ARMADILHAS E CAMINHOS NA REGULAÇÃO DA MODERAÇÃO DE CONTEÚDO

Diagnósticos & Recomendações # 5
Agosto de 2021

Artur Pericles Lima Monteiro
Francisco Brito Cruz
Juliana Fonteles da Silveira
Mariana G. Valente

INTERNETLAB
pesquisa em direito e tecnologia

ARMADILHAS E CAMINHOS NA REGULAÇÃO DA MODERAÇÃO DE CONTEÚDO

AGOSTO 2021

COMO CITAR ESTE DOCUMENTO
Artur Pericles Lima Monteiro, Francisco Brito Cruz, Juliana Fonteles da Silveira e Mariana G. Valente, "Armadilhas e caminhos na regulação da moderação de conteúdo", Diagnósticos & Recomendações (São Paulo: InternetLab, 2021).

EQUIPE DO PROJETO
AUTORES Artur Pericles Lima Monteiro, Francisco Brito Cruz, Juliana Fonteles da Silveira e Mariana G. Valente
COLABORAÇÃO Heloisa Massaro
COMUNICAÇÃO Karina Oliveira
DIAGRAMAÇÃO Gabriela Rocha

INTERNETLAB
pesquisa em direito e tecnologia

SUMÁRIO

APRESENTAÇÃO

O QUE É O INTERNETLAB?

O InternetLab é um centro independente de pesquisa interdisciplinar, que produz conhecimento e promove o debate em diferentes áreas que envolvem tecnologia, direitos e políticas públicas. Somos uma entidade sem fins lucrativos baseada em São Paulo, que atua como ponto de articulação entre pesquisadores e representantes dos setores público, privado e da sociedade civil. Partimos da ideia de que a formulação de boas políticas públicas depende de diagnósticos mais precisos sobre a relação entre as novas tecnologias de informação e comunicação – como a internet – e os direitos das pessoas.

Veja mais no nosso site: www.internetlab.org.br

QUAL É O OBJETIVO DESTE DOCUMENTO?

Esta é mais uma intervenção do InternetLab para contribuir com o debate público em torno da regulação de plataformas, intensificado com a emergência de propostas ventiladas pelo Poder Executivo em 2021 (como a publicação de uma minuta de decreto, preparada pelo Ministério do Turismo, que modificaria o regulamento do Marco Civil da Internet). Da forma como foi apresentado, o texto proibiria a "moderação de conteúdo" realizada por plataformas digitais tal qual ela ocorre hoje.

Embora reconheçamos as preocupações legítimas sobre o arbítrio em relação à expressão online, especialmente quanto à atuação das empresas que têm maior número de usuários e poder econômico, neste documento apontamos as razões pelas quais entendemos que simplesmente abolir práticas de moderação de conteúdo, tal como proposto, seria equivocado se analisado do ponto de vista de direitos dos usuários e da proteção de um ambiente aberto e democrático na internet.

Este documento ainda indica outro caminho para que possamos responder aos anseios de diferentes setores quanto à moderação de conteúdo, que leva em conta os pontos acima e pode ser mais promissor, embasado numa abordagem procedimental de proteção a direitos dos usuários em face da aplicação de políticas de conteúdo por grandes empresas de internet.

PRINCIPAIS PONTOS

1 O Marco Civil da Internet foi concebido tendo em vista proteger a liberdade de expressão, afastando dos provedores de aplicações a responsabilidade jurídica por conteúdo de usuários. A moderação de conteúdo (entendida como a elaboração, implementação e aplicação de políticas próprias de conteúdo pelos provedores de aplicação de internet) não estava entre as principais questões em discussão no momento de sua edição, em 2014;

2 Uma lição do processo de elaboração do Marco Civil, celebrado como uma conquista para a internet no Brasil, é a participação ampla de diferentes setores da sociedade. Essa participação permitiu a formação de consensos amplos e equilibrados para a proteção de direitos fundamentais e de segurança jurídica;

3 Hoje, insatisfações justas sobre moderação de conteúdo têm se multiplicado, particularmente quanto à atuação das grandes empresas de tecnologia. Entre outras, destacam-se questões sobre falta de transparência na elaboração e aplicação das políticas, problemas no uso de tecnologias de automação, impacto desigual sobre grupos minorizados e influência desmedida sobre a esfera pública por parte de atores econômicos;

14 Ao mesmo tempo, é preciso reconhecer que a moderação de conteúdo tem um papel central na construção de um ecossistema com pluralidade de ambientes digitais, proporcionando diferentes formas de expressão e interação que contribuem com a promoção de democracia e do pluralismo;

15 Inviabilizar a elaboração, implementação e aplicação de políticas de conteúdo por provedores prejudicaria não só a expressão de usuários nas plataformas, mas também o acesso a informações, uma vez que os provedores não seriam mais capazes de manter em operação, da melhor forma que encontraram para servir os usuários espaços tão variados quanto sites de receita e a Wikipédia;

16 Uma regulação que adota conceitos revestidos de incerteza e deixa nas mãos do Executivo sua fiscalização cria um perigo real de captura das plataformas pelo governo, em violação da liberdade de expressão;

17 O direito autoral não pode ser usado como um argumento contra a moderação de conteúdo gerado por usuários, porque nem todo conteúdo é protegido por direitos autorais e mesmo o conteúdo protegido não cria automaticamente um dever de publicação e manutenção pelas plataformas;

18 Duas chaves devem guiar qualquer regulação para moderação de conteúdo: direitos dos usuários e defesa de um ambiente aberto e democrático na internet. Isso significa reconhecer deveres a plataformas, mas sem deixar de enxergar seu papel;

19 Uma abordagem procedimental, embasada principalmente em possibilidade de recurso e transparência, pode ser um caminho mais promissor para responder às legítimas preocupações quanto à questão da moderação de conteúdo por plataformas de internet.

INTRODUÇÃO

O que pode ser dito em plataformas da internet? E quem decide? Essas questões são levantadas por mais vozes, em cada vez mais alto e bom som, e têm apenas crescido em importância.

Parte substancial das respostas a tais questões hoje gira em torno daquilo que as grandes plataformas e quem as estuda chama de "moderação de conteúdo".[1] Esse é o termo empregado para nos referirmos às regras e aos procedimentos e sistemas usados pelas plataformas para remover, limitar alcance, rotular conteúdo como desinformação, assim como suspender ou remover contas. Num momento em que as preocupações e as desconfianças a respeito das novas tecnologias têm superado o otimismo, cresce a insatisfação com os termos em que se dão as respostas a tais questões tão básicas e tão fundamentais sobre a esfera pública digital. O peso das decisões de algumas poucas plataformas que concentram números exorbitantes de usuários e como essas decisões são tomadas são pontos cruciais nesse debate. O limitado acesso a informações sobre moderação de conteúdo, mesmo entre pesquisadoras e pesquisadores, contribui para um clima de suspeitas.

