REGULAÇÃO DE “FAKE NEWS” NO BRASIL

Juliano Maranhão, Ricardo Campos, Jessica Guedes,
Samuel Rodrigues de Oliveira e Maria Gabriela Grings

prefácio

REGULAÇÃO DE "FAKE NEWS" NO BRASIL

O Brasil não fica alheio aos impactos que a relação cada vez mais complexa entre sociedade, tecnologia e internet tem trazido nas sociedades contemporâneas. Por óbvio, fatos que ocorram no exterior podem ter impactos e até mesmo serem replicados no país. Neste sentido, um dos principais temas discutidos no âmbito regulatório-legislativo e social nos últimos anos são as *fake news*.

No País, a discussão legislativa vem sendo travada no âmbito do Projeto de Lei 2630/2020, que propõe a Lei de Transparência e Responsabilidade na Internet, que vem sendo chamada de Lei de "fake news". O projeto já passou por intensos debates, não só na Câmara e no Senado, mas em uma série de audiências e eventos, envolvendo diversos atores da sociedade civil, com a divulgação de diversos relatórios e notas técnicas de associações como Data Privacy Brasil[1], Internet Lab[2], Laboratório de Políticas Públicas e Internet (LAPIN)[3], Coalizão Direitos na Rede4 e Electronic Frontier Foundation (EFF)[5].

O objetivo do presente documento é trazer contribuições pontuais do Instituto Legal Grounds, aprofundando e, assim, propondo aperfeiçoamentos a dois temas que entendemos centrais nesse debate. O primeiro deles refere-se à atividade de moderação de conteúdo pelas plataformas digitais. O segundo refere-se ao modelo de autorregulação regulada, proposto originalmente no Substitutivo submetido pelo Senador Antônio Anastasia (PSDB/MG), e que foi incorporado no texto atual, mas que merece algumas complementações relevantes para que o instituto possa exercer plenamente sua função.

A seguir, traremos, na primeira Seção, sobre o panorama do texto atual e seu histórico legislativo. Na segunda Seção, abordaremos a moderação de conteúdo, aliada a uma proposta de redação de artigos pertinentes como estratégia de aperfeiçoamento do PL 2630. Ainda, trataremos do Instituto de Autorregulação Regulada, fazendo, ao fim, sugestões de redação ao PL capazes de traduzir as teses aqui defendidas.

Sumário

1

Histórico Legislativo no Brasil

1.1. Alguns conceitos básicos sobre "fake news" e o foco da discussão legislativa

Apesar da divulgação de informações falsas não ser algo novo, especialmente no contexto político, o tema começou a ser discutido de forma mais profunda após a eleição de Donald Trump nos Estados Unidos em 2016, tanto que, no ano seguinte, o termo foi considerado a expressão do ano pela editora Collins[6].

Todavia, o termo 'fake news" é amplo e abarca um grande guarda-chuva de ações. Por isso, para tratar no âmbito de política pública e legislativa, é preciso ter cuidado com o conceito para conseguir analisar o problema de forma efetiva e, consequentemente, traçar soluções que sejam capazes de impedir e/ou reduzir os impactos do problema. Assim, estudo realizado por Claire Wardle e Hossein Derakhshan estabeleceu parâmetros teóricos sobre o assunto e três conceitos centrais[7], quais sejam:

i) Desinformação: a **informação falsa** que é criada para **prejudicar** uma pessoa, um grupo social, uma organização ou um país.

i) *Mis-information*: a informação que é **falsa**, mas que **não foi criada** com o objetivo de causar danos.

i) *Mal-information:* é a informação baseada na **realidade**, mas que é usada para **causar danos** a pessoas, organização ou países.

No atual momento, as regulações sobre o tema se centram na desinformação, que é a criação/disseminação de uma notícia falsa com o objetivo de atingir alguém ou algum grupo.

1.2. Primeiros debates sobre regulação

No Brasil, possivelmente o primeiro contato fático mais profundo que a sociedade teve com o tema foi nas eleições gerais de 2018, ocasião na qual a Folha de São Paulo noticiou o uso da estratégia de disseminação de *fake news* por parte de um dos candidatos à Presidência da República[8].

Todo esse cenário conduziu os intuitos de regular o tema no país. O parecer do Conselho de Comunicação Social n. 01/2018[9] (PCS 1/2018) é um importante documento acerca do histórico de regulação de "fake news" no país, pois, além de expor os projetos de legislação até ali vigentes, também apresentou sinalizações necessárias para direcionar a elaboração de uma legislação eficaz sobre o assunto.

O PCS 01/2018 apresenta que, até abril de 2018, existiam 14 (quatorze) projetos de lei sobre *fake news* no Congresso Nacional, sendo que, somente um deles estava no Senado Federal e os demais na Câmara dos Deputados. Interessante mencionar que somente 2 (dois) dos projetos buscam criar

legislação nova, quais sejam, o PL 7604/2017 e o PL 6812/2017, ambos de autoria do Deputado Luiz Carlos Hauly (PSDB/PR). Atualmente, esses dois PLs estão apensados ao PL 2630/2020.

Os outros 12 (doze) projetos de lei buscaram alterar o Código Penal, o Código Eleitoral, o Marco Civil da Internet e a então Lei de Segurança Nacional para acrescentar disposições que tratassem sobre as *fake news*. O PCS 01/2018 destaca que os PLs 7604/2017 e 9647/2018 eram os únicos que previam a responsabilização das plataformas e determinavam a remoção do conteúdo. Os demais limitavam a responsabilização para quem divulga, compartilha ou dissemina notícias falsas. Do rol dos 14 (catorze) PLs citados, 12 (doze) apresentaram a conceituação de *fake news* e a determinação de aplicação de sanção penal pela conduta.

O PCS 1/2018 também elenca cinco contribuições que devem ser observadas pelas legislações que pretendam regular o assunto, e, atualmente, ao menos quatro delas ainda se revelam pertinentes: notícias falsas devem ser rebatidas com mais informação, a legislação brasileira em vigor deve ser considerada, as plataformas devem ser neutras e transparentes e políticas públicas de educação para a mídia se fazem urgentes.

De 2018 para cá, muita coisa mudou acerca do debate da regulação das *fake news* no Brasil, mas destacaríamos duas. A primeira mudança é relativa ao aumento de números de projeto de lei sobre o tema. Na Câmara dos Deputados, a busca por proposição com o termo retorna 80 (oitenta) resultados de projetos de lei em tramitação sobre *fake news*. A segunda mudança decorreu do avanço no processo legislativo que agregou diferentes perspectivas consolidadas no PL 2630/2020.

1.3. Um breve panorama do Projeto de Lei 2630/2020

O PL n. 2630/2020 foi apresentado pelo Senador Alessandro Vieira (CIDADANIA/SE) em 13/05/2020 e aprovado pelo Senado Federal em 30/06/2020, o que demonstra a velocidade do trâmite do referido PL naquela casa[10].

O PL n. 2630/2020 chegou à Câmara dos Deputados em 03/07/2020 e ficou parado até abril de 2021, quando a mesa indicou a tramitação do PL pelas comissões de Ciência e Tecnologia, Comunicação e Informática (CCTCI); Finanças e Tributação e Constituição e Justiça e de Cidadania.

Assim, a primeira comissão na qual o projeto tramitou foi a CCTCI, com a relatoria do Deputado Paulo Ganime (NOVO/RJ). Porém, em 06/07/2020, foi instituído Grupo de Trabalho (GT) específico denominado "Aperfeiçoamento da legislação brasileira – internet" para estudar o PL e todos os seus apensos com o objetivo de aprimorar o projeto. O GT tem coordenação da Deputada Bruna Furlan, relatoria do Deputado Orlando Silva e a participação dos membros titulares Filipe Barros, Paulo Eduardo Martins, Silvio Costa Filho, Sóstenes Cavalcante, Gustavo Fruet, Felipe Rigoni, Lídice da Mata, Luiza Erundina, Natália Bonavides, Rui Falcão e Vinicius Poit. Atualmente, o GT ainda está em funcionamento e está organizando diversas audiências públicas sobre o tema[11].

O PL n. 2630/2020 é composto por 36 (trinta e seis) artigos e 7 (sete) capítulos e pretende instituir a "Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet". De forma geral, o PL pretende estabelecer um regime de responsabilização das redes sociais e dos serviços de mensageria privada com mais de 2 (dois) milhões de usuários registrados no Brasil.

O PL estabelece nove princípios, quais sejam: i) liberdade de expressão e de imprensa; ii) garantia dos direitos de personalidade, da dignidade, da honra e da privacidade do indivíduo; iii) respeito ao usuário em sua livre formação de preferências políticas e de uma visão de mundo pessoal; iv) responsabilidade compartilhada pela preservação de uma esfera pública livre, plural, diversa e democrática; v) garantia de confiabilidade e da integridade dos sistemas informacionais; vi) promoção do acesso ao conhecimento na condução dos assuntos de interesse público; vii) acesso amplo e universal aos meios de comunicação e à informação; viii) proteção dos consumidores; e ix) transparência nas regras para veiculação de anúncios e conteúdos pagos.

O PL também estabelece quatro objetivos: i) o fortalecimento do processo democrático por meio do combate ao comportamento inautêntico e às redes de distribuição artificial de conteúdo e do fomento ao acesso à diversidade de informações na internet no Brasil; ii) a defesa da liberdade de expressão e o impedimento da censura no ambiente online; iii) a busca por maior transparência das práticas de moderação de conteúdos postados por terceiros em redes sociais, com a garantia do contraditório e da ampla defesa; e iv) a adoção de mecanismos e ferramentas de informação sobre conteúdos impulsionados e publicitários disponibilizados para o usuário.

O art. 5º estabelece o rol de conceitos do PL n. 2630/2020 apresentando nove definições para melhor aplicação da legislação. Dentro dos conceitos ali apresentados, é necessário destacar o de rede social e o de serviço de mensageria privada por serem centrais para a lógica de construção do PL.

- Rede social é considerada como a aplicação de internet que se destina a realizar **a conexão de usuários entre si**, permitindo e tendo como centro da atividade a **comunicação**, o **compartilhamento** e a **disseminação** de conteúdo em um mesmo sistema de informação, através de **contas conectadas ou acessíveis entre si de forma articulada**.

