretamente, a forma de atuação das plataformas. Neste passo, inegável a importância dos direitos humanos para definir o posicionamento das plataformas e das iniciativas governamentais em eventual regulação, como forma de colocar os direitos individuais em primeiro plano. Por isso, importante que os Estados editem leis em sentido estrito regulamentando o direito humano à liberdade de expressão, com base nos *standards* de direitos humanos, sob os critérios de razoabilidade e proporcionalidade.

É importante, também, que os Estados considerem as técnicas de aumento da manipulação da informação, chamados comportamentos não autênticos, tais como o uso de *bots, cyborgs e trolls*. Nesse ponto, é necessário cobrar transparência das plataformas quanto ao uso de algoritmos.

Também é importante a adoção de um conceito padrão de desinformação, com base nas definições mais consensuais de *disinformation, misinformation e malinformation*, numa demonstração de aceitação de uma linearidade de pensamento sobre o enfoque conceitual. Os Estados devem também implementar de forma eficaz as chamadas leis de proteção de dados, com mais transparência e reforço ao consentimento dos usuários, de forma que ele possa verificar com facilidade quais dados do seu perfil são expostos e compartilhados, e decidir sobre isso. Na mesma linha, muitos estão lucrando com a desinformação, sendo imperioso para o auxílio da moderação de conteúdo e do combate à desordem informacional, que ocorra a desmonetização de fornecedores que amplifiquem o conteúdo fraudulento. O olhar deve sempre ter como foco o modelo de negócios, como dito em mais de uma oportunidade.

Diretamente na moderação de conteúdo, as mídias devem pautar sua atuação com transparência, apresentando os números das postagens removidas ou suspensas e, ainda, avisar adequadamente os usuários, dando-lhe oportunidade de recursos com revisão humana. A moderação, ainda, deve partir de regras claras e de fácil acesso ao usuário.

Os esforços devem vir no sentido de capacitar os usuários, pesquisadores e comunidade de verificação de fatos, com a promoção de práticas efetivas de alfabetização digital e do pensamento crítico. Para isso, as plataformas de mídias sociais fornecer os dados necessários aos pesquisadores, com o fim de abordar adequadamente a desinformação.

Por fim, é certo que estamos vivenciando um momento disruptivo, de alto impacto tecnológico, com modificação radical na forma de comunicação, sendo que o fenômeno da desinformação é complexo e uma grave ameaça à convivência pacífica e à própria democracia, de forma que o combate à desordem informacional deve vir pautado em diversas frentes, na medida proporcional e necessária para garantir os direitos e liberdades individuais.

REFERÊNCIAS

AIETA, Vânia Siciliano. O impacto eleitoral resultante da manipulação das notícias falsas no universo das redes sociais: a construção da desinformação. **Revista Interdisciplinar de Direito,** Valença, v. 18, n. 1, p. 213-233, jan./jun. 2020.

ALEMANHA. Bundesministerium der Justiz. **NetzDG:** Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken (Netzwerkdurchsetzungsgesetz - NetzDG). 2017. Disponível em: https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html. Acesso em: 21 nov. 2023.

ANG, Benjamin; ANWAR, Nur Diyanah; JAYAKUMAR, Shashi. Disinformation & Fake News: meanings, present, future*. In:* ANG, Benjamin; ANWAR, Nur Diyanah; JAYAKUMAR, Shashi. **Disinformation and Fake News:** meanings, present, future. Singapure: Palgrave Macmillan, 2021. p. 3-20. Disponível em: https://doi.org/10.1007/978-981-15-5876-4. Acesso em: 20 nov. 2023.

ASWAD, Evelyn. Losing the freedom to be human. **Columbia Human Rights Law Review**, Columbia, v. 52, p. 306-371, jul. 2020. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635701. Acesso em: 11 nov. 2023.

BARBOSA, Mafalda Miranda. Fake News e Fact-Checkers: uma perspectiva jurídico-civilista. **Revista de Direito e Responsabilidade,** *[s.l.],* ano 3, p. 733-766, 2021.

BENTO, Leonardo Valles. Parâmetros internacionais do direito à liberdade de expressão. **Revista de Informação Legislativa**, Brasília, a. 53, n. 210, p. 93-115, abr./jun. 2016.

BICKERT, Monika. Charting a way forward on online content regulation. **Meta**, [*s.l.*], fev. 2020. Disponível em: https://about.fb.com/news/2020/02/online-content-regulation/. Acesso em: 16 ago. 2022.

BRASIL. **Lei nº 12.965, de 23 de abril de 2014**. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Brasília, 23 abr. 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 11 nov. 2023.