Esse clima é ainda alimentado por relatos reiterados de pessoas prejudicadas por erros na moderação de conteúdo que resultam em posts retirados do ar, assim como de pessoas lesadas e até mesmo silenciadas quando a moderação de conteúdo falta. A desinformação e as estratégias adotadas pelas plataformas para enfrentá-las também tornam o debate ainda mais complexo, em particular com a deterioração do espaço cívico e da polarização política vivida no país, cuja crise democrática tem contaminado até mesmo a resposta à pandemia. As plataformas são cobradas a fazer mais para manter um ambiente digital sadio e, ao mesmo tempo, criticadas por interferirem demais e cercearem a expressão de seus usuários. Se em algo os diferentes parecem concordar é que as plataformas têm deixado a desejar no trabalho que elas mesmo inventaram — o de criar e aplicar regras sobre a expressão online de populações de milhões de pessoas.

Nesse cenário, talvez retirar esse poder das plataformas pudesse soar como um caminho intuitivo para agradar gregos e troianos. Se a moderação de conteúdo tem problemas, por que não acabar com ela? Essa é a noção por trás de algumas propostas regulatórias recentes no Brasil, tanto no âmbito do Congresso Nacional quanto do Executivo. Este documento mostra por que esse caminho na verdade é uma armadilha — e busca apontar rumos mais promissores.

1 "Moderação de conteúdo consiste em processo por meio do qual plataformas de internet agem sobre contas ou conteúdos que violem seus termos de uso, impactando sua disponibilidade, visibilidade e/ou credibilidade. A moderação pode envolver diferentes medidas, tais como remoção, suspensão temporária, redução artificial de alcance ou proeminência, superposição de tela de aviso, adição de informação complementar, dentre outras". Thiago Dias Oliva, Victor Pavarin Tavares e Mariana G Valente, "Uma solução única para toda a internet? Riscos do debate regulatório brasileiro para a operação de plataformas de conhecimento", Diagnósticos & Recomendações (São Paulo: InternetLab, 2020), 11, internetlab.org.br/.... É importante lembrar que "moderação de conteúdo" é um termo empregado também para se referir à atividade comunitária, ou de usuários em determinados espaços como grupos e fóruns, com a mesma finalidade de aplicar regras a respeito de conteúdos alheios. Neste material, primordialmente nos referimos à atuação das plataformas; subsidiariamente, àquela comunitária, quando esse é o modelo principal de moderação de plataformas.

O CENÁRIO ATUAL: O MARCO CIVIL DA INTERNET

O CONTEXTO DA APROVAÇÃO DA LEI

Não há como discutir moderação de conteúdo no Brasil sem antes entender o que diz o Marco Civil da Internet (Lei nº 12.965/2014). O Marco Civil foi a primeira legislação brasileira integralmente voltada para regular a internet, determinando direitos e garantias para usuários, e estabelecendo deveres e obrigações para as empresas fornecedoras de serviços na rede e o Estado. É, sobretudo, a política mais bem estruturada até o momento no setor, inaugurando um quadro de governança da internet no Brasil.

As principais discussões relacionadas à liberdade de expressão na época da aprovação do Marco Civil não focavam nos conteúdos restringidos pela moderação de conteúdo por atividade dos próprios provedores de aplicação de internet (que é um conceito que abrange desde blogs até as plataformas). O cerne da questão era outro, quase o oposto: evitar incentivos jurídicos que fizessem com que provedores agissem mais contra conteúdos gerados por seus usuários para evitarem ser responsabilizados por danos causados por eles.

Isso porque a jurisprudência brasileira tendia a admitir que esses provedores fossem responsabilizados pelo conteúdo postado por usuários sempre que, tendo apenas uma notificação requerendo sua remoção, mantivessem-no disponível. Esse regime de responsabilização de intermediários, conhecido como *notice and takedown*, levantava preocupações por criar um incentivo às plataformas para realizar a remoção a despeito da procedência da queixa, considerando o risco que assumiriam caso deixassem o conteúdo no ar.[2]

De fato, o Superior Tribunal de Justiça (STJ), em entendimento consolidado em 2012, portanto antes da aprovação do Marco Civil, estabeleceu que ao provedor não é atribuída responsabilidade pelo conteúdo gerado por terceiros, desde que "uma vez notificado de que determinado texto ou imagem possui conteúdo ilícito, o provedor [retire] o material do ar no prazo de 24 horas, sob pena de responder solidariamente com o autor direto do dano, pela omissão praticada".[3]

2 Thiago Oliva, "Responsabilidade de intermediários e a garantia da liberdade de expressão na rede", *InternetLab* (blog), 23 de abril de 2019, internetlab.org.br/...

3 STJ, 3ª Turma, REsp 1.323.754/RJ, rel. min. Nancy Andrighi, j. 19.jun.2012.

Particularmente considerando o contexto brasileiro de falta de parâmetros fortes sobre liberdade de expressão no Brasil, esse modelo de responsabilidade de intermediários recebeu críticas justas por conta de seus efeitos nocivos à internet como fórum plural, dado o incentivo econômico para restrições excessivas por parte das plataformas para que elas fugissem de passivos gerados por notificações extrajudiciais mobilizadas por indivíduos e organizações com maior acesso a recursos.

O caminho adotado pelo Marco Civil da Internet, em seu artigo 19, foi o de evitar esse cenário de ameaça à liberdade de expressão. O Brasil seguiu um modelo em que os provedores de aplicação apenas serão responsabilizados por danos decorrentes de conteúdo veiculado em suas plataformas em caso de omissão após determinação judicial de remoção.

UMA LIÇÃO DO MARCO CIVIL: PROCESSO DEELABORAÇÃO ABERTO, COM AMPLA PARTICIPAÇÃO SOCIAL

Essa importante conquista do Marco Civil da Internet é fruto de um processo extenso e complexo envolvendo diversos atores da sociedade, dentre eles legisladores e membros do Poder Executivo, pesquisadores, empresas e organizações ativistas por direitos digitais. Em oposição ao um processo legislativo ordinário insulado em anéis burocráticos, o Marco Civil foi forjado em uma elaboração normativa participativa e que privilegiou análises técnicas aprofundadas sob prismas diversos, sendo submetido a consulta pública e norteado pelas diretrizes construídas em fórum multissetorial, o Comitê Gestor da Internet no Brasil (CGI.br).

Para além da participação em si, a consulta pública tinha um objetivo estratégico de formular consensos equilibrados e tecnicamente adequados e conferir legitimidade à futura legislação. O resultado foi atingido, transformando a aprovação em experiência modelo na regulação de tecnologia que promova direitos fundamentais. Essa é uma lição importante que o Marco Civil traz para qualquer proposta de regulação da internet: a participação social, num processo aberto, é indispensável, tanto para obtenção de subsídios dos mais diversos setores afetados, quanto para o próprio sucesso e efetividade da norma. A própria lei reconhece isso, ao estabelecer que o Comitê Gestor da Internet no Brasil, órgão multissetorial com atribuição na governança da internet no Brasil, participe do desenvolvimento de políticas públicas relacionadas ao setor (art. 24, III).