- Serviço de mensageria privada é a aplicação de internet que viabiliza o envio de mensagens para **destinatários certos e determinados**, inclusive protegidas por criptografia de ponta a ponta, a fim de **que somente remetente e destinatário da mensagem tenham acesso ao seu conteúdo**, excluídas aquelas prioritariamente destinadas a uso corporativo e os serviços de correio eletrônico.

O art. 6º do PL estabelece que os provedores de redes sociais e os serviços de mensageria privada devem adotar medidas para vedar o funcionamento de contas inautênticas, vedar contas automatizadas não identificadas como tal e identificar todos os conteúdos impulsionados e publicitários que a distribuição tiver sido feita por pagamento. Ainda neste artigo, o PL prevê que os provedores devem adotar medidas técnicas para identificar contas de *bots*.

Sobre os cadastros de contas, o art. 7º estabelece que os provedores podem solicitar a apresentação de documento de identidade do usuário quando houver denúncia que viole a Lei, indícios de contas inautênticas ou ordem judicial. No mesmo sentido, o art. 8º prevê que os serviços de mensageria privada devem suspender contas de usuários do serviço que funcione exclusivamente por número de telefone.

Especificamente para os serviços de mensageria privada, o PL apresenta que devem ser elaboradas quatro políticas básicas: i) manter a natureza

interpessoal do serviço, ii) limitar o número de encaminhamentos de uma mesma mensagem e o número máximo de membros por grupo, iii) instituir mecanismos para aferir consentimento prévio do usuário para inclusão em grupos e correlatos; e iv) desabilitar, por padrão, o aceite para inclusão em grupos e semelhantes. Sendo que, o PL classifica o encaminhamento em massa como o envio de uma mesma mensagem por mais de 5 (cinco) usuários em um intervalo de até 15 (quinze) dias para grupos e semelhantes.

O art. 10 do PL é um dos principais pontos de tensão do projeto por tratar da chamada rastreabilidade. O citado artigo prevê que os serviços de mensageria privada devem guardar os registros de mensagens veiculadas em encaminhamento em massa pelo prazo de 3 (três) meses, sendo que, os registros devem contar com data, hora e quantitativo total de usuários que receberam a mensagem, que deve ser no mínimo 1000 (mil) usuários. Os acessos aos registros somente podem ocorrer para investigação criminal e em instrução processual penal.

Sobre a moderação de conteúdo, o art. 12 do PL prevê que em caso de dano imediato ou de difícil reparação, de segurança da informação ou do usuário, de violação a direitos de crianças e adolescentes, de ocorrência de crimes tipificados na Lei 7716/1989 (define os crimes resultantes de preconceitos de raça ou cor), ou de grave comprometimento da usabilidade, integridade ou estabilidade da aplicação, os provedores podem retirar o conteúdo sem notificar o usuário. Sendo que, nos outros casos, os provedores devem notificar o usuário informando a fundamentação, o processo de análise e de aplicação da medida, e, em todos os casos, deve garantir a possibilidade do usuário recorrer da indisponibilização de conteúdos e contas.

O PL estabelece obrigação específica para as redes sociais apresentarem relatórios trimestrais de transparência que devem conter diversas informações, dentre elas, o número total das medidas de moderação de contas e conteúdos, o número total de contas automatizadas identificadas, e os dados relacionados a engajamentos ou interações com conteúdos identificados como irregulares.

Os relatórios devem ser disponibilizados em até trinta dias do final do trimestre e, em caso de ausência de apresentação, as redes sociais devem apresentar a justificativa prévia para tanto. Igualmente, os provedores de redes sociais devem facilitar o compartilhamento dos dados com instituições de pesquisa.

Os arts. 14 a 17 do PL estabelecem regras para o impulsionamento e a publicidade nas redes sociais. Todo o conteúdo que se enquadre nessa perspectiva deve identificar a conta responsável e permitir que o usuário tenha acesso ao contato do responsável. No período eleitoral, aplicam-se regras específicas como a divulgação do valor gasto com o impulsionamento e a identificação do responsável pela contratação. Ademais, as redes sociais devem permitir que o usuário acesse o histórico de seis meses de conteúdo impulsionado que teve contrato e exigir a identificação dos contratantes.

O Capítulo III (arts. 18 – 24) versam sobre a atuação do Poder Público estabelecendo critérios de identificação dos perfis da Administração e medidas que devem ser adotadas pelos entes públicos. Todas as contas de redes sociais de entidades e órgãos da Administração Pública, direta ou indireta, e dos agentes políticos com competência proferida pela Constituição Federal (ex: deputados e senadores) são consideradas contas de interesse público que devem respeitar os princípios da Administração. Essas contas não podem impedir

que outras contas tenham acesso às suas publicações e, caso o agente político tenha mais de um perfil na mesma plataforma, ele pode indicar qual será a representante do seu mandato.

Igualmente, todos os portais da transparência devem informar dados relacionados com a contratação de serviços de publicidade e propaganda e impulsionamento na internet, como o valor do contrato e a lista de locais nas quais o recurso financeiro foi aplicado (p. ex. página e aplicativo). Neste sentido, interessante previsão do art. 20 aponta que deve ser coibida publicidade para sites e perfis que promovam atos de incitação à violência contra pessoa ou grupo.

Os órgãos públicos devem editar normas internas sobre a estratégia de comunicação social, incluindo perspectiva que permita ao público requerer a revisão ou remoção de postagens. Na mesma linha, os órgãos podem criar manual de boas práticas com recomendações para uso de servidores exclusivamente no exercício de suas funções.

Para além dos regramentos sobre o comportamento da Administração Pública nas redes sociais, o PL também prevê algumas obrigações do Poder Público, quais sejam: i) realizar campanhas educacionais para o uso consciente e responsável da internet; ii) criar formas de responsabilização por danos coletivos em caso de descumprimento do PL em conjunto com o Ministério Público e o Poder Judiciário e iii) impedir a perseguição ou prejuízo ao servidor por conteúdo compartilhado por ele em caráter privado, fora do exercício de suas funções e que não constitua material cuja publicação tenha vedação prevista em lei.

O Capítulo IV do PL trata sobre o Conselho de Transparência e Responsabilidade na Internet, que deve ser instituído pelo Congresso Nacional em até 60 (sessenta) dias de publicação da Lei, terá atribuição de realizar estudos, pareceres e recomendações sobre o tema da legislação. O PL estabelece um rol de onze competências para o Conselho, entre elas, elaborar código de conduta para redes sociais e serviços de mensageria privada a ser avaliado e aprovado pelo Congresso Nacional e avaliar a adequação das políticas de uso adotadas pelos provedores.

O Conselho contará com 21 (vinte e um) conselheiros com mandatos de 2 (dois) anos admitida uma recondução. Entre os conselheiros, as cadeiras foram divididas da seguinte forma: i) com uma cadeira: Senado Federal, Câmara dos Deputados, Conselho Nacional de Justiça, Conselho Nacional do Ministério Público, Comitê Gestor da Internet, Setor de Telecomunicações, Conselho Nacional dos Chefes de Polícia Civil, Departamento de Polícia Federal, Agência Nacional de Telecomunicações, Conselho Nacional de Autorregulamentação Publicitária; ii) com duas cadeiras: academia e comunidade técnica, provedores de acesso/aplicações/conteúdo de internet, comunicação social; iii) com cinco cadeiras: representantes da sociedade civil. Todavia, membros dos três Poderes, ocupantes de cargos públicos que sejam demissíveis *ad nutum* e pessoas vinculadas ou filiadas a partido político não podem integrar o Conselho.

O art. 30 do PL é dedicado a autorregulação regulada ditando que os provedores de redes sociais e serviços de mensageria privada podem criar instituição voltada para transparência e responsabilidade no uso da internet. O PL estabelece 6 (seis) possíveis atribuições para essas instituições, as quais destacamos incluir uma ouvidoria independente com a finalidade de receber críticas e avaliar as atividades da instituição e desenvolver boas práticas de suspensão de contas de usuário em conjunto com as empresas de telefonia móvel.

A instituição também pode encaminhar relatórios trimestrais informando as políticas de uso e de monitoramento do volume de conteúdo compartilhado pelos usuários e serviços de mensageria privada para o Conselho, assim como aprovar resoluções e súmulas para regular seu procedimento de análise. O Conselho deve certificar a instituição de autorregulação.

O art. 31 trata sobre as sanções estabelecendo que o descumprimento ao PL pode ser punido com advertência com indicação de prazo para adoção de medidas corretivas e multa de até 10% (dez por cento) do faturamento do grupo econômico no Brasil no seu último exercício, para além da possibilidade de aplicação das demais sanções civis, criminais e administrativas. O referido artigo determina que a autoridade judicial deve observar a proporcionalidade considerando a condição econômica do infrator, as consequências da infração na esfera coletiva e a reincidência, que é considerada quando o agente repetir conduta anteriormente sancionada no prazo de 6 (seis) meses.

O Capítulo VII abarca as disposições finais determinando que os provedores de redes sociais e de serviços de mensageria devem ter sede e nomear representantes legais no Brasil e manter acesso remoto dos seus bancos de dados a partir do Brasil, especialmente para o cumprimento de ordens da autoridade judicial brasileira. Os valores aplicados a título de multa devem ser destinados do FUNDEB (Fundo de Manutenção e Desenvolvimento da Educação Básica e de Valorização dos Profissionais de Educação) e devem ser empregados em ações de educação e alfabetização digitais.

2

Moderação de Conteúdo pelas Plataformas

2.1. O que é e como vem sendo feita a moderação de conteúdo

A moderação de conteúdo deve ser entendida como o conjunto de mecanismos de governança que estruturam a participação em uma plataforma para facilitar cooperação e prevenir abusos.[12] Essa atividade vem sendo largamente empregado pelos provedores de redes sociais e de aplicações que permitam compartilhamento de conteúdo na internet, bem como nos resultados de buscas.[13] Abaixo trazemos alguns dados sobre moderação pelas principais plataformas de internet:

2.1.1. YouTube

Em consonância com o postulado pelos Princípios de Santa Clara, o YouTube fornece relatórios trimestrais relativos à remoção de conteúdo da plataforma[14]. Em termos de canais removidos, foram 2.230.310 de janeiro a março de 2021, que continham 59.301.978 vídeos. De abril a junho, 4.187.640 canais foram removidos, totalizando 67.007.971 vídeos. Quando um canal é removido, todos os seus vídeos também o são. No total, 6.417.950 canais, contendo 126.309.949 vídeos, foram removidos do YouTube de janeiro a junho de 2021. Os principais motivos para remoção foram: 1) Spam, conteúdo enganoso e golpes (90,3%); 2) Nudez ou conteúdo sexual (4,3%); 3) Assédio e bullying virtual (1,9%).