BUHMANN, Karin; OLIVERA, Roxana. Human rights and social media platforms: the corporate responsibility to respect human rights in regard to privacy infringements involving photo posting. **Australian Journal of Human Rights**, [*s.l.*], v. 26, n. 1, p. 124-141, 2020.

CANADÁ. **Paris call for trust and security in cyberspace**. 2020. Disponível em: https://www.canada.ca/en/democratic-institutions/services/paris-call-trust-security-cyberspace.html. Acesso em: 23 maio 2022.

CASTELLS, Manuel. **A sociedade em rede**. 6.ed. rev. e ampl. São Paulo: Paz e Terra, 2002. V. 1.

CENTER FOR AMERICAN PROGRESS *et al*. **Change the terms**: reducing hate online. 2021. Disponível em: https://www.changetheterms.org/. Acesso em: 14 ago. 2023.

CHAOUCH, Sana. **La communication RSE sur les réseaux sociaux: comment crédibiliser son message? Cas Le Petit Marseillais**. 2022. Dissertação (Mestrado em Sciences de Gestion) – Université Catholique de Louvain, Louvain, 2022. Disponível em: http://hdl.handle.net/2078.1/thesis:36448. Acesso em: 21 nov. 2023.

CHIRWA, Candice; MANYANA, Zimkhitha. The rise of fake news: surveying the effects of social media on informed democracy. **The Thinker**, [*s.l.*], v. 88, p. 59-65, 2021. Disponível em: https://journals.uj.ac.za/index.php/The_Thinker/article/view/604/378. Acesso em: 19 out. 2023.

COMITÊ GESTOR DA INTERNET NO BRASIL. **Resolução CGI.br/RES/2009/003/P**. São Paulo, 2009. Disponível em: https://www.cgi.br/resolucoes/documento/2009/003/. Acesso em: 14 mar. 2023.

D'ANCONA, Matthew. **Pós-verdade, a nova guerra contra os fatos em tempos de fake news.** São Paulo: Faro, 2018.

DATAREPORTAL. **All the numbers you need**. Kepios, 2023. Disponível em: https:// https://datareportal.com/. Acesso em: 03 jun. 2023.

DUNKER, Christian. Subjetividade em tempos de pós-verdade. *In:* DUNKER, Christian; TEZZA, Cristóvão; FUKS, Julián; TIBURI, Marcia; SAFATLE, Vladimir. **Ética e pós-verdade**. São Paulo: Brasiliense, 2017. p. 10-45.

ELECTRONIC FRONTIER FOUNDATION *et al*. **Manila Principles on Intermediar Liability**. [*S.l.*], 2015. Disponível em: https://manilaprinciples.org/pt-br.html. Acesso em: 14 ago. 2022.

ELECTRONIC FRONTIER FOUNDATION *et al.* **The Santa Clara Principles on Transparency and Accountability in Content Moderation**. Washington D.C., maio 2018. Disponível em: https://santaclaraprinciples.org/scp1/. Acesso em: 11 nov. 2023.

EMPOLI, Giuliano da. **Os engenheiros do caos**. São Paulo: Vestígio, 2019.

ESTADOS UNIDOS. Legal Information Institute. **U.S. Constitution**: First Amendment. Washington D.C., 1791. Disponível em https://www.law.cornell.edu/constitution/first_amendment. Acesso em: 16 jun. 2023.

ESTARQUE, Marina; ARCHEGAS, João Victor. **Redes Sociais e Moderação de Conteúdo:** criando regras para o debate público a partir da esfera privada. Rio de Janeiro: Instituto de Tecnologia e Sociedade do Rio – ITS, 2021. Disponível em: https://itsrio.org/wp-content/uploads/2021/04/Relatorio_RedesSociaisModeracaoDeConteudo.pdf. Acesso em: 14 maio 2023.

FACEBOOK. **Facebook submission to UN Special Rapporteur on Freedom of Opinion and Expression for Report on Disinformation**. [*S.l.*], 2021. Disponível em: https://www.ohchr.org/sites/default/files/Documents/Issues/Expression/disinformation/4-Companies/Facebook.pdf. Acesso em: 16 ago. 2022.

FLORIDI, Luciano. **The Fourth Revolution**. Reino Unido: Oxford University Press, 2014.

FLORIDI, Luciano**. The Onlife Manifesto:** being human in a hyperconnected era. Reino Unido: Spring Open, 2015.