MUDANÇA NOS VENTOS: MODERAÇÃO DE CONTEÚDO NO FOCO

A moderação de conteúdo, como já dito, não estava no centro do processo que culminou com a edição do Marco Civil, em 2014. Embora nos últimos meses tenham surgido posições contrárias, a interpretação que sempre prevaleceu é que a lei se limita a regular expressamente a responsabilidade de intermediários no caso de conteúdos que não foram removidos após ordem judicial; o Marco Civil não regula e, por isso, permite o reverso, a moderação de conteúdo, pela qual os provedores restringem conteúdo publicado por usuários.[4]

Recentemente, no entanto, algumas propostas regulatórias têm adotado uma outra leitura, que também passou a ser defendida pelo Executivo, particularmente após a minuta de decreto divulgada em maio deste ano. Segundo essa posição, em razão de o art. 19 da lei isentar os provedores de aplicação de responsabilidade por conteúdo de terceiros a não ser quando descumprem decisão judicial que determina a indisponibilização, os provedores também não poderiam remover conteúdo sem ordem judicial para tanto. Como vimos, essa leitura está em dissonância com o que motivou a inclusão de tal dispositivo na lei, e não encontra amparo sólido na própria legislação.

4 Clara Iglesias Keller, "Policy by judicialisation: the institutional framework for intermediary liability in Brazil", *International Review of Law, Computers & Technology*, 2020, 1-19, doi.org/10.1080/13600869.2020.1792035.

Além disso, o texto da proposta também considera que o presidente da República pode regular a moderação de conteúdo por decreto. O fundamento para isso seria o disposto no art. 8º do Marco Civil, segundo o qual são nulas cláusulas contratuais que violem a liberdade de expressão. Um decreto presidencial poderia, segundo essa lógica, dispor sobre a moderação de conteúdo a título de mera especificação desse dispositivo. Isso ignoraria, no entanto, o quão complexa e repleta de difíceis escolhas é a regulação da moderação de conteúdo.

Parece completamente oposto ao processo aberto de construção que levou ao Marco Civil que o presidente da República unilateralmente modifique — sem ouvir a sociedade, pesquisadores, organizações da sociedade civil, grandes plataformas comerciais e não comerciais, assim como outros pequenos provedores — a regulação dessa importante prática que é a moderação de conteúdo, ainda mais a pretexto de expedir regulamento para fiel execução da lei, como estabelecido na Constituição a respeito de decretos (art. 84, IV). Também seria um grave erro excluir o Legislativo da elaboração normativa sobre esse assunto.

Nos meses seguintes, as ideias fundamentais dessa proposta de decreto foram vocalizadas por parlamentares e pelo próprio Executivo, abrindo margem para que o tema fosse avaliado pelo parlamento.

Seja qual for a via regulatória, tais ideias trazem armadilhas concretas para direitos, como veremos adiante. Em 2021, o cenário é outro; estamos distantes do momento de edição do Marco Civil, e o debate se modificou consideravelmente. Hoje, a moderação de conteúdo está no foco.

MODERAÇÃO DE CONTEÚDO E SEUS DESAFIOS

Diferentes setores têm expressado preocupação com as decisões de moderação de conteúdo, principalmente por parte de empresas de tecnologia que possuem enorme poder econômico. Desinformação, discurso de ódio e violência política são fontes de apreensão. A circunstância da pandemia e do contexto eleitoral estadunidense entre 2020 e 2021, que se desdobrou no escalonamento da desinformação e de seus impactos na saúde coletiva, assim como os eventos no Capitólio dos Estados Unidos,[5] têm sido apontadas como exemplos de como as empresas de tecnologia têm feito menos do que deveriam para coibir determinados discursos.

Mas se as plataformas têm sido criticadas por deixar de agir, o inverso também é verdadeiro: é comum a crítica e a apreensão por medidas de moderação de conteúdo consideradas excessivas ou equivocadas, particularmente quanto ao tamanho do poder exercido por um pequeno número de empresas. Isso se revelou especialmente quando a moderação de conteúdo passou a interferir em publicações de líderes políticos, que por

5 A exemplo do movimento de invasão do Capitólio. Cf. Charlie Savage, "Incitement to riot? What Trump told supporters before mob stormed Capitol", *New York Times*, 1o de outubro de 2021, nytimes.com/....

muito tempo foram alvo de tímida intervenção.[6] Algumas plataformas mantiveram ou mantêm políticas mais permissivas para essas figuras públicas, inclusive, a partir de um entendimento específico de que haveria mais interesse público nas suas manifestações.

No caso mais conhecido e publicizado, a suspensão do então presidente dos Estados Unidos, Donald Trump foi recebida com receio mesmo entre quem não se alinhava com a postura do mandatário, como a chanceler alemã, Angela Merkel,[7] e o presidente francês, Emmanuel Macron.[8]

E, ainda que os holofotes tenham se voltado ao tema quando as pessoas mais poderosas do planeta se viram afetadas, escolhas questionáveis, falhas e arbítrios na moderação de conteúdo têm impactado cidadãos comuns há muito tempo. São cotidianos os episódios de "falsos positivos" e "falsos negativos" no processo de moderação, ou seja, aquele conteúdo que foi excluído mesmo sem violar regra da plataforma e aquele que inequivocamente viola regra e foi mantido, respectivamente.

Isso ocorre pois essa atividade acarreta impactos enormes a direitos humanos, em especial quando realizada com finalidades econômicas e em escala global. Fundamentada em emitir determinados juízos de valor, a moderação de conteúdo ocasiona inevitavelmente desacordos sobre os limites do discurso e a providência que lhe é direcionada e acaba por resvalar em zonas cinzentas. E, embora as plataformas tenham buscado adotar políticas de conteúdo mais detalhadas e com regras mais claras, muitas vezes elas também não são capazes de trazer consensos e uniformizar a resposta a casos fronteiriços.

Ao mesmo tempo, muitos casos exigem mais informações sobre o contexto para se tomar uma decisão adequada ao exercício de direitos. Assim, é possível distinguir uma paródia para entretenimento de um conteúdo que se afigura como uma reprodução ilegal ou uma expressão legítima de protesto de uma incitação à violência.

Esse ideal, no entanto, torna-se impraticável quando se está em questão o serviço em escala que as grandes plataformas oferecem. Em razão do volume, as empresas desenvolveram ferramentas de inteligência artificial que são determinantes na decisão sobre visibilidade e disponibilidade de conteúdo. Tais sistemas automatizados podem não possuir a mesma precisão na capacidade de interpretar contextos do que revisores humanos, por exemplo. Ao falhar em identificar a emoção e a intenção do usuário, as tecnologias de aprendizado de máquina e processamento de linguagem natural utilizadas na moderação oferecem mais uma complexa camada de possível equívoco ou viés abusivo na tomada de decisão. O resultado disso é que podem comprometer o direito do usuário de se manifestar. Ou ainda, fragilizam o direito à dignidade e equidade quando essas imprecisões levam a publicações de discurso de ódio serem preservadas, enquanto manifestações de ativismo são excluídas.

6 Cf. "InternetLab apresenta contribuição ao Comitê de Supervisão do Facebook sobre a suspensão de Donald Trump das plataformas da empresa", *InternetLab* (blog), 24 de fevereiro de 2021, internetlab.org.br/...

7 Enrique Müller, "Merkel acha 'problemática' a suspensão de contas de Trump nas redes sociais", *El País*, 1o de dezembro de 2021, brasil.elpais.com/...

8 Dave Lawler, "Emmanuel Macron blasts social media platforms for banning Trump", *Axios*, 2 de abril de 2021, axios.com/...