Em se tratando de vídeos isolados, de janeiro a março de 2021 foram 9.569.641 vídeos removidos, contando a detecção automatizada de conteúdo; 478.326 vídeos foram manualmente removidos no período. De abril a junho, foram 6.278.771 vídeos removidos de maneira automatizada, e 351.570 manualmente. No total, foram 15.848.412 vídeos removidos por detecção automatizada e 829.896 removidos manualmente. Os principais motivos para remoção foram: 1) Segurança infantil (29,9%); 2) Nudez ou conteúdo sexual (22,4%); 3) Conteúdo violento ou explícito (16,8%). "Spam, conteúdo enganoso ou golpes" ocupa a 4ª posição, com 14,1%.

Quanto aos comentários removidos: 1.032.365.719 de janeiro a março, e 1.162.156.293 de abril a junho de 2020. No total, 2.194.522.012 comentários removidos.

De janeiro a junho de 2021, os 5 países que mais tiveram conteúdo removido foram, nesta ordem: Índia, Estados Unidos, Brasil, Indonésia e Rússia. No Brasil, 1.552.384 vídeos foram removidos, número relativamente próximo ao dos EUA, com 1.769.708. A Índia, por sua vez, teve mais vídeos removidos que Brasil e EUA juntos: 3.583.194.

2.1.2. Facebook

O relatório[15] mais recente do Facebook se refere ao segundo trimestre de 2021. O documento traz em destaque dados relativos à desinformação sobre COVID-19: mais de 20 milhões de conteúdos foram removidos do Facebook e Instagram, enquanto mais de 3 mil contas, páginas e grupos foram removidas.

Além disso, o relatório traz dados sobre demais conteúdos removidos. Mais de 31,5 milhões de conteúdos com discurso de ódio foram removidos no Facebook, ante 25,2 milhões no primeiro trimestre deste ano. Em se tratando do Instagram, foram 9,8 milhões de conteúdo removido no segundo trimestre de 2021, versus 6,3 no primeiro.

O Facebook ainda afirma ter "tomado medidas" em relação à segurança infantil. Segundo o relatório, foram tomadas medidas sobre 2,3 milhões de conteúdos de nudez infantil e abuso físico no Facebook, e 458 mil no Instagram, no segundo trimestre deste ano. Em relação a conteúdo de exploração sexual infantil, foram tomadas medidas sobre 25,7 milhões de conteúdos no Facebook e 1,4 milhões no Instagram.

2.1.3. Twitter

O último relatório de transparência[16] publicado pelo Twitter traz dados relativos ao período de 1° de janeiro a 30 de junho de 2020. Todavia, já estão disponíveis dados relativos ao período de julho a dezembro desse mesmo ano[17]. De 3.538.093 contas notificadas, 1.009.083 (28,5%) foram removidas. Quanto ao conteúdo, a plataforma indica que 4.470.600 "conteúdos" foram removidos no período.

De 1° de julho a 31 de dezembro de 2020, a plataforma removeu 3.8 milhões de *tweets* que violavam suas normas de uso. Dos tweets removidos, 77% haviam recebido menos de cem "impressões" (métrica de alcance da plataforma). 17% haviam recebido entre cem e mil impressões, ao passo que "apenas" 6% dos tweets removidos haviam recebido mais de mil impressões – o Twitter não informou qual o número máximo de impressões que as postagens removidas receberam. No total, os tweets cujo conteúdo violavam as normas da plataforma e que foram removidos correspondiam a menos de 0,1% de todas as impressões totais da plataforma no período.

Em termos percentuais, os números representam um aumento de 9% na quantidade de contas removidas em comparação ao relatório anterior, relativo a janeiro-junho de 2020. O aumento no quantitativo de conteúdo removido foi de 132% em relação ao mesmo período. Embora tenha sido divulgado o total de contas reportadas (12.2 milhões) e as razões das denúncias, o relatório não indica quantas contas foram suspensas em decorrência das denúncias.

2.2. Propostas internacionais para regulação da atividade de moderação

Pela posição central das plataformas digitais na construção do espaço de debate público atual, a tendência, na Europa[18] e nos Estados Unidos da América[19], consiste em ver a moderação como central não só para a promoção da qualidade dos serviços disponibilizados, que traduz o interesse privado dos provedores, mas também para a garantia de liberdade de expressão do indivíduo, de direitos fundamentais e do ambiente democrático, razão pela qual, há uma necessidade de convergência dos termos de uso com a ordem jurídica Estatal, em democracias constitucionais contemporâneas.

Por outro lado, cada vez mais se reconhece que essa curadoria de conteúdo implica responsabilidade das plataformas pela adoção dos procedimentos adequados de governança e meios técnicos disponíveis para evitar abusos e mitigar impactos provocados por conteúdo nocivo propagado. Ou seja, aponta-se para deveres inerentes à moderação de conteúdo pelas plataformas terem se tornado, de fato, as anfitriãs do espaço para o discurso público e "curadoras" da liberdade de expressão, de modo que suas regras não podem ficar restritas à relação contratual com os usuários, devendo incorporar a legislação nacional, sendo compatíveis com as regras sobre licitude de conteúdo difundido na esfera pública.[20] Isso inclui não só os ilícitos criminais,[21] como também as regras nacionais sobre publicidade[22] nos veículos de comunicação, nas práticas de propaganda online e impulsionamento de conteúdo.

Tendo em vista o interesse público do qual se cerca, os sistemas privados de governança montados pelos provedores para a moderação de conteúdo na internet têm sido centro de debates internacionais, que apontam preocupação quanto à falta de transparência e arbítrio pelos provedores.[23] Nesse sentido, entidades da sociedade civil têm promovido iniciativas e promulgado documentos no sentido de estimular a adoção de mecanismos de transparência, dentre eles o *Santa Clara Principles*[24], o *Corporate Accountability Index*[25] e os princípios sobre filtragem de conteúdo da *Eletronic Frontier Foundation*[26].

Esses documentos elencam a necessidade de provedores de aplicação divulgarem os números de medidas aplicadas (*flagging*, *blocking*, *takedown*) e de notificarem o usuário afetado pela medida, oferecendo as razões correspondentes, bem como a oportunidade para sua contestação. Iniciativas legislativas também incorporaram alguns deveres procedimentais da transparência, como a lei alemã de redes *NetzDG*[27] e a proposta de regulação europeia sobre prevenção de disseminação online de conteúdo terrorista[28] e a Recomendação sobre a Responsabilidade de Intermediários na Internet do Conselho Europeu.[29]

Nos itens a seguir, descreveremos algumas iniciativas relevantes sobre moderação e sua regulação, a saber, os Princípios de Manila, diretrizes da Eletronic Frontiers Foundation, os Princípios de Santa Clara, o Ranking Digital Rights e a Legislação Alemã (*NetzDG*).

2.2.1. Princípios de Manila

Os Princípios de Manila Sobre Responsabilidade dos Intermediários[30] são um conjunto de diretrizes que delineia salvaguardas a ser aplicadas em todas as regulamentações legais sobre responsabilidade de intermediários. O documento foi lançado na RightsCon, conferência realizada na capital das Filipinas no ano de 2015, por uma coalizão de ativistas de direitos na Internet e organizações da sociedade civil. O principal objetivo dos Princípios de Manila é incentivar o desenvolvimento de regimes de responsabilidade interoperáveis e harmonizados que possam promover a inovação, respeitando os direitos dos usuários, de acordo com a Declaração Universal dos Direitos Humanos, o Pacto Internacional sobre Direitos Civis e Políticos e os Princípios Orientadores das Nações Unidas sobre Negócios e Direitos Humanos.

2.2.2. Eletronic Frontier Foundation

Nesse sentido, organizações como Electronic Frontier Foundation (EFF), Article 19, Derechos Digitales, entre outras, postularam seis princípios a serem considerados por legisladores e intermediários ao desenvolver, adotar e analisar normas, políticas e práticas que tratam da responsabilidade dos intermediários por conteúdo de terceiros. São eles:

1) Os intermediários devem ser protegidos por lei da responsabilização por conteúdos produzidos por terceiros;
2) Não se deve solicitar a remoção de conteúdos sem a ordem de uma autoridade judicial;
3) Requisições de restrição de conteúdos devem ser claras, não ambíguos e seguir o devido processo;
4) Leis, ordens e práticas de restrição de conteúdos devem seguir os testes de necessidade e proporcionalidade;
5) Leis, políticas e práticas de restrição de conteúdo devem respeitar o devido processo;
6) Transparência e prestação de contas devem ser integradas em leis e em políticas e práticas de restrição de conteúdos.

2.2.3. Princípios de Santa Clara

Os Princípios de Santa Clara sobre Transparência e Responsabilidade na Moderação de Conteúdo[31] (*Santa Clara Principles*) possuem um objetivo semelhante. Elaborados por um pequeno grupo de organizações, ativistas e acadêmicos, à ocasião da Primeira Conferência de Moderação de Conteúdo em Escala – que ocorreu em Santa Clara, na Califória, em 2018 –, os princípios servem como um ponto de partida ao delinear níveis mínimos de transparência e responsabilidade sobre a matéria.

Visando à obtenção de transparência e à criação de um regime de responsabilidade significativo em torno da moderação cada vez mais agressiva do conteúdo gerado pelo usuário nas plataformas de Internet, os *Santa Clara Principles* estabelecem:

1) Números: As empresas devem publicar os números de postagens removidas e contas permanente ou temporariamente suspensas devido a violações de suas diretrizes de conteúdo. Esses dados devem ser fornecidos em um relatório regular, idealmente trimestral, em um formato aberto e digitalmente acessível.

2) Aviso: As empresas devem notificar cada usuário cujo conteúdo seja retirado ou cuja conta seja suspensa sobre o motivo da retirada ou suspensão. Os avisos devem estar disponíveis em uma forma duradoura e acessível mesmo que a conta do usuário seja suspensa ou encerrada. Os usuários que denunciam conteúdo também devem ter acesso a um registro do conteúdo que relataram e os resultados dos processos de moderação.