FORNASIER, Mateus de Oliveira; BECK, Cesar. Cambridge analytica: escândalo, legado e possíveis futuros para a democracia. **Revista Direito em Debate**, Ijuí, v. 29, n. 53, p. 182-195, jan./jun. 2020. Disponível em: https://doi.org/10.21527/2176-6622.2020.53.182-195. Acesso em: 14 ago. 2022.

FRANÇA. Numérique. **Creating a French Framework to make social media platforms more accountable:** acting in France with a European vision. Paris: French Secretary of State for Digital Affairs, 2019. Disponível em: https://thecre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf Acesso em: 14 ago.

FRANÇA; NOVA ZELÂNDIA. **Cristchurch call to eliminate terrorista & violent extremist contente online**. 2019. Disponível em: https://www.christchurchcall.com/. Acesso em: 14 ago. 2023.

FUKUYAMA, Mayumi. Society 5.0: Aiming for a New Human-Centered Society. **Japan SPOTLIGHT**, [*s.l.*], p. 47-50, jul./ago. 2018.

GIL, Antônio Carlos. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.

GONZÁLES-IZA, Daniela. Human rights in the digital Era: challenges and opportunities from the United Nations Human Rights System. **Jurnal Pengajian Umum Asia Tenggara,** *[s.l.],* v. 22, p. 1-13, 2021.

GOOGLE. Direitos Humanos. **Google Inc.**, [*s.l.*], 2020. Disponível em: https://about.google/intl/pt-BR/human-rights/. Acesso em: 16 out. 2022.

GORWA, Robert. The platform governance triangle: conceptualizing the informal regulation of online content. **Internet Policy Review,** [*s.l.*], v. 8, n. 2, jun./2019. Disponível em: https://policyreview.info/articles/analysis/platform-governance-triangle-conceptualising-informal-regulation-online-content. Acesso em: 25 jun. 2023

HARARI, Yuval Noah. **Homo Deus**: uma breve história do amanhã. São Paulo: Companhia das Letras, 2016.

HARAYAMA, Yuko. Society 5.0: Aiming for a new human-centered society: Japan's science and technology policies for addressing global social challenges. **Hitachi Review**, [*s.l.*], v. 66, n. 6, p. 8-13, ago. 2017. Disponível em: http://www.hitachi.com/rev/archive/2017/r2017_06/pdf/p08-13_TRENDS.pdf. Acesso em: 18 out. 2022.

JONES, Kate. Online disinformation and political discourse: applying a human rights framework**. Chatham House**, Londres, nov. 2019. Disponível em: https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf. Acesso em: 30 abr. 2023.

JØRGENSEN, Rikke Frank. A human rights-based approach to social media platforms. **Berkley Center for Religion, Peace & World Affairs**, Washington, fev. 2021. Disponível em: https://berkleycenter.georgetown.edu/responses/a-human-rights-based-approach-to-social-media-platforms. Acesso em: 30 abr. 2023.

KAKUTANI, Michiko. **A morte da verdade:** notas sobre a mentira na era Trump. 1 ed. Rio de Janeiro: Intrínseca, 2018.

KUMAR, Srijan; SHAH, Neil. False Information on web and social media: a survey. **ArXiv**, [*s.l.*], v. 1, n. 1, p. 1-35, abr. 2018. Disponível em: https://arxiv.org/pdf/1804.08559.pdf. Acesso em: 30 abr. 2023.

LAZER, David M. J. *et al*. The science of fake news: addressing fake news requires a multidisciplinary effort. **Science**, [*s.l.*], v. 359, n. 6380, p. 1094-1096, mar. 2018. Disponível em: https://www.science.org/doi/10.1126/science.aao2998. Acesso em: 25 abr. 2023.

LÉVY, Pierre. **Cibercultura**. Tradução: Carlos Irineu da Costa. São Paulo: 34, 1999. 264p.

MADAKAM, Somayya Madakam; TRIPATHI, Siddharth. Social Media/Networking: Applicattion, Technologies, Theories. **Journal of Information Systems and Technology Management – Jistem USP,** São Paulo, v. 18, p. 1-19, 2021. Disponível em: https://www.scielo.br/j/jistm/a/MxH4kbZ4rwpWSGt3KCLkfxs/?lang=en&format=pdf. Acesso em: 19 nov. 2023.

MARCONI, Maria de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica.** 6. ed. São Paulo: Atlas, 2017.

MCINTYRE, Lee C. **Post-truth.** Cambridge: MIT Press, 2018. 241 p.

META. Corporate Human Rights Policy. **Meta Inc.**, [*s.l.*], 2021. Disponível em: https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf. Acesso em: 16 ago. 2022.