Pesquisa do InternetLab demonstrou que tecnologias desenvolvidas e utilizadas por empresas de tecnologia para aferir a "toxicidade" de uma publicação podem não ser capazes de diferenciar conteúdo de ódio dirigido a LGBTQs do conteúdo publicado pelos próprios membros da comunidade LGBTQ utilizando-se de linguagem pseudo-ofensiva como forma de prepará-los para lidar com a hostilidade que lhes é frequentemente dirigida. Utilizando uma ferramenta de inteligência artificial semelhante às empregadas pelas plataformas,[9] a pesquisa indicou que a tecnologia considerou um número significativo de contas de *drag queens* no Twitter com níveis mais altos de toxicidade do que os nacionalistas brancos nos Estados Unidos.

Portanto, operando em contextos culturais, políticos e sociais radicalmente diversos e lidando com milhões de usuários, as equipes de elaboração de políticas e diretrizes, os sistemas automatizados e equipes de revisores não têm dado conta de realizar a moderação sem cometer erros e abusos à liberdade de expressão de seus usuários ou reforçar discursos discriminatórios e violentos ilegais.

Um exemplo brasileiro relevante dessas falhas diz respeito a uma pessoa que, em outubro de 2020, no Brasil publicou uma foto no Instagram, contendo oito fotografias que mostravam sintomas de câncer de mama, das quais cinco exibiam mamilos, com o título em português que indicava o apoio à campanha Outubro Rosa. Embora as políticas sobre nudez do Facebook incluam expressamente uma exceção permitindo publicações de mamilos para conscientização sobre o câncer de mama, o post foi removido pelo sistema automatizado por meio de aplicação do Padrão da Comunidade sobre Nudez Adulta e Atividade Sexual do Facebook.[10] O caso foi parar no Comitê de Supervisão criado pelo Facebook para reavaliar, de forma independente, decisões de moderação de conteúdo tomadas pela plataforma. Embora o próprio Facebook tenha reconhecido o erro antes da decisão no caso, o comitê enfatizou como a automação cria riscos para a liberdade de expressão — em muitos casos com prejuízos irreversíveis, como no caso em questão, dado que o conteúdo só foi restaurado após o mês de realização da campanha.[11]

Aos desafios de falha de identificação e detecção em escala e das decisões automatizadas somam-se o baixo *accountability* e a transparência insuficiente desse modelo. A insuficiência de dados divulgados pelas plataformas sobre essa atividade gera impactos em duas dimensões: interesse público e direito de defesa do usuário. Da perspectiva mais abrangente do público na medida em que as empresas exercem poder de mediação sobre o que é visto, falado ou suprimido no debate público mediante filtragem, etiquetagem

9 A pesquisa usa "Perspective", uma tecnologia de IA desenvolvida pela Jigsaw (anteriormente Google Ideas), para medir os níveis de toxicidade de tuítes de *drag queens* famosas nos Estados Unidos. Cf. Thiago Dias Oliva, Dennys Marcelo Antonialli e Alessandra Gomes, "Fighting hate speech, silencing drag queens? Artificial intelligence in content moderation and risks to LGBTQ voices online", *Sexuality & Culture* 25, no 2 (2021): 700–732, doi.org/10.1007/s12119-020-09790-w.

10 "Comitê de Supervisão revoga decisão original do Facebook: caso 2020-004-IG-UA", Comitê de Supervisão, *Notícias* (blog), janeiro de 2021, oversightboard.com/....

11 O InternetLab apresentou contribuição para a decisão do Comitê, alertando que, ainda que o conteúdo seja posteriormente restaurado, as consequências negativas decorrentes da remoção podem ser irreversíveis, como mostra este caso: o post, parte de uma campanha de conscientização sobre o câncer de mama em outubro de 2020, só foi restaurado no início de dezembro. Artur Pericles Lima Monteiro et al., "InternetLab apresenta contribuição para consulta pública do Comitê de Supervisão do Facebook", *InternetLab* (blog), 2 de abril de 2021, internetlab.org.br/....

e exclusão de conteúdo e conta, presume-se que existem compromissos a serem cumpridos, tais como informar a sociedade a partir de relatórios de transparência que contenham dados agregados sobre essas atividades. Na perspectiva do usuário, a ausência de informações básicas (sobre o motivo e critérios da indisponibilização, as políticas específicas tidas por violadas e as oportunidades e mecanismos de recurso) dificulta que o usuário se defenda de forma apropriada, e podem esconder possíveis arbitrariedades e falhas na decisão que levou à remoção e, portanto, limitações equivocadas à liberdade de manifestação do usuário.

Mas, se parece claro que a transparência deve ser aperfeiçoada, até nesse ponto há desafios. Oferecer informações extremamente detalhadas sobre a atividade de moderação pode miná-la, por fornecer elementos estratégicos para aqueles que tem como objetivo burlar a forma como as regras são aplicadas — o que é conhecido como *gaming the system*. Assim, se agentes de disseminação de spam souberem como empresas de internet detectam tais mensagens não-solicitadas, aprenderão como burlar essas regras e tornar a lotar *timelines* e caixas de mensagens com conteúdo publicitário indesejado ou fraudulento, por exemplo.

AS ARMADILHAS EM PROPOSTAS CONTRA A MODERAÇÃO DE CONTEÚDO

São muitas e diversas as críticas e preocupações quanto à moderação de conteúdo realizada pelas grandes empresas. Assegurar a liberdade de expressão de usuários de modo a proporcionar o livre debate de opiniões no ambiente digital e, sobretudo, nas redes sociais é primordial e está no centro da preocupação de todos os setores envolvidos. Em algumas propostas regulatórias recentes, no entanto, essas muitas questões recebem uma única resposta.

Assim, a proposta de decreto que circulou publicamente em maio de 2021 é um exemplo de uma corrente de ideias que se repetem. Essa corrente defende que seja estabelecida uma regra geral proibindo que provedores de aplicação controlem os ambientes que mantêm nas redes. Nesse tipo de proposta, a não ser em alguns casos excepcionais, provedores são proibidos de impor restrições sem ordem judicial tanto a contas quanto a conteúdo de usuários.

Se o objetivo é proteger usuários, no entanto, acabar com a moderação de conteúdo não é o caminho. Na verdade, muito pelo contrário: esse tipo de proposta cai em várias armadilhas e, mirando no poder de algumas empresas, acaba acertando outro alvo, prejudicando a liberdade de expressão.

DIFERENTES AMBIENTES PODEM SER CONSTRUÍDOS A PARTIR DA MODERAÇÃO DE CONTEÚDO

À primeira vista, a moderação de conteúdo poderia ser lida apenas como uma forma de limitar o que usuárias e usuários podem dizer em plataformas e outros espaços online. Mas isso não retrata a realidade integralmente. A moderação de conteúdo também é decisiva na configuração da internet tal como conhecemos. Por ser parte integrante da arquitetura digital das plataformas, ela tem o poder de construir diferentes ambientes e modelos de serviço online. A capacidade de desenvolver formas diversas de realizar tal atividade é chave para que os espaços criados na rede possuam regras e vocações próprias e diferentes umas das outras, criando a internet plural que nós conhecemos.