3) Recurso: As empresas devem oferecer uma oportunidade válida para o recurso oportuno contra qualquer remoção de conteúdo ou suspensão de conta. A longo prazo, podem ser estabelecidos processos de revisão externa independente também enquanto um mecanismo para que os usuários possam buscar reparação.

2.2.4. Ranking Digital Rights

Outra iniciativa afim é aquela proposta pelo projeto Ranking Digital Rights[32] ("Classificando Direitos Digitais", em tradução livre), que reúne pesquisadores de diversos países para elaborar metodologias de avaliação e classificação das empresas mais relevantes no campo das tecnologias da informação e comunicação (TICs) no que diz respeito às suas políticas no âmbito da liberdade expressão e privacidade. Com base em 58 indicadores, divididos em três categorias – governança, liberdade de expressão e de informação, e privacidade –, o projeto avalia os compromissos e práticas relacionados a direitos digitais por parte de empresas como Apple, Samsung, Facebook, Twitter, Amazon, Baidu, entre outras.

Na pesquisa *Corporate Accountabiity Index* de 2020, nenhuma das 26 empresas ranqueadas obtiveram resultados satisfatórios nos quesitos de transparência e *accountability*, baseados em padrões internacionais de direitos humanos. O Twitter foi a "melhor dos piores"[33], com a nota 53 (de 100), ficando apenas um ponto à frente da Verizon Media, proprietária do Yahoo. Google e Facebook obtiveram as notas 48 e 45, respectivamente. Dentre as plataformas digitais analisadas, a pior nota foi para a Amazon: 20%. Uma das principais informações trazidas no relatório do projeto referente ao ano de 2020 foi que Facebook e Google estavam entre as empresas com o maior *gap* entre as políticas que declaravam praticar e o que de fato faziam na prática.

2.2.5. NetzDG

Um importante e muito discutido marco global no tema tem sido a lei alemã *Netzwerkdurchsetzungsgesetz* (NetzDG). A lei, promulgada em 1° de setembro de 2017, entrou em vigor no mês seguinte no dia 1 de outubro com uma previsão de um período de adaptação para as empresas que se esgotou no dia 1° de janeiro de 2018. A lei causou uma grande reação da comunidade da internet à época, porém aos poucos ganhou perenidade e se tornou um dos principais marcos globais de regulação da internet de países democráticos.

Seu mecanismo de funcionamento é simples, porém eficaz. A lei impõe algumas obrigações para provedores de redes sociais como obrigações referentes a relatórios de transparência, estabelecimento de um sistema de gerenciamento de reclamações e a obrigação de designar um representante legal no país.

Dentre as obrigações impostas, duas tocam diretamente o tema da moderação de conteúdo: a) relatórios de transparência e b) estabelecimento de um sistema de gerenciamento de reclamações. Especialmente a obrigação de um gerenciamento de queixas traz uma grande novidade para o tema da moderação de conteúdo. Ilícitos elencados no §1 inciso 3 da NetzDG servem de baliza para que a proteção dos usuários ocorra primeiramente num procedimento simplificado na própria plataforma. A gestão desse sistema de denúncias tem um duplo efeito: por um lado garante um direito de acesso à justiça para os usuários dentro do espirito do mundo digital e por outro lado torna a moderação de conteúdo, ou seja, a administração privada da liberdade expressão da população, uma atividade mais transparente e balizada por ditames mínimos de direito público.

No contexto da lei NetzDG, há uma previsão de uma avaliação independente após 3 (três) anos de vigência visto a novidade em termos legislativos para um ambiente de inovação. O parecer foi então encomendado pelo Ministério da Justiça alemão ao Prof. Martin Eifert que o concluiu neste ano de 2021 trazendo vários importantes elementos como a constatação que não houve qualquer indicio do grande temor inicial: o de a lei causar um *chilling effect*, ou seja, um apagamento em massa das empresas privadas. Um outro desenvolvimento interessante foi uma nova reforma na lei alemã no sentido de aperfeiçoar criando novos mecanismos como a necessidade de as empresas noticiarem crimes ocorridos em suas plataformas à autoridades públicas, criação de formas de disputas arbitrais para conflitos entre usuários e plataforma e acesso à dados das plataformas para fins de pesquisa. Todas essas mudanças entraram em vigor dia 28 de junho deste ano.

2.3 Críticas e recomendações atuais à atividade de moderação

Embora as plataformas tenham passado a divulgar números de atividade de moderação, induzidas por pressões governamentais,[34] foram tímidas em relação à disponibilização de mecanismos efetivos de contestação, permanecendo arbitrárias as decisões pelos provedores quanto à exclusão de conteúdos e contas.[35]

Por outro lado, o debate acadêmico e de ativistas digitais[36] tem apontado a insuficiência dessas medidas. Note-se que todas elas se referem a boas práticas *após a tomada de decisão* pelo provedor de aplicação acerca de medida sobre conteúdos ou contas. Mas uma série de preocupações tem sido levantada sobre a adequação e impactos a direitos fundamentais decorrentes do próprio processo decisório adotado pelas plataformas. Esse processo de tomada de decisão envolve, basicamente, equipes de moderadores e ferramentas computacionais para a detecção ou decisão automatizada sobre o conteúdo.

2.3.1. Sobre os Moderadores

Quanto aos moderadores, há três ordens de questionamento. A primeira diz respeito à adequação dos padrões de moderação, tipicamente

modulados para democracias constitucionais, assumindo-se, por exemplo, que a desinformação e o ódio são criados e tolerados apenas por uma minoria, quando na verdade podem ter o suporte da maioria da população e ter inclusive origem governamental.[37] A segunda concerne à necessidade de competência dos moderadores para lidar com questões locais, incluindo o conhecimento da língua e aspectos culturais e políticos relevantes, considerando que, por vezes, a moderação não é nacional e até mesmo pode ser executada de modo terceirizado para empresas estrangeiras.[38]

Recentemente, o Facebook fez esforço louvável em constituir e divulgar seu *Oversight Board*[39] para moderação de conteúdo, enfatizando a equipe escolhida, formada por membros conhecidos por sua defesa da liberdade de expressão e direitos humanos. Porém, o Oversight Board decide pequeno número de casos selecionados, com o objetivo de oferecer diretrizes para as equipes de moderadores, que efetivamente tomarão decisões sobre moderação de conteúdo na rede social. Pouco se sabe sobre a composição, treinamento e capacidades das equipes de moderadores, em qualquer das redes sociais hoje dominantes. O episódio de Mianmar,[40] em que o *Facebook* foi incapaz de detectar como a rede social foi utilizada como arma para incitar violência contra a minoria étnica *Rohingya*, ilustra bem essas duas ordens de preocupação. A terceira preocupação concerne às condições de trabalho do moderador, que é exposto constantemente a imagens e conteúdo degradantes, o que pode afetar sua saúde psíquica.[41]

2.3.2. Automação da Moderação

Por sua vez, a automação na moderação para a detecção e tomada de decisão de conteúdo abusivo, tanto ex ante quanto ex post, tem sido estimulada pela regulação como o *Copyright Directive* da União Europeia, o *NetzDG* alemão e a lei australiana *Abhorrent Violent Material*, que obrigam a adoção de medidas em tempo eficiente, por vezes, dentro de 24 horas. Porém, críticos dessa legislação apontam para o risco de excesso de confiança nessas ferramentas[42] e que a discricionariedade das plataformas no seu uso pode levar a resultados contraproducentes,[43] como violações à liberdade de expressão, pelo excesso de bloqueios,[44] e a outros direitos humanos, caso não haja transparência e prestação de contas em relação a essas ferramentas,[45] em particular quanto a seu papel na tomada de decisão, quanto a sua explicabilidade, i.e. critérios relevantes para a classificação de postagens como *fake news* ou *hate speech*, muitas vezes tornada opaca pelo algoritmo empregado,[46] seu nível de acurácia e a possível incorporação de traços sociais discriminatórios[47] ou distorções em acurácia conforme gênero ou raça.[48]

Por fim, considerando que, embora não respondam por sua elaboração, a propagação dos conteúdos é viabilizada pela plataforma, de modo que tem sido destacada a necessidade dos provedores de aplicação adotarem os meios técnicos a seu alcance visando a mitigar impactos eventualmente provocados pelas ações ou omissões da moderação. Tal atuação inclui não só restaurar contas e conteúdos equivocadamente excluídos, como também empregar meios disponíveis para corrigir a informação ou informar sobre o caráter abusivo do conteúdo àqueles que a ele foram expostos, ou no mínimo, que com ele interagiram.[49]

2.3.3. Diagnóstico sobre a moderação de conteúdo no PL 2630/20

A partir do debate internacional recente, é possível observar que o tratamento dado pelo PL 2630/2020 à moderação de conteúdo é incompleto. Isso porque, além de não estipular a necessidade dos termos de serviço dos provedores de aplicação incorporarem parâmetros de conteúdo ilícito, conforme legislação nacional sobre a comunicação e sobre a publicidade, tratam apenas de um aspecto da moderação, a saber, aquilo que ocorre após a tomada de decisão dos provedores sobre o conteúdo, com a obrigatoriedade de oferecer oportunidade de defesa e recurso, além da divulgação de métricas da moderação.

Faltam regras de transparência e prestação de contas naquilo que ocorre no processo de tomada de decisão pela plataforma, seja em relação aos moderadores, seja em relação aos algoritmos empregados, bem no que diz respeito à disponibilização de mecanismos claros para reclamação para os usuários ofendidos e ao empenho dos provedores no sentido de mitigar impactos do conteúdo nocivo que foi propagado na rede.

2.3.4. Pontos de atenção sobre moderação de conteúdo

Nesse sentido, os seguintes pontos merecem especial atenção no tema da moderação de conteúdo privado pelos provedores:

1) As regras elaboradas pelo provedor de aplicação sobre moderação de conteúdo devem ser debatidas de modo amplo, à luz de seus impactos específicos sobre o discurso público, envolvendo diferentes representantes da sociedade civil, considerando-se o papel relevante do discurso veiculado nas plataformas para o exercício de direitos individuais, para a coesão social e para a democracia.[50] Tanto as regras como as ferramentas para sua aplicação, incluindo a moderação automática e humana devem ser descritas de modo transparente. As plataformas devem responder tanto pelas regras quanto por sua aplicação, assegurando que adota os procedimentos e medidas adequadas, tendo em vista seu papel essencial como fóruns do discurso público e como palcos para exercício de uma gama de direitos humanos online, incluindo liberdade de expressão, mas também a liberdade de informação, direito à dignidade e à privacidade.