MONTEIRO, Artur Pericles Lima; CRUZ, Francisco Brito; SILVEIRA, Juliana Fonteles da; VALENTE, Mariana G. **Armadilhas e caminhos na regulação da moderação de conteúdo:** diagnósticos & recomendações. São Paulo: InternetLab, 2021. 32p.

NOVA IORQUE. **H2O Law.** Stratton Oakmont, inc. v Prodigy Services Co. Supreme Court, Nassau County, New York, Trial IAS Part 34. Nova Iorque, maio 1995. Disponível em: https://h2o.law.harvard.edu/cases/4540. Acesso em: 27 fev. 2022.

NOVA IORQUE. **Justia**: US Law. Cubby, Inc. v. CompuServe Inc., 776 F. Supp. 135 (SDNY 1991). Nova Iorque, out. 1991. Disponível em: https://law.justia.com/cases/federal/district-courts/FSupp/776/135/2340509/. Acesso em: 11 jun. 2023

OEA. Organização dos Estados Americanos. **Joint declaration on freedom of expression and "fake news", disinformation and propaganda.** Washington, D.C., mar. 2017. Disponível em: https://www.osce.org/files/f/documents/6/8/302796.pdf. Acesso em: 15 maio 2022.

OLIVA, Thiago Dias; TAVARES, Victor Pavarin; VALENTE, Mariana G. **Uma solução única para toda a internet?:** riscos do debate regulatório brasileiro para a operação de plataformas de conhecimento. São Paulo: InternetLab, 2020. 23p.

ONU. Organização das Nações Unidas. **Declaração Universal dos Direitos Humanos**. Paris: Assembleia Geral das Nações Unidas, 10 de dezembro de 1948. Disponível em: https://www.ohchr.org/EN/UDHR/Pages/Language.aspx?LangID=por. Acesso em: 15 set. 2023.

ONU. Organização das Nações Unidas. **Direito Humanos e Eleições:** um manual sobre normas internacionais de direitos humanos relativas a eleições. Nova Iorque: Alto-Comissariado das Nações Unidas para os Direitos Humanos, 2021a. V. 2.

ONU. Organização das Nações Unidas. **Disinformation and freedom of opinion and expression**: report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Irene Khan. Geneva: UN, 2021b. Disponível em: https://digitallibrary.un.org/record/3925306#record-files-collapse-header. Acesso em: 16 ago. 2022.

ONU. Organização das Nações Unidas. **Pacto Internacional dos Direitos Civis e Políticos.** Nova York: Assembleia Geral das Nações Unidas, 16 dez. 1966. Disponível em: https://www.cne.pt/sites/default/files/dl/2_pacto_direitos_civis_politicos.pdf. Acesso em: 15 set. 2023.

OSÓRIO, Aline. **Direito eleitoral e liberdade de expressão.** 2 ed. Belo Horizonte: Forum, 2022.

PARIS. **Paris Call for Trust and Security in Cyberspace**. 2018. Disponível em: https://pariscall.international/en/. Acesso em: 14 ago. 2022.

PASSARELLI, Brasilina; GOMES, Ana Paula Fernandes. Transliteracias: a terceira onda informacional nas humanidades digitais. **Revista Ibero-Americana de Direito da Informação,** Brasília, v. 13, n. 1, p. 253-275, jan./abr. 2020.

PIERRE, Cristina. What is the difference between social media and social network? **Blog Medium,** [*s.l.*], 2018. Disponível em: https://medium.com/@christina_meyer/what-is-the-difference-between-social-media-and-social-network-e6dd5ad28d8f. Acesso em: 28 mar. 2023.

RAIS, D.; FALCÃO, D.; GIACCHETTA, A. Z.; MENEGUETTI, P. **Direito Eleitoral Digital.** São Paulo: Thompson Reuters Brasil, 2018.

RECUERO, Raquel. **Redes sociais na internet.** Porto Alegre: Sulina, 2009. 191p.

REINO UNIDO. **Online Harms White Paper:** full government response to the consultation. Londres: Secretary of State for Digital, Culture, Media and Sport, dez. 2020. Disponível em: https://www.gov.uk/government/consultations/online-harms-white-paper/outcome/online-harms-white-paper-full-government-response. Acesso em: 01 mar. 2022.

RUBIN, Victoria L. **Misinformation and disinformation**: detecting fakes with the eye and AI. Switzerland: Springer Nature, 2022. 305p.

RUEDIGER, Marco Aurélio; GRASSI, Amaro (Coord.). **Redes sociais nas eleições 2018**. Rio de Janeiro: FGV DAPP, 2018. Disponível em: https://bibliotecadigital.fgv.br/dspace/handle/10438/25737. Acesso em: 23 abr. 2023.