Esse desenvolvimento progressivo de espaços variados na internet está diretamente associado ao fato de os provedores adotarem políticas e práticas de moderação distintas umas das outras, cujo conteúdo visa a definir a adequação dos usuários ao escopo da plataforma e demais regras de comportamento para o funcionamento utilitário e sustentável do serviço. Em suma, não só a existência e aplicação de regras é fundamental, mas também a capacidade para que regras e procedimentos de moderação de cada um dos serviços possam ser diferentes uns dos outros.

Assim, algumas plataformas permitem que o próprio usuário controle sua experiência, moderando comentários feitos em suas publicações. O Instagram, por exemplo, prevê ferramentas para que os usuários decidam que comentários em suas fotos que possam ser considerados inadequados, ofensivos, intimidadores ou spam, mesmo que não violem Diretrizes da Comunidade, sejam ocultados.[12] Além disso, também é possível desativar comentários de todos os demais usuários, ou ainda apenas daqueles que não o "seguem".[13] O Twitter também disponibilizou, recentemente, ferramentas para que os usuários possam restringir que pessoas que não seguem comentem uma determinada publicação.[14]

Outro exemplo são as políticas quanto à nudez. Enquanto algumas plataformas, como Twitter, permitem nudez (e o usuário pode esperar ver esse tipo de conteúdo), em outras, como Facebook e Instagram, quase todo tipo de nudez é vedado.

12 "Como faço para ocultar comentários ou solicitações de contato que não quero ver no Instagram?", Instagram, [s.d.], help.instagram.com/....

13 João Kurtz, "Instagram permite escolher quem pode comentar nas suas fotos", *TechTudo*, 26 de setembro de 2017, techtudo.com.br/....

14 "Como controlar sua experiência no Twitter", Twitter, [s.d.], help.twitter.com/....

Essa variedade de regras e práticas de moderação dinamiza a produção e disseminação de conteúdo e de plataformas na internet e fornece ao usuário maior abrangência de oportunidades de relacionamento e de uso de serviço. Isso impacta não só o que os usuários podem postar e ver, mas também como recebem e se relacionam com o mesmo conteúdo.

Para voltar ao exemplo da nudez: uma plataforma em que esse tipo de conteúdo é permitido pode também ser um espaço em que mesmo publicações não relacionadas com nudez adotarão outro tipo de linguagem, mais adulto; já uma plataforma em que a nudez é proibida pode ser lida pelos usuários como um espaço em que a linguagem deve ser apropriada a todas as idades.

Além disso, até plataformas que adotem as mesmas políticas de conteúdo podem ser espaços diferentes a depender de como a moderação é conduzida, quanto tempo medidas levam a ser impostas e quão severas são as punições impostas a pessoas que violam as regras. De maneira geral, podemos pensar que uma plataforma mais rígida na aplicação de suas políticas de conteúdo pode incutir nos usuários um senso de autocontenção que se manifesta mesmo quanto a publicações que seriam permitidas pelas regras do provedor. Ao mesmo tempo, num outro extremo, uma plataforma mais tolerante na aplicação de suas políticas pode criar um ambiente em que as verdadeiras regras observadas pelos usuários são informais, não escritas, mas depreendidas a partir da experiência. Diferentes combinações são possíveis entre esses dois extremos.

Um paralelo pode ilustrar melhor o que queremos dizer. Uma sala de concerto e uma casa de shows são dois espaços voltados para a apreciação de música, em que a grande regra é não atrapalhar outras pessoas no público que querem aproveitar o espetáculo. Ainda assim, são dois ambientes decididamente diferentes.

Na sala de concerto, não só se espera que as pessoas não fiquem conversando, mas também que evitem qualquer tipo de ruído ou interferência — como mexer no celular ou abrir a embalagem de uma bala. Na casa de shows, falar ao telefone pode ser uma quebra de expectativa, mas ninguém espera silêncio absoluto; poucas pessoas se incomodariam com quem faz um comentário em voz baixa, e o público chega a cantar junto com os artistas. Novamente, o objetivo é o mesmo — aproveitar a música —, mas a experiência é completamente diferente, até mesmo em como consumimos e nos relacionamos com a apresentação: mais sisuda numa sala de concerto, mais descontraída numa casa de shows. Engessar a moderação de conteúdo pode prejudicar a construção desses diferentes espaços na internet, assim como as múltiplas formas de expressão e interação que proporcionam.

PROIBIR A MODERAÇÃO DE CONTEÚDO PREJUDICA O ACESSO À INFORMAÇÃO

Quando se fala em regular a moderação realizada pelos provedores, comumente se remete unicamente às redes sociais de propriedade de empresas multinacionais. Não raro as propostas levam em consideração os riscos e questões inerentes essencialmente a esse tipo de aplicação, ignorando as particularidades de outros provedores. Em 2020, outro número desta série de Diagnósticos & Recomendações do InternetLab foi dedicado a mostrar como "plataformas de conhecimento", como Wikipédia, Internet Archive e GitHub, têm organização e funcionamento distintos, o que inclui suas versões da moderação de conteúdo.[15] Alguns desses serviços instituem formas mais participativas de moderação de conteúdo, para que usuários membros da comunidade possam tomar parte na aplicação de regras, em contraposição ao modelo predominante adotado pelas multinacionais, em que a tarefa fica exclusivamente na mão de funcionários contratados ou de sistemas automatizados.

Abolir a moderação de conteúdo como conhecemos também prejudicaria essas plataformas, assim como outras que não costumam receber a devida atenção no debate regulatório. É preciso lembrar que o conceito de provedor de aplicação de internet do Marco Civil é extremamente abrangente;[16] inclui também sites em que usuários postam e comentam receitas, como o TudoGostoso, além das plataformas de conhecimento como a Wikipédia, os repositórios acadêmicos SSRN e Academia.edu, portais de educação como edX e Coursera, entre outras.

Se medidas de moderação de conteúdo são obstaculizadas, serviços e comunidades como esses podem ser impossibilitados, em claro prejuízo não só às oportunidades de expressão de usuárias e usuários, mas também ao acesso à informação proporcionada por tais provedores. O valor prático desses espaços seria anulado caso, por exemplo, o TudoGostoso não fosse capaz de eliminar textos publicados no site que não sejam receitas culinárias. E não seria preciso que tais textos tivessem nada de intrinsecamente ilícito. Um belo poema ou um ótimo conto não têm lugar num site de receitas, que só pode existir se puder fazer valer suas regras sobre conteúdo.

O mesmo problema sofreria a Wikipédia caso os editores não pudessem desfazer edições em dissonância com os padrões editoriais da enciclopédia.[17] É normal e até mesmo esperado que editores mais experientes apaguem ou corrijam informações de outros editores para garantir que as regras daquele espaço — oferecer informação confiável, imparcial e relevante — sejam cumpridas. Excepcionalmente, editores podem também ser banidos, por vandalismo ou spam nos artigos, por exemplo, como no caso de um vendedor utilizar diversos espaços de artigos da plataforma para promover publicidade a respeito de seu produto.

15 Oliva, Tavares e Valente, "Uma solução única".

16 Lei 12.965/2014: "Art. 5º. Para efeitos desta Lei, considera-se [...] VII - aplicações de internet: conjunto de funcionalidades que podem ser acessadas por meio de um terminal conectado à internet; [...]"