2) A moderação de conteúdo *online* contra a desinformação deve abranger não somente serviços de redes sociais, que permitam ao usuário publicar conteúdos *online* para redes de relacionamentos ou ao público em geral (*Facebook, Twitter, etc.*), como também serviços de postagem e compartilhamento de vídeos (*Youtube*), bem como serviços de busca que direcionem o usuário para conteúdo desinformativo (*Google search*). Assim, o artigo sobre moderação não deve limitar sua referência a provedores de redes sociais, mas abranger os provedores de aplicação em geral;

3) Atenção deve ser dada, também, à moderação de publicidade. O provedor tem a oportunidade e o dever de filtrar o conteúdo, previamente, ao seu impulsionamento, considerando que obterá retorno financeiro

pela difusão. Esse dever precisa estar aliado às regras de transparência para identificação das fontes de financiamento de propagação de conteúdos ilícitos, submetendo-se à regulação existente sobre publicidade no Brasil;

4) A atividade de moderação humana deve contar com equipe de revisores adequada ao volume de comunicações, com competência linguística e conhecimento da cultura local adequados para o exame do conteúdo;

5) Ferramentas computacionais que empreguem inteligência artificial cada vez mais têm se mostrado eficazes e efetivas em custos para detectar e classificar conteúdo ilícito. Porém, são falíveis e devem atender a parâmetros éticos para evitar que sejam contraproducentes ou ameacem outros direitos fundamentais, dentre eles, padrões aceitáveis de acurácia para seu emprego, transparência, explicabilidade e controle de vieses;

6) Dado o caráter falível da moderação e a possibilidade de exclusão ou marcação da expressão pública de conteúdo lícito e legítimo, o poder de moderação das plataformas não pode ser arbitrário, devendo incorporar também mecanismos online de contestação, facilmente acessíveis pelo usuário e que garantam o efetivo contraditório, além de respostas rápidas e justificadas pelos provedores quanto a sua decisão final;

7) Deve haver transparência pelos provedores em relação à equipe empregada para moderação de conteúdo, empresas contratadas e aos meios técnicos, em particular, sobre os sistemas de inteligência artificial utilizados;

8) Os provedores de aplicação devem, na medida do possível naquilo que for tecnicamente viável, buscar informar ao usuário que teve contato com conteúdo que violou os termos e usos do contrato ou que alvo de medidas de checagem por fontes independentes. Essa medida informativa pela plataforma vem em consonância com seus deveres de tráfego visando a contribuir com a atenuação dos efeitos da circulação e propagação de desinformação ou conteúdo abusivo. Essa medida deve atingir, no mínimo, aqueles que visualizaram ou interagiram explicitamente com o conteúdo abusivo pelo período em que foi mantido na rede e quando possível, todos aqueles que foram expostos ao conteúdo. Provedores de aplicação, em suas especificidades técnicas, utilizam métricas diversas para mensurar engajamento e monitoramento dos usuários que efetivamente tiveram contato com os diferentes tipos de conteúdos disponibilizados. Nesse contexto, é importante que o texto consiga abarcar a pluralidade de métricas para garantir a eficácia da norma e o alcance da medida para o máximo de usuários afetados, conforme a viabilidade técnica[51].

2.4. Proposta de redação para o art. 12 do PL 2630/2020

Art. 12. Os procedimentos de moderação de conteúdo consistem nos mecanismos de governança que estruturam a

participação em plataforma de internet para prevenir abusos e violações aos termos de uso dos provedores de aplicação e compreendem medidas para a classificação, tomada de decisão sobre o conteúdo postado ou contas em operação, devendo abranger:

I - regras adequadas que observem e promovam direitos fundamentais e o Estado de Direito no ambiente online;

II - equipe adequada de moderadores e ferramentas computacionais confiáveis;

III - canal online efetivo para recebimento e apuração de denúncias por usuários;

IV - procedimento em plataforma online para recurso contra decisões pelos provedores de aplicação que garanta o devido processo; e

V - meios técnicos para mitigar o impacto de conteúdo ilícito propagado na plataforma.

§ 1º Os termos de uso dos provedores de aplicação deverão incluir, dentre os conteúdos e as contas em operação sujeitos à aplicação de medidas, aqueles que veiculem discurso ou práticas manifestamente ilegais, que constituam desinformação, calúnia, difamação, injúria, ameaça e incitação à violência, exercício ilegal da medicina, arte dentária ou farmacêutica, charlatanismo, curandeirismo, crimes resultantes de preconceito de raça ou de cor, exploração sexual de menores ou crimes eleitorais, nos termos da legislação vigente.

§ 2º Os termos de uso em relação à publicidade online e impulsionamento de conteúdo deverão observar as normas de publicidade no País, em especial as Leis nº 4.680/65 e nº 12.232/10, além da Lei nº 8.078/90, devendo os provedores de serviço de aplicação distinguir claramente perante o usuário o conteúdo patrocinado e publicidade, bem como moderar seu conteúdo previamente à difusão publicitária ou impulsionamento.

§ 3º A equipe de moderadores deverá ter dimensão proporcional ao volume de usuários da plataforma, receber treinamento adequado para a função e acesso a tratamento profissional para preservação de sua saúde psíquica, possuir competência linguística e conhecimento da cultura local para o exame do conteúdo.

§ 4º A aplicação de medidas sobre conteúdo e contas em operação com base em decisões automatizadas deve ser transparente quanto ao papel da automatização, bem como propiciar informações gerais sobre os seus critérios de operação, padrões de acurácia e medidas para controle de vieses.

§ 5º Deverá ser disponibilizado aos usuários canal facilmente acessível para denúncias sobre conteúdo e contas em operação.

§ 6º Os provedores de aplicação de internet submetidos a esta Lei devem assegurar, nos processos de elaboração e aplicação de

medidas com base em seus termos de uso, mecanismos de recurso e devido processo.

§ 7º Em caso de denúncia ou de medida aplicada em função dos termos de uso das aplicações ou da presente Lei que recaia sobre conteúdos e contas em operação, o usuário deve ser notificado sobre a fundamentação, o processo de análise e a aplicação da medida, assim como sobre os prazos e procedimentos para sua contestação.

§ 8º Os provedores dispensarão a notificação aos usuários se verificarem risco: de violação a direitos de crianças e adolescentes; de crimes tipificados na Lei nº 7.716, de 5 de janeiro de 1989; de grave comprometimento da usabilidade, integridade ou estabilidade da aplicação. de postagem reiterada de conteúdo idêntico ou similar a conteúdo sobre o qual já foi adotada medida pelo próprio provedor ou por outros provedores de aplicação.

§ 9º Deve ser garantido pelo provedor o direito do usuário recorrer de decisões adotadas pelo provedor de aplicação sobre conteúdos e contas, bem como corrigir classificações equivocadas de conteúdo, em tempo útil e eficaz, com a restauração de contas indevidamente suspensas ou excluídas por violações aos termos de uso.

§ 10º O prazo de defesa será diferido nos casos de conteúdo que use imagem ou voz manipuladas para imitar a realidade, com o objetivo de induzir a erro acerca da identidade de candidato a cargo público, ressalvados o ânimo humorístico ou de paródia.

§ 11º A decisão do procedimento de moderação deverá assegurar ao ofendido o direito de resposta na mesma medida e alcance do conteúdo considerado abusivo.

§ 12º O provedor de aplicação deverá notificar os usuários acerca de adoção de medida sobre conteúdo com o qual tenha havido interação ou tenha sido visualizado, por meio de qualquer ferramenta disponível na plataforma, fornecendo, quando possível, informação acurada sobre o tema, proveniente de fonte independente.

§ 13º Os relatórios de transparência, previstos no § 1º do art. 13, deverão informar sobre os procedimentos e resultados da moderação, incluindo informações sobre empresas eventualmente contratadas para moderação, quantidade de pessoas envolvidas, bem como sobre métricas de publicidade e conteúdo impulsionado.

3 Autorregulação Regulada

3.1. Sobre o instituto da autorregulação regulada e sua adequação para o tema das *fake news*

O alvoroço global nos últimos anos em torno do tema *fake news* é perfeitamente compreensível, quando se torna transparente e nítida a função da esfera pública nas democracias modernas. A esfera pública moderna denota a forma como a constituição e a circulação da informação política e social é estruturada dentro de um Estado, tendo-se tornado parte essencial e basilar de todas as democracias liberais modernas[52]. Sem esse pré-requisito perde-se de vista o porquê da recorrência do tema do perigo das notícias fraudulentas nos meios digitais.

Um dos problemas centrais que surge ao se debater, cogitar e propor uma regulação para as redes sociais advém do caráter dinâmico do mundo digital, que aumenta a incerteza sobre a eficácia de uma possível regulação. De fato, a incerteza e celeridade das transformações do mundo digital exige uma maior criatividade ou experimentalismo de uma regulação para poder lidar com tamanha tarefa.

Uma forma moderna de lidar com a crescente incerteza do ponto de vista regulatório encontra-se no instituto da autorregulação regulada[53]. Este procura trabalhar no liame entre duas tradicionais formas de regulação: autorregulação e regulação por um terceiro-normalmente o Estado ("*Fremdregulierung*"). Por um lado, a autorregulação tem a vantagem da eficiência pela disposição do conhecimento interno e dinâmica de constante revisão de conceitos. Por outro, tem a desvantagem por não necessariamente perseguir interesses e valores públicos. Já a regulação por terceiro ("Fremdregulierung") tem a vantagem de poder ser implementada por coerção em nome do interesse público e a desvantagem de, em ambientes dinâmicos, não dispor de conhecimento suficiente para lograr êxito na persecução do objetivo[54].