SALGUES, Bruno. **Society 5.0**: Industry of the future, technologies, methods and tools. London: Iste, 2018. V. 1.

SANDER, Barrie. Democratic disruption in the age of social media: between marketized and structural conceptions of human rights law. **The European Journal of International Law,** Oxford, v. 32, n. 1, p. 159-193, 2021.

SARLET, Ingo Wolfgang; HARTMANN, Ivar Alberto Martins. Proteção de dados e inteligência artificial: perspectivas éticas e regulatórias. **RDU**, Porto Alegre, v. 16, n. 90, p. 85-108, nov./dez. 2019.

SARLET, Ingo Wolfgang. Liberdade de expressão e o problema da regulação do discurso do ódio nas mídias sociais. **Revista Estudos Institucionais**, [*s.l.*], v. 5, n. 3, p. 1207-1233, set./dez. 2019.

SCHWAB, Klaus. **A quarta revolução industrial.** Tradução: Daniel Moreira Miranda. São Paulo: Edipro, 2016.

SCHWAB, Klaus; DAVIS, Nicholas Davis. **Aplicando a quarta revolução industrial.** Tradução: Daniel Moreira Miranda. São Paulo: Edipro, 2019.

SIEBERT, Silvânia; PEREIRA, Israel Vieira. A pós-verdade como acontecimento discursivo. **Linguagem em (Dis)curso,** Tubarão, v. 20, n. 2, p. 239-249, maio/ago. 2020. Disponível em: https://www.scielo.br/j/ld/a/vykt83t8h8874gJT7ys46sy/?lang=pt&format=pdf. Acesso em: 29 abr. 2023.

SILVA, Fernanda dos Santos Rodrigues; GERTRUDES, Júlia Maria Caldeira. **Governança da moderação de conteúdo online:** percepções sobre o papel dos atores e regimes. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2023. Disponível em: https://bit.ly/3WHMUIg. Acesso em: 25 jun. 2023.

STARTS, Janis. Disinformation as a threat to national security. *In:* ANG, Benjamin; ANWAR, Nur Diyanah; JAYAKUMAR, Shashi. **Disinformation and Fake News:** meanings, present, future. Singapure: Palgrave Macmillan, 2021. p. 23-33. Disponível em: https://doi.org/10.1007/978-981-15-5876-4. Acesso em: 20 nov. 2023.

TANDOC JUNIOR, Edson C. Tools of disinformation: how fake news gets to deceive. *In:* ANG, Benjamin; ANWAR, Nur Diyanah; JAYAKUMAR, Shashi. **Disinformation and Fake News:** meanings, present, future. Singapure: Palgrave Macmillan, 2021. p.35-46. Disponível em: https://doi.org/10.1007/978-981-15-5876-4. Acesso em: 20 nov. 2023.

TELLES, André. **A revolução das mídias sociais:** cases, conceitos, dicas e ferramentas. São Paulo: M. Books, 2010.

TURCILO, Lejla; OBRENOVIC, Mladen. **Misinformation, disinformation, malinformation**: causes, trends, and their influence on democracy. Sarajevo: Fundação Heirich Böll, 2020. 38p.

UNIÃO EUROPEIA. European Comission. **The Strengthened Code of Practice on Disinformation 2022**. Bruxelas, 2022. Disponível em: https://ec.europa.eu/newsroom/dae/redirection/document/87585. Acesso em: 22 nov. 2023.

WARDLE, Claire; DERAKHSHAN, Hossein. **Information disorder:** toward an interdisciplinary framework for research and policymaking. Estrasburgo: Council of Europe, 2017. *E-book*. Disponível em: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html#. Acesso em: 29 abr. 2023.

WARDLE, Claire; DERAKHSHAN, Hossein. Reflexões sobre a “desordem da informação”: formatos da informação incorreta, desinformação e má-informação. *In*: IRETON, Cherilyn; POSETTI, Julie (Ed.). **Jornalismo, Fake News & Desinformação:** manual para educação e treinamento em jornalismo. Paris: UNESCO, 2019. p. 46-58.

WERTHEIN, Jorge. A sociedade da informação e seus desafios. **Revista Ciência da Informação**. Brasília, v. 29, n. 2, p. 71-77, maio/ago. 2000.

X. Platform Use Guidelines: defending and respecting the rights of people using our service. **X Corp.**, [*s.l.*], 2023. Disponível em: https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice. Acesso em: 16 ago. 2022.