17 Oliva, Tavares e Valente, "Uma solução única", 10.

O caso da Wikipédia parece fundamental para a compreensão de como termos de uso, regras sobre conteúdo e a atividade de moderação podem existir para garantir a liberdade de informação e decorrentemente de expressão. O efeito de uma legislação que impede a possibilidade de excluir ou suspender conta ou conteúdo, sem ordem judicial, seria extremamente adverso para os serviços como a Wikipédia, que teriam seu funcionamento distorcido e sua finalidade alterada, tendo em vista que a moderação nessas plataformas é condição essencial para preservar a confiabilidade, qualidade e existência do serviço. É a cuidadosa curadoria do que o público e a comunidade oferecem que fazem esses sites serem o que são.

Sem que possam exercer o controle próprio a tais espaços, esses sites não seriam como conhecemos. Com isso, além das pessoas que ali publicam e contribuem, também seriam prejudicadas as milhões de pessoas que acessam tais plataformas para obter a impressionante informação que estamos acostumados a encontrar na internet — que já esperamos encontrar.

O PROBLEMA DO *SPAM*

A contenção no envio de *spam* - sigla em inglês para "envio de publicidade em massa" - e de outros conteúdos fraudulentos ou que representem insegurança aos usuários representa um dos motivos mais elementares que justifica a prática de moderação de conteúdo em grandes plataformas de internet. Para se ter uma ideia, no primeiro trimestre de 2021 mais de 80% dos canais removidos no YouTube foram indisponibilizados por se enquadrarem na proibição de spam, fraudes ou conteúdo enganoso pelas políticas da plataforma de vídeos.[18] Por sua vez, o Facebook informou que retirou quase 1 bilhão de conteúdos da rede social neste mesmo período.[19] A remoção ágil destas publicações (e perfis) dedicadas a inundar usuários com conteúdo indesejado é parte do que torna os serviços de tais empresas minimamente seguros (em relação a fraudes e crimes) e úteis nas finalidades que se propõe.

Ao mesmo tempo, "publicidade indesejada" não é um conteúdo necessariamente ilegal - muito pelo contrário. Visto como conteúdos individuais, peças de *spam* podem ser totalmente inofensivas e protegidas pela liberdade de expressão, por exemplo. É o caso de um anúncio de venda de doces e bolos; não há nada nele de problemático, mas se só isso dominar a experiência de um usuário em uma rede social, a rede social poderá ter perdido sua utilidade.

Para combater este tipo de problema à sua integridade, utilidade e à segurança de seus usuários, geralmente as empresas de internet aplicam políticas que lhes dão alguma discricionariedade sobre o que pode ser considerado *spam* ou conteúdo enganoso - especialmente para que agentes maliciosos ou fraudadores busquem utilizar brechas

18 "Cumprimento das diretrizes da comunidade do YouTube", Google Transparency Report, [s.d.], transparencyreport.google.com/....

19 "Community Standards Enforcement Report. Spam", Facebook Transparency Center, [s.d.], transparency.fb.com.

nas políticas para continuar a espalhar *spam*. Desta maneira, tais políticas que possuem certo grau de abertura e dificilmente conseguiriam ser contempladas em uma lista restrita e exaustiva de conteúdos que poderiam ser objeto de ação das plataformas sem ordem judicial.

Abolir ou engessar a moderação de conteúdo nestes casos pode tornar a atividade de provedores de aplicação de internet inviável, e seus serviços inúteis aos seus usuários. A ação contra *spam* é uma das armadilhas ocultas mais relevantes à regulação da moderação de conteúdo.

SE MODERAÇÃO É SEMPRE CENSURA, COMO JUSTIFICAR EXCEÇÕES PARA CONTEÚDO LÍCITO?

A corrente de ideias exemplificada na proposta de decreto que circulou em maio de 2021 parece admitir que não é possível simplesmente retirar todo o controle das plataformas. A minuta de decreto tornada pública pelo Executivo, ao mesmo tempo em que proibia restrições sem ordem judicial, criava exceções para "nudez ou representações explícitas ou implícitas de atos sexuais", "fabricação ou consumo, explícito ou implícito, de drogas ilícitas", "prática ou o ensino do uso de computadores ou tecnologia da informação com o objetivo de roubar credenciais, invadir sistemas, comprometer dados pessoais ou causar danos", entre outros.

Isso parece incompatível com a própria compreensão, manifestada pelo governo, de que qualquer tipo de moderação de conteúdo seria censura. Nudez e sexo não são ilícitos *per se* no Brasil. O Marco Civil adota um regime especial de responsabilidade apenas no caso de disseminação não consentida de conteúdo de "nudez ou de outros atos sexuais de caráter privado" (art. 21, *caput*); o dispositivo não se aplica para nudez em geral, que muitas vezes é representada em obras artísticas, por exemplo. Se conteúdo de nudez é lícito e, portanto, faz parte da expressão protegida pela Constituição, como poderia se negar a esse tipo de manifestação a proteção contra moderação de conteúdo que o governo considera necessária? Por coerência, seria de se esperar que qualquer conteúdo lícito fosse abrangido pela proposta regulatória.

O desfavorecimento da nudez na proposta regulatória pode ser considerado como uma violação da liberdade de expressão por cercear uma forma particular de manifestação, particularmente quando isso não é instituído em lei, como exigido pela Convenção Americana de Direitos Humanos (art. 13.2) e o Pacto Internacional sobre Direitos Civis e Políticos (art. 19.3), que fazem parte do direito brasileiro e estabelecem que restrições à expressão devem ser "expressamente previstas em lei". O mesmo se aplica a conteúdo que retrate consumo de drogas, que pode ser usado para fins artísticos ou acadêmicos, por exemplo, e para conteúdo pedagógico sobre segurança da informação, que pode noções sobre "invadir sistemas". Nada disso é necessariamente ilícito, e ao criar categorias desprivilegiadas das proteções oferecidas a outro tipo de conteúdo, a proposta regulatória pode ser considerada inconstitucional.

MAIS ARMADILHAS NAS EXCEÇÕES: O PERIGO DA CAPTURA DAS PLATAFORMAS

Mesmo quanto a conteúdo de fato ilícito, como violação da imagem e da privacidade, por exemplo, persistem tensões nas exceções à regra geral de vedação da moderação de conteúdo proposta no texto do Ministério do Turismo. O que caracteriza esses ilícitos está longe de ser incontroverso. E isso tem consequências que vão muito além.

Por exemplo, sabemos que pessoas públicas têm menos privacidade, mas como isso se traduz para a moderação de conteúdo? Uma foto do prefeito publicada por um cidadão para criticá-lo viola a privacidade do primeiro? E se a foto mostra o mandatário com sua família, num contexto reservado? Postar um *meme* com a imagem do presidente da República viola seu direito da personalidade? E se o *meme* retrata um guarda de trânsito? Em muitas situações, teremos diferentes respostas para essas perguntas.

O grande problema aqui se revela quando pensamos em como as plataformas reagirão diante dessas incertezas. Elas abrem espaço para responsabilização tanto caso provedores removam conteúdo que não deviam — por exemplo, porque avaliaram que uma publicação se enquadrava numa exceção à regra geral —, quanto não removam conteúdo que deviam — porque consideraram que não se aplicava uma exceção. Erros nesse juízo sobre regra geral e exceções sujeitaria plataformas e outros provedores a duras sanções, que incluiriam até mesmo o fim das operações no Brasil, segundo a minuta divulgada pelo Executivo em maio.