A autorregulação regulada oferece outra e nova possibilidade de lidar com as incertezas, ao conciliar vantagens das duas abordagens alternativas: a regulação versus a autorregulação. Foca-se no importante momento de auto-organização conforme expertise e dinâmica própria da indústria, estimulando-se, porém, alguns parâmetros gerais de interesses públicos caros ao Estado e sociedade. Nesse sentido, a autorregulação regulada consegue "induzir" o setor privado a contribuir para o cumprimento de tarefas públicas. Essa forma de regulação pode lidar melhor com uma sociedade que cada vez mais se locomove, e se distancia, de uma sociedade centrada em organizações conseguindo absorver melhor as incertezas e construir parâmetros melhores de eficácia na regulação.

Niklas Luhmann, em em seu livro de 1995 sobre a realidade da comunicação de massa, conseguiu resumir na conhecida frase uma importante dimensão da esfera pública

moderna: "Tudo que nós sabemos sobre nossa sociedade, e também sobre o mundo, no qual nós vivemos, sabemos através dos meios de comunicação de massa"[55]. Naturalmente, Luhmann estava expondo a centralidade da comunicação de massas em nossas vidas, e para época, década de noventa, especialmente a centralidade da produção da informação por organizações ou empresas, como jornal impresso, rádio e televisão. O que torna meios de comunicação de massa ("*Massenmedien*")[56] é sua capacidade de produzir uma "interrupção no contato" entre o emissor e receptor, ou seja, o nível de interação entre as organizações e o público geral torna-se secundário frente as formas de seleção e rotina da comunicação[57], sendo feita essa interação indireta de forma indireta através de gêneros de programas, pesquisa de audiência etc. O aspecto da massificação da informação significa que na produção da informação não há um direcionamento individual da informação. Ela passa a ser cada vez mais geral e acessível.

Nessa sociedade das organizações, a "veracidade", ou melhor, a qualidade da informação estava estritamente ligada por um lado, aos *standards* jornalísticos advindo da formação profissional dos jornalistas com suas técnicas de checagem e padrões éticos, e por outro lado também pela responsabilidade civil e penal do redator-chefe do jornal. Em outras palavras, a dimensão da organização tinha ao mesmo tempo o poder, de dentro das redações dos jornais, vincular um *ethos* jornalístico com a vinculação às normas e precedentes quanto ao que pode ser publicado no tocante às notícias não verídicas que esbarrem em calúnia, difamação etc. A estrita ligação entre o padrão profissional e organização com deveres legais correlatos, ou seja, de duas dimensões, deveres jurídicos organizacionais e padrões profissionais do jornalismo, ao caminhar lado a lado moldaram e conformaram a formação de uma esfera pública das democracias liberais marcadas pela pluralidade e possibilidade de controle judicial e político, que por sua vez começam a ruir com a entrada de um outro meio de comunicação: a Internet[58].

Com a popularização do mundo digital e o papel cada vez mais relevante dos usuários não só no consumo, mas na própria produção do conteúdo (os chamados "produmidores"), a centralidade das organizações aos poucos perde seu valor. Esse fato se dá acima de tudo porque aquele "contato interrompido" entre emissor (grandes empresas jornalísticas) e receptor (público geral), que caracterizava a esfera pública na sociedade das organizações, volta a se restabelecer. Com as redes sociais, a produção de informação nova, pode-se dar de forma desvinculada das organizações jornalísticas, ou seja, da interação entre o emissor e o receptor, interação entre leitores em *blogs*, em posts no *Facebook* etc. Isso, sem que a informação produzida de modo pulverizado perca seu alcance, que não só tem profusão abrangente, como também tem sua eficácia ampliada pela possibilidade de direcionamento para públicos específicos.

Daí decorre não somente o problema das notícias fraudulentas, mas também de seu combate eficaz via jurídica e política. A produção de uma informação fraudulenta ou a produção em escala industrial por *sites* de notícias fraudulentas, que circulam rapidamente nas redes sociais, não encontram a mesma forma eficaz de resposta jurídica através de uma decisão de remoção de conteúdo e direito de resposta como na sociedade das organizações.

Se antes a esfera pública era centrada na produção de informação por jornais impressos e televisão, a esfera pública atualmente é centrada em redes e plataformas digitais. O foco da regulação não pode ser nem uma organização tradicional, nem somente o indivíduo que produziu o conteúdo. Como na economia das plataformas[59] a acumulação de dados pelas plataformas se torna a mercadoria mais importante, a regulação, ou os parâmetros gerais da regulação, deve se concentrar em estabelecer alguns parâmetros gerais, os quais as próprias redes sociais devem criar uma organização para cumprir tais parâmetros. Somente se os parâmetros forem atingidos e respeitados, a organização é reconhecida como instituição da autorregulação regulada. A forma como é aplicado e executado os objetivos públicos, fica a cargo dos mecanismos tecnológicos disponíveis pelas redes sociais[60].

Um ponto central de uma organização reconhecida como autorregulação regulada nesse contexto seria de estabelecer um procedimento transparente e em plataforma digital[61], no qual possam ser reclamadas mas também contestadas a adoção de medidas de controle pelos provedores de redes sociais. Com isso, pode-se traçar a curto prazo, de forma transparente e com direito de defesa, de onde se emana grande parte das notícias fraudulentas. Acima de tudo, pode-se aos poucos concretizar regras sociais e a própria cognição quanto a aceitabilidade de determinadas notícias em determinados meios.

Resumidamente, pode-se afirmar que, diante da complexidade e incertezas advindas do mundo digital, as opções de regulação estatal ficam bem restritas. Porém, a opção do instituto da autorregulação regulada apresenta-se como viável em lidar com os desafios das notícias fraudulentas nos meios eletrônicos visto que ela reúne duas características importantes que uma regulação deve ter:

1) a participação do objeto da regulação na implementação dos objetivos públicos, visto que o Estado não possui conhecimento técnico para suprir tal demanda;

2) o estabelecimento de determinados parâmetros a serem seguidos pela instituição da autorregulação regulada, parâmetros esses advindos do interesse público.

3.2. Limitações da previsão da autorregulação regulada no PL 2630/2020

O instituto da autorregulação regulada, como já indicado acima, foi introduzido no debate por meio do Substitutivo proposto pelo Senador Antônio Anastasia e tem sua inspiração na NetzDG alemã. A proposta teve êxito em sua recepção pelo atual PL2630/20, porém, aspectos cruciais para que o instituto prospere, presentes na proposta original, não foram incorporados. A autorregulação regulada vem prevista, no projeto atual, no art. 30, que transcrevemos abaixo:

> *Art. 30. Os provedores de redes sociais e de serviços de mensageria privada poderão criar instituição de autorregulação voltada à transparência e à responsabilidade no uso da internet, com as seguintes atribuições:*

I – criar e administrar plataforma digital voltada à transparência e à responsabilidade no uso da internet, que contenha regras e procedimentos para decidir sobre a adoção de medida informativa, atendendo ao disposto nesta Lei;

II – assegurar a independência e a especialidade de seus analistas;

III – disponibilizar serviço eficiente de atendimento e encaminhamento de reclamações;

IV – estabelecer requisitos claros, objetivos e acessíveis para a participação dos provedores de redes sociais e serviços de mensageria privada;

V – incluir em seu quadro uma ouvidoria independente com a finalidade de receber críticas e avaliar as atividades da instituição; e

VI – desenvolver, em articulação com as empresas de telefonia móvel, boas práticas para suspensão das contas de usuários cuja autenticidade for questionada ou cuja inautenticidade for estabelecida.

§ 1º A instituição de autorregulação deverá ser certificada pelo Conselho de Transparência e Responsabilidade na Internet.

§ 2º A instituição de autorregulação poderá elaborar e encaminhar ao Conselho de Transparência e Responsabilidade na Internet relatórios trimestrais em atendimento ao disposto nesta Lei, bem como informações acerca das políticas de uso e de monitoramento de volume de conteúdo compartilhado pelos usuários dos serviços de mensageria privada.

§ 3º A instituição de autorregulação aprovará resoluções e súmulas de modo a regular seus procedimentos de análise.

A principal limitação presente na redação atual está na falta de um mecanismo de incentivo para que as plataformas possam aderir à instituição de autorregulação. Esse mecanismo de incentivo é a isenção de punição às plataformas que aderirem à instituição.

Como a proposta legislativa traz uma série de obrigações procedimentais com as melhores práticas de governança na moderação de conteúdo, isso significa que uma instituição de autorregulação homologada assegura a adequação a tais práticas, de modo que seus membros, ao seguir seu código e procedimentos, estará conforme à lei. Daí porque se legitima a isenção de punição.

Outros aspectos que não foram incorporados também fazem falta para a efetividade do instituto. Faltam delineamentos mínimos sobre a estrutura da instituição de autorregulação para que seja reconhecida pelo Conselho de Transparência e Responsabilidade na Internet, incluindo plataforma online para tomada de decisão sobre medidas de moderação de conteúdo com membros representativos de diferentes setores da sociedade civil.

Por outro lado, é importante resgatar o papel da Instituição de autorregulação na elaboração e atualização de código de conduta, não só para orientar a atividade de moderação pelas plataformas, como também para trazer subsídios e parâmetros técnicos para decisões judiciais sobre moderação.

De modo a suprir essas lacunas e garantir o efetivo funcionamento do instituto de autorregulação regulada, propomos a seguinte redação.

3.3. Proposta de redação para autorregulação regulada

DA AUTORREGULAÇÃO REGULADA

Art. 30º. O Conselho de Transparência e Responsabilidade na Internet reconhecerá, como instituição de autorregulação, a entidade formada por provedores de aplicação associados que se enquadrem no art. 1º, § 1º desta Lei, que satisfaça os seguintes requisitos:

I – crie e administre plataforma digital voltada ao recebimento de denúncias sobre conteúdos ou contas e tomada de decisão sobre medidas de moderação a serem implementadas por seus associados, bem como a revisão de decisões de moderação de conteúdo e contas por seus associados, por meio de provocação por aqueles afetados diretamente pela decisão;

II- contenha órgão competente para tomar decisões, em tempo útil e eficaz, sobre as denúncias e revisão de medidas de moderação adotadas pelos associados formado por analistas representativos de diferentes setores da sociedade civil, incluindo, entre outros, representantes dos consumidores, da imprensa, do jornalismo audiovisual, de empresas provedoras de conexão e de aplicações de internet, de entidades acadêmicas e organizações não governamentais em campos ligados à temática desta Lei;

III – assegure a independência e a especialidade de seus analistas;

IV – disponibilize serviço eficiente de atendimento e encaminhamento de reclamações;

V – estabeleça requisitos claros, objetivos e acessíveis para a participação dos provedores de redes sociais e serviços de mensageria privada;

VI – inclua em seu quadro uma ouvidoria independente com a finalidade de receber, encaminhar e solucionar solicitações e críticas e avaliar as atividades da instituição;

VII – desenvolva, em articulação com as empresas de telefonia móvel, boas práticas para suspensão das contas de usuários cuja autenticidade for questionada ou cuja inautenticidade for estabelecida;

VIII- estabeleça e divulgue em seu sítio na internet Código de Conduta para a implementação desta Lei, vinculante para seus

associados, resoluções sobre seus procedimentos de análise e súmulas interpretativas, com base na experiência de seu órgão decisório;

IX- o Código de Conduta deverá ser revisado periodicamente de modo a refletir os parâmetros interpretativos e experiência de decisões sobre moderação de conteúdo e contas.