Considerando o risco de se exporem a sanções gravíssimas, confrontados com dúvidas sobre o que entra na regra geral e o que entra nas exceções, os provedores têm incentivo para adotar a visão de quem será responsável pela aplicação da regulação. Se órgãos sem independência do Executivo são os encarregados por isso, o perigo de captura da moderação de conteúdo é real. Isso pode ser considerado em si como uma violação da liberdade de expressão, porque é central a esse direito que as pessoas sejam livres particularmente para criticar justamente o governo e suas políticas.

Essa tensão sobre o conteúdo que se enquadra na regra geral de proibição e o conteúdo excepcionado pode gerar tanto um incentivo à não-remoção de determinados conteúdos por razões inadequadas quanto um incentivo à remoção indevida. Em outras palavras, tanto ao decidir manter quanto ao decidir remover conteúdo, o natural é que o provedor busque evitar riscos adotando a orientação que imagina que não criaria problemas com quem tem o poder de impor sanções.

Por exemplo, ao avaliar o que é "incitação de atos ameaça ou violência" (uma das exceções previstas na proposta de decreto já mencionada, por exemplo), a plataforma pode ser mais permissiva com conteúdo favorável ao governo — mesmo que o conteúdo fosse mesmo ilícito. Ao mesmo tempo, pode considerar que imagens de um protesto convocado por movimentos sociais sejam uma forma de "enaltecimento ou ajuda a organizações criminosas" (outra exceção) — ainda que se trate de uma entidade legítima, mas crítica ao governo. Em resumo, um modelo regulatório que incumba sua fiscalização a órgãos do Executivo cria um perigo genuíno de que as plataformas, como forma de autopreservação, restrinjam ou deixem no ar conteúdo segundo o que esperam que iria evitar choques com o governo. Isso dá a essas autoridades uma indevida influência sobre a esfera pública e viola a liberdade de expressão.

UM ESTRANHO NO NINHO: DIREITO AUTORAL E MODERAÇÃO DE CONTEÚDO

O direito autoral sempre teve um papel de destaque nas discussões sobre regras para o funcionamento dos provedores de aplicação de internet. Houve muito debate sobre qual modelo regulatório deveria ser aplicado nesse tema no Marco Civil da Internet, por exemplo. Nesse caso, a lei acabou estabelecendo que o disposto no art. 19 a respeito da limitação da responsabilidade de provedores ao descumprimento de ordem judicial de indisponibilização "depende de previsão legal específica" no caso de infrações a direitos de autor ou conexos (art. 19, § 2º). Também definiu que, até a edição dessa norma, a responsabilidade de provedores de aplicações "continuará a ser disciplinada pela legislação autoral vigente" (art. 31). Embora esse dispositivos já tenham sido interpretados como se o direito brasileiro atualmente adotasse um modelo de *notice and takedown* em matéria de direito autoral, a discussão segue aberta e persistem as críticas a essa opção regulatória, por conta de seu potencial para censura e limitação indevida do uso de obras por terceiros.[20]

20 Mariana G Valente, "Direito autoral e plataformas de internet: um assunto em aberto", *InternetLab* (blog), 18 de abril de 2019, internetlab.org.br/...

O direito autoral é uma disciplina jurídica que serve para promover relações justas no campo da cultura e do conhecimento, mas que importa à garantia da liberdade de expressão (e, especialmente, com a liberdade de expressão no ambiente virtual).

Isso porque é comum que conteúdos legítimos, protegidos pela liberdade discursiva do usuário e pela legislação de direito autoral, sejam removidos pelas plataformas sob o argumento de defesa do direito do titular da obra, especialmente porque muitas vezes o sistema de moderação dessas empresas apresenta falhas nos seus mecanismos automatizados ou lacunas interpretativas nas regras que o orientam.

Vítimas frequentes são a publicação de paródias e uso de pequenos trechos de obras – o que se dinamizou com as enormes possibilidades para criatividade proporcionadas por tecnologias digitais. Tal prática gera um custo alto para a liberdade de manifestação e liberdade artística dos usuários, que poderiam encontrar nessas plataformas uma ferramenta profícua para a difusão de suas ideias e encontram em obras técnicas ou artísticas uma via para elaborar críticas, fundamentar um argumento ou criar uma nova obra.

Recentemente, no entanto, especialmente na minuta de decreto sobre o Marco Civil, o direito autoral começou a aparecer na discussão com outro sentido — como um argumento para inviabilizar a moderação de conteúdo, a partir da noção de que todo conteúdo postado por usuários em redes sociais e outros sites seria protegido por direito autoral.

Esse argumento tem pelo menos dois problemas. Em primeiro lugar, é equivocado quando supõe que todo conteúdo publicado na internet é protegido por direitos autorais: somente são protegidas obras que apresentem alguma originalidade[21] — nos dizeres da Lei de Direitos Autorais, "criações do espírito" (art. 7º, lei 9.610/1998). A lei expressamente exclui de sua proteção, por exemplo, ideias, esquemas, planos, etc (art. 8º). Assim, por exemplo, não será uma obra protegida um tuíte que deseje "Feliz aniversário!" a alguém.

Em segundo lugar, mesmo que o conteúdo seja de fato protegido por direito autoral, isso por si só não significa que um provedor de aplicações seja obrigado a hospedá-lo. Ainda que os termos de uso e instrumentos assemelhados das plataformas incluam cláusulas pelas quais os usuários concedem uma licença para o provedor de aplicações quanto a conteúdo que inclua obras protegidas de que sejam titulares,[22] essas disposições criam uma faculdade de uso da obra. A licença não cria um dever de uso da obra[23] e, portanto, é compatível com moderação de conteúdo, inclusive mediante indisponibilização.

21 José de Oliveira Ascensão, *Direito autoral*, 2o ed (Rio de Janeiro: Renovar, 1997), 39-40; Pedro Paranaguá e Sérgio Branco, *Direitos autorais* (Rio de Janeiro: FGV, 2009), 24.

22 Ascensão, *Direito autoral*, 39-40.

23 Ascensão, *Direito autoral*, 311-12.

OUTRO CAMINHO PARA A REGULAÇÃO DA MODERAÇÃO DE CONTEÚDO

Se partirmos de um diagnóstico de que a atividade de moderação de conteúdo é necessária, mas complexa e tendente a colocar direitos fundamentais em risco, outros caminhos para regulá-la devem ser considerados por sociedades democráticas. Evitar as armadilhas que foram expostas acima não pode ser razão para imobilismo nesse debate – em especial se forem considerados o cerceamento ao exercício de direitos decorrentes de erros, arbítrios e opacidades na atividade de grandes plataformas de internet.

DOIS PONTOS-CHAVE

Dois pontos são chave para pensar uma regulação voltada a aperfeiçoar a moderação de conteúdo sem cair nas armadilhas apontadas acima: direitos dos usuários e defesa de um ambiente aberto e democrático na internet.