§1º O usuário poderá realizar a solicitação à ouvidoria prevista no inciso VI por, pelo menos, meio telefônico ou eletrônico.

§2º O prazo de solução da solicitação do usuário deve ser de 5 (cinco) dias úteis.

§3º A instituição de autorregulação deverá elaborar e encaminhar ao Conselho de Transparência e Responsabilidade na Internet relatórios trimestrais em atendimento ao disposto nesta Lei;

§4º A instituição de autorregulação aprovará resoluções e súmulas de modo a regular seus procedimentos de análise.

§5º Os provedores de aplicação membros da instituição de autorregulação reconhecida pelo Conselho de Transparência e Responsabilidade na Internet não estarão sujeitos às penalidades previstas no artigo 33, caput e incisos I e II desta Lei, acerca de comportamento que seja conforme às determinações da instituição, sem prejuízo de sanções civis e penais.

NOTAS

1. Cf.: DATA PRIVACY BRASIL. "RASTREABILIDADE, METADADOS E DIREITOS FUNDAMENTAIS: NOTA TÉCNICA SOBRE O PROJETO DE LEI 2630/2020". Disponível em: https://www.dataprivacybr.org/wp-content/uploads/2020/07/Data-Privacy-Brasil.-Rastreabilidade-e-Direitos-Fundamentais.-PL-2630.2020.pdf.

2. Cf.: INTERNETLAB. "ESTRATÉGIAS DE PROTEÇÃO DO DEBATE DEMOCRÁTICO NA INTERNET". Disponível em: https://www.internetlab.org.br/wp-content/uploads/2020/07/il_policypaper2_estrategias-de-protecao_20200715.pdf.

3. Cf.: LAPIN. "NOTA TÉCNICA SOBRE O PL 2.630.2020 | SOBRE A INCLUSÃO DE MECANISMOS DE TRANSPARÊNCIA ALGORÍTMICA NO PL DAS 'FAKE NEWS'". Disponível em: https://lapin.org.br/2020/08/19/nota-tecnica-pl-2-630-2020-sobre-a-inclusao-de-mecanismos-de-transparencia-algoritmica-no-pl-das-fake-news/.

4. Cf.: COALIZÃO DIREITOS NA REDE. "PL 2630/20: propostas da CDR para uma lei efetiva e democrática". Disponível em: http://plfakenews.direitosnarede.org.br/pl-2630-20-propostas-da-cdr-para-uma-lei-efetiva-e-democratica/.

5. Cf.: ELECTRONIC FRONTIER FOUNDATION. "FAQ: Por que o Projeto de Lei Brasileiro de Tornar Obrigatória a Rastreabilidade em Aplicativos de Mensageria Privada Frustrará a Expectativa de Privacidade e Segurança dos Usuários". Disponível em: https://www.eff.org/pt-br/deeplinks/2020/08/faq-why-brazils-plan-mandate-traceability-private-messaging-apps-will-break-users.

6. Disponível em: https://www.bbc.com/portuguese/internacional-41843695.

7. WARDLE, Claire; DERAKHSHAN, Hossein. Information Disorder: Toward na interdisciplinar framework for research and policy making. Council of Europe, 2017; Disponível em: https://rm.coe.int/information-disordertoward-an-interdisciplinary-framework-for-researc/168076277c.

8. Disponível em: https://www1.folha.uol.com.br/poder/2018/10/empresarios-bancam-campanha-contra-o-pt-pelo-whatsapp.shtml.

9. Disponível em: https://legis.senado.leg.br/sdleg-getter/documento?dm=7740092&ts=1593906687173&disposition=inline.

10. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/141944.

11. Disponível em: https://www2.camara.leg.br/atividade-legislativa/comissoes/grupos-de-trabalho/56a-legislatura/aperfeicoamento-da-legislacao-brasileira-internet.

12. GRIMMELMANN, James. The virtues of moderation. Yale Journal of Law & Technology, 2015, v. 17. Disponível em: https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=1110&context=yjolt.

13. Barrett, Paul. M. Who moderates the social media giants? In: New York University Stern Center for Business and Human Rights, 2020. Disponível em: https://bhr.stern.nyu.edu/blogs/2020/6/4/who-moderates-the-social-media-giants.

14. GOOGLE. Transparency Report. Cumprimento das diretrizes da comunidade do YouTube. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR.

15. FACEBOOK. Relatório de Aplicação dos Padrões da Comunidade, agosto de 2021. Disponível em: https://about.fb.com/br/news/2021/08/relatorio-de-aplicacao-dos-padroes-da-comunidade-agosto-de-2021/.

16. TWITTER. Insights from the 17th Twitter Transparency Report. Disponível em: https://blog.twitter.com/en_us/topics/company/2020/ttr-17.

17. TWITTER. Transparency. Rules enforcement. Disponível em: https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec.

18. JØRGENSEN, Rikke Frank. When Private Actors Govern Human Rights. Research Handbook on Human Rights and Digital Technology 346, 363. In: Ben Wagner, Matthias C. Kettemann, Kilian Vieth (Orgs.) 2019; ROBERTS, Sarah T. Behind the Screen. Content moderation in the shadows of social media, Edward Elgar: Londres, 2019 p. 33;

19. Barett, Paul, Regulating Social Media: The Fight Over Section 230 — and Beyond, September 2020. Senator Hawley Introduces Bill to Remove Section 230 Immunity from Behavioral Advertisers, Office of Senator Josh Hawley, July 28, 2020, https://www.hawley.senate.gov/senator-hawley-introduc- es-bill-remove-section-230-immunity-behavioral-advertisers, CITRON, Danielle K; FRANKS, Mary Anne. The internet as a speech machine and other myths confounding section 230 Reform. Boston University School of Law Public Law and Legal Theory Paper, 2020, n. 20-8; CITRON, Danielle K; WITTES, Benjamin. The internet will not break: Denying bad Samaritans § 230 immunity. Fordham Law Review, v. 86, pp. 401–423.

20. WIELSCH, Dan. Os ordenamentos das redes: Termos e condições de uso - Código - Padrões da comunicação. em: Ricardo Campos, Georges Abboud, Nelson Nery Jr. (Orgs.) Fake News e Regulação, 2a. Edição, RT: Sao Paulo 2020, p. 91 ss.

21. No Brasil, no mínimo, as práticas capituladas nos tipos penais presentes nos artigos 138,139,140, 147, 282, 283 e 284 da Lei 7.209 de 11 de julho de 1984 (Código Penal), na Lei 7.716 de 5 de janeiro de 1989 e artigos 323 e 337 do Código Eleitoral e artigos 33, parágrafo 4o e 34, parágrafo 3o da Lei 9.504 de 30 de setembro de 1997.

22. No Brasil, naquilo que for cabível, as Leis n° 4.680/65 e a Lei n° 12.232/10 e ao disposto no art. 36 da Lei 8.078/90.

23. HAMILTON, Rebecca J. Governing the Global Public Square. Harvard International Law Journal, v. 62, 2021 (no prelo); JØRGENSEN, Rikke Frank; ZULETA, Lumi. Private Governance of Freedom of Expression on Social Media Platforms. Nordicom Review, 2020, pp. 51-67; CAPLAN. Ronyn; GILLESPIE, Tarleton. Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy. Social Media + Society, 2020; GORWA, Robert; BINNS, Reuben; KATZENBACH, Christian. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 2020; KETTEMANN, Matthias. Menschenrechte und politische Teilhabe im digitalen Zeitalter, 2020; TIEDEKE, Anna Sophia; KETTEMANN, Matthias. Back up: can users sue platforms to reinstate deleted content? Internet Policy Review Jounal on internet regulation, vol. 9, 2020.

24. THE SANTA CLARA PRINCIPLES on Transparency and Accountability in Content Moderation. Disponível em: <https://santaclaraprinciples.org/>. Acesso em: 26.ago 2020.

25. Ranking Digital Rights Header Branding. Disponível em: <https://rankingdigitalrights.org/index2019/>. Acesso em: 26.ago 2020.

26. GEBHART, Gennie. Who has your Back? Censorship Edition 2019. Electronic Frontier Foundation. Disponível em: < https://www.eff.org/wp/who-has-your-back-2019>. Acesso em: 26.ago 2020.

27. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, Seção 2, que requer a produção de relatórios de transparência. Ver traducao da lei alema para o português em: Ricardo Campos, Georges Abboud, Nelson Nery Jr. (Orgs.) Fake News e Regulação, 2a. Edição, RT: Sao Paulo 2020, p. 337.

28. EUROPEAN COMMISSION. Regulation of the European Parlament and of the Council on

preventing the dissemination of terrorist content online, 2018. O Art. 8(2) do referido regulamento obriga a publicação de requerimentos de remoção de conteúdo.

29. Disponível em: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680790e14

30. ELECTRONIC FRONTIER FOUNDATION. Princípios de Manila Sobre Responsabilidade dos Intermediários. Disponível em: https://www.eff.org/files/2015/07/02/manila_principles_1.0_pt.pdf

31. Cf.: "The Santa Clara Principles On Transparency and Accountability in Content Moderation". Disponível em: https://santaclaraprinciples.org/.

32.Cf.: "2020 Ranking Digital Rights Corporate Accountability Index". Disponível em: https://rankingdigitalrights.org/index2020/.

33.BROUILLETTE, Amy. Key findings: Companies are improving in principle, but failing in practice.

Disponível em: https://rankingdigitalrights.org/index2020/key-findings.

34. MASNICK, Mike. How Government Pressure Has Turned Transparency Reports from Free Speech Celebrations to Censorship Celebrations, Techdirt. Disponível em: https://www.techdirt.com/blog/?d=17&m=4&y=2018.

35. GILLESPIE, Tarleton. Custodians of the Internet. Platforms, content moderation, and the hidden decisions that shape social media, New Haven 2018, p. 24 ss.; KLONICK, Kate. The New Governors: The People, Rules, and Processes Governing Online Speech" 131. Harvard Law Review, 1598.

36. Cf.: https://globalvoices.org e https://datasociety.net/.

37. WU, Tim. Is the First Amendment Obsolete? Michigan Law Review, vol. 117, 2018, pp. 547-557; HAMILTON, Rebecca J. Governing the Global Public Square. Harvard International Law Journal, v. 62, 2021 (no prelo).

38. GILLESPIE, Tarleton. Custodians of the Internet. Platforms, content moderation, and the hidden decisions that shape social media, New Haven. Yale University Press, 2018; ROBERTS, Sarah T. Digital detritus: 'Error' and the logic of opacity in social media content moderation. First Monday, 23(3); SUZOR, Nicolas P.; WEST, Sarah Myers; QUODLING, Andrew; YORK, Jillian. What do We mean when We talk about transparency? Toward meaningful transparency in commercial content moderation. International Journal of Communication 2019, vol 13. Disponível em: https://ijoc.org/index.php/ijoc/article/view/9736.

39.Disponível em: https://www.oversightboard.com/news/announcing-the-first-members-of-the-oversight-board/?fbclid=IwAR1oZWZjr3eJqa06kwLkWJZTaSG7H8M9bnPz2qNu0RlbBNSXopEjPnZwZuI.

40. STECKLOW, Steve. Inside Facebook's Myanmar Operation Hatebook: A Reuter's Special Report, REUTERS. Disponível em: https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ [https://perma.cc/2QWF-D6L8.

41. ROBERTS, Sarah T. Behind the Screen. Content moderation in the shadows of social media, Yale University Press: Londres, 2019 p. 33.

42. LI, Sydney; WILLIAMS, Jamie. Despite What Zuckerberg's Testimony May Imply, AI Cannot Save Us. Electronic Frontier Foundation.. Disponível em: https://www.eff.org/deeplinks/2018/04/despite-what-zuckerbergstestimony-may-imply-ai-cannot-save-us.

43. BLOCH-WHEBA, Hannah. Automation in Moderation, Cornell Internattional Law Journal 2020.

44. BARZIV, Sharon; ELKIN-KOREN, Niva. Behind the scenes ofonline copyright enforcement: Empirical evidence on notice & takedown. Connecticut Law Review, 2018; URBAN, Jennifer; KARAGANIS, Joe; SCHOFIELD, Brianna. Notice and takedown in everyday practice. UC Berkeley Public Law Research Paper, 15.mar 2016.

45. GORWA, Robert; BINNS, Reuben; KATZENBACH, Christian. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society 2020.

46. ANANNY, Mike.; CRAWFORD, Kate. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society , 2018, 20(3), pp. 973–989. GORWA, Robert; Garton ASH, Timothy Garton. Democratic Transparency in the Platform Society. In: Persily N, Tucker Josh, Social Media and Democracy: The State of the Field and Prospects for Reform. Cambridge, UK: Cambridge University Press, 2020.

47. BINNS, Reuben, VEALE, Michael, KLEEK, Max Van; SHADBOLT, Nigel. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In: International conference on social informatics, pp. 405–415. Berlin: Springer, 2017.

48. ANGWIN, Julia; GRASSEGGER, Hannes. Facebook's Secret Censorship Rules Protect White Men, But Not Black Children. In: ProPublica. Disponível em: <org/article/facebook-hate-speech-censorship-internaldocuments-algorithms>. Acesso em: 25.ago 2020; HOFFMANN, Anna Lauren. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. Information, Communication & Society 2019, vol. 22, ed. 7, pp. 900–915.

49.ECKER, Ullrich K H; SWIRE-THOMPSON, Briony. Misinformation and its Correction: Cognitive Mechanisms and Recommendations for Mass Communication (2018). Disponível em: https://www.researchgate.net/publication/317603082_Misinformation_and_its_Correction_Cognitive_Mechanisms_and_Recommendations_for_Mass_Communication

50. Cf. KETTEMANN/SCHULZ, Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System (Leibniz Institute for Media Research Working Paper #2, 2020).

51. Sobre as métricas utilizadas, veja:

Facebook (https://www.facebook.com/business/help/743427195703387),

Youtube (https://support.google.com/youtube/answer/2991785?hl=pt-BR), Twitter (https://help.twitter.com/pt/using-twitter/media-studio-analytics);

52. HABERMAS, Jürgen. Strukturwandel der Öffentlichkeit Suhrkamp. Frankfurt am Main, 1990. p. 82 e 86; HONNETH, Axel. Das Recht der Freiheit. Frankfurt am Main, 2011. p. 474 ss. Para uma alternativa crítica à construção frankfurtiana centrada na política, da esfera pública ver: BLANNING, Tim. The Culture of Power and the Power of Culture. Old Regime Europe 1660 – 1789. New York: Oxford University Press, 2002.

53. COLLIN, Peter; STOLLEIS, Michael et al. (Orgs.). Regulierte Selbstregulierung in der westlichen Welt des späten 19. und führen 20. Jahrhunderts. Vittorio Klostermann. Frankfurt am Main, 2014.

54.LADEUR, Karl-Heinz. Die Regulierung von Selbstregulierung und die Herausbildung einer "Logik der Netzwerke". Rechtliche Steuerung und die beschleugnite Selbstorganisation der postmodernen Gesellschaft. Zeitschrift für Verwaltung und Verwaltungswissenschaft, Caderno 4, p. 59 ss., 2001.

55. LUHMANN, Niklas. Die Realität der Massenmedien. Opladen: Westdeutscher Verlag, p. 9.

56. A tradução brasileira do livro em questão exclui "massa" tanto do título quanto nos capítulos. Isso prejudica a compreensão visto que o livro trata justamente do aspecto da massificação da informação pelos meios de comunicação e sua diferenciação dentro da sociedade.

57. LUHMANN, Niklas. Die Realität der Massenmedien. Opladen: Westdeutscher Verlag, p. 11 e 33.

58. Sobre o papel dos meios pelos quais o direito circula e sua influência sobre a normatividade jurídica e social ver: VISMANN, Cornelia. Medien der Rechtsprechung. Fischer Verlag Frankfurt am Main, 2011. p. 97 ss.; VESTING, Thomas. Computernetzwerke. Velbrück Verlag, Weilerswist, 2015.

59. SRNICEK, Nick. Platform Capitalism. Polity Press Cambridge, 2017

60. THOMA, Anselm Christian. Regulierte Selbstregulierung im Ordnungsverwaltungsrecht. Duncker & Humblot, Berlim, 2007. p. 66 ss.

61. Ver especialmente os exemplos de Ethan Katsch no livro: KATSCH, Ethan. Digital Justice. Technology and the Internet Disputes. Oxford University Press, Cambridge, 2017.

Equipe do Instituto Legal Grounds responsável pelo documento

Juliano Maranhão

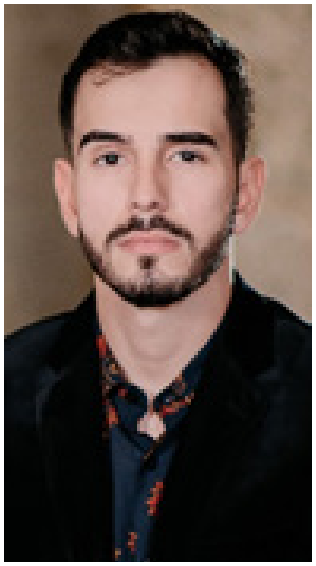
Diretor do Instituto Legal Grounds. Bacharel, doutor e livre docente pela Faculdade de Direito da Universidade de São Paulo, onde atualmente é professor associado do Departamento de Filosofia e Teoria Geral do Direito. Pesquisador da Fundação Alexander von Humboldt e professor convidado da Goethe- Universität Frankfurt am Main. Foi pesquisador visitante nas universidades de Miami, Leipzig e Maastricht. Pós-doutorado no departamento de ciência da computação da Universidade de Utrecht. Membro do Comitê Executivo da *International Association for Artificial Intelligence and Law* (IAAIL) e do Conselho Editorial do *Artificial Intelligence and Law Journal*.

Ricardo Campos

Diretor do Instituto Legal Grounds. Mestre e doutor pela Goethe Universität Frankfurt am Main. Docente em proteção de dados e novas mídias Goethe Universität. Docente no mestrado da Academia Europeia de Teoria do Direito. Vencedor do prêmio Werner Pünder (2021) com trabalho sobre regulação de serviços digitais.

Jessica Guedes

Pesquisadora no Instituto Legal Grounds. Mestranda em Direito pela Universidade de Brasília (UnB). Especi-alista em Direito Constitucional pelo Instituto Brasiliense de Direito Público (IDP) e Graduada em Direito pela mesma Instituição. Cofundadora do Portal Bot Jurídico. Advogada.



Samuel Rodrigues de Oliveira

Pesquisador no Instituto Legal Grounds. Doutorando em Teoria do Estado e Direito Constitucional pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio). Mestre em Direito e Inovação e bacharel em Direito pela Universidade Federal de Juiz de Fora (UFJF). Especialista em Relações Internacionais.

Maria Gabriela Grings

Pesquisadora do Instituto Legal Grounds. Mestre e Doutora em Direito Processual pela Faculdade de Direito da Universidade de São Paulo (USP). Bacharel em Direito pela Universidade Federal do Paraná (UFPR). Membro do Instituto Brasileiro de Direito Processual (IBDP). Advogada.'

Como citar: MARANHÃO, Juliano; CAMPOS, Ricardo; GUEDES, Jéssica; OLIVEIRA, Samuel Rodrigues de; GRINGS, Maria Gabriela. Regulação de "Fake News" no Brasil. São Paulo: Instituto Legal Grounds, 2021.