Direitos dos usuários

É preciso definitivamente reconhecer que usuárias e usuários têm direitos perante as plataformas e outros provedores. Propostas regulatórias recentes como a minuta de decreto divulgada em maio acertam nesse ponto, mas acabam indo longe demais. Diante do espaço ocupado no debate democrático e na vida social por plataformas como Facebook, YouTube e Twitter, é preciso deixar para trás uma lógica de que elas, como donas dos serviços, ditam as regras e sua aplicação sem dar satisfação a ninguém. Essa posição tem perdido espaço mesmo nos Estados Unidos, onde o direito constitucional à liberdade de expressão é visto como um limite à ação do Estado mas não de particulares. Aqui a virada envolve enxergar como expressão o que muitas vezes as empresas descrevem em termos de negócios como conteúdo.

Essa maneira de ver as questões também revela limites importantes, já que poucos provedores de aplicação funcionam como arenas de debate público como as plataformas mantidas pelas grandes empresas. Sites de receitas como TudoGostoso, por exemplo, não são voltados a isso, nem têm a mesma centralidade para a democracia — têm outra missão, o que não lhes faz ter menos valor.

Plataformas colaborativas que compõem a infraestrutura informacional democrática como a Wikipédia, por sua vez, têm suas próprias dinâmicas e devem ser vistos como espaços de construção e expressão coletiva. Não faz sentido pensar da mesma forma as garantias de usuários em canais para manifestação individual como redes sociais e direitos de usuários em ambientes como uma enciclopédia virtual aberta. Ainda que redes sociais também sejam palco para importante mobilização de grupos, não podem ser concebidas como uma obra colaborativa como a Wikipédia.

Em outras palavras, algumas redes sociais são como governantes de espaços quase públicos[24] e é apropriado pensar em direitos dos governados em relação a esses novos governantes. Já em plataformas colaborativas cruciais à democracia como a Wikipédia seria equivocado instituir os mesmos direitos entre usuários,[25] que mantêm entre si uma relação diferente, mais próxima à de coautores de uma obra, que se ajustam mediante concessões recíprocas e em busca de consenso em torno da empreitada coletiva, não por meio do exercício de direitos e pretensões jurídicas, um contra o outro. Em resumo, a lógica de direitos dos usuários tem seu lugar, mas não deve ser aplicada de modo indiscriminado e sem reflexão.

Proteção de um ambiente aberto e democrático na internet

A segunda chave é que a regulação não deve comprometer a internet como um espaço aberto e democrático no Brasil. Ao contrário, proteger tais valores deve ser um norte para qualquer modelo. E isso passa por entender que provedores de aplicação têm um papel importante a desempenhar. Propostas regulatórias recentes no Brasil erram ao desconsiderar isso. Acima, apontamos armadilhas num formato regulatório que vise impedir que provedores façam moderação de conteúdo. Aqui, queremos levantar uma questão mais geral.

A regulação da moderação de conteúdo não deve se dar à custa da experimentação e da inovação. Como destacamos em outro documento desta série Diagnósticos & Recomendações,[26] esse é um risco real especialmente quando práticas das grandes empresas são tomadas como modelos para toda a regulação. Além do caso de plataformas colaborativas e sem fins lucrativos como a Wikipédia, mesmo plataformas comerciais podem adotar outros modelos. Esse é o caso do Reddit, que tem uma espécie de federalismo na moderação de conteúdo: adota tanto uma política global, aplicada pela própria plataforma, quanto políticas locais a cada *subreddit* (como são chamadas as comunidades), geridas pelos próprios usuários. E esses são apenas alguns exemplos notáveis e atualmente existentes. Outras possibilidades de moderação de conteúdo podem ser imaginadas, desenvolvidas e testadas, o que é algo a ser preservado por qualquer regulação.

24 Kate Klonic, "The new governors: the people, rules, and processes governing online speech", *Harvard Law Review* 131, no 6 (abril de 2018): 1598–1670.

25 Oliva, Tavares e Valente, "Uma solução única", 21.

26 Oliva, Tavares e Valente, "Uma solução única".

Deveres regulatórios muito exigentes também podem acabar criando ainda mais concentração, por exemplo, ao estabelecer um aparato que apenas as empresas de capital multibilionário serão capazes de atender. Isso teria o efeito contrário ao buscado, dado que uma parte importante das preocupações com moderação de conteúdo está justamente no acúmulo de poder num pequeno número de empresas.

A chave da defesa do ambiente aberto e democrático na internet indica também o que pode ser exigido das plataformas. O foco deve estar na manutenção de um ecossistema sadio para comunicação e interação. Mais informações e maior transparência, particularmente por parte das grandes empresas, são componentes necessários para que o regulador seja até mesmo capaz de compreender o que merece sua atenção e resposta. Nem sempre essa resposta deve se dar por meio de intervenção nas próprias plataformas. No caso de desinformação, por exemplo, um importante instrumento regulatório é investir em estratégias de comunicação de conteúdo de qualidade, o que leva até mesmo a discussões sobre financiamento para jornalismo de alto padrão.

Ao mesmo tempo, vista a partir da ótica da saúde do ambiente na internet, essa prestação de contas por parte das plataformas não deverá ter como destinatários apenas reguladores. Uma sociedade democrática deve contar com pesquisadores independentes para compreender os problemas do ecossistema digital e formular saídas. Acadêmicos e organizações voltados à pesquisa desempenham função fundamental, mas dependem de acesso a dados. Ainda que não no mesmo grau que pesquisadores — que assumem compromissos éticos especiais quando trabalham com dados —, também usuários e cidadãos em geral devem ter acesso a informações a serem prestadas pelas plataformas.

Essa diversidade de envolvidos e interessados é central para a chave da defesa de um ambiente aberto e democrático na internet. Isso inclui as próprias plataformas e outros provedores de aplicação. Aqui podemos tirar uma lição de veículos de comunicação tradicionais, com quem o impacto das grandes plataformas tem sido comparado. A importância desses veículos para a democracia é inquestionável, como demonstrado pelo próprio fato de serem concessões públicas. Canais de TV, por exemplo, têm papel central até mesmo para o processo eleitoral, e sua atuação independente é vista como essencial para a legitimidade do feito.

Seria implausível ditar cada aspecto de suas atividades, também porque um ambiente de comunicação sadio só pode existir com uma pluralidade de visões sobre práticas e padrões desse campo. É crucial para a democracia que esses veículos sejam independentes. O mesmo se aplica ao ambiente digital. O objetivo da regulação não deve ser controlar diretamente os provedores de aplicação.

PERSPECTIVA PROCEDIMENTAL

Como propostas regulatórias podem equacionar essas duas chaves? Uma abordagem que pode ser promissora parte do que chamamos de uma perspectiva procedimental.

Essa abordagem seria uma forma de disciplinar a atuação das plataformas quando elas agirem sobre seus usuários, sem, contudo, engessá-la. No lugar de tachar qualquer moderação como censura, essa perspectiva busca regular esse processo, reconhecendo direitos aos usuários e proporcionando mais transparência, tanto aos usuários envolvidos numa controvérsia quanto à sociedade em geral.

Uma vantagem apontada para essa abordagem é que, de um lado, evita a determinação imprecisa e controversa do conteúdo (ou conta) a ser identificado e moderado – por remoção, redução de alcance ou rotulagem –, de outro, restringe o poder das empresas privadas de governarem o discurso sem escrutínio público. Ela permite, ainda, que os usuários possam entender melhor as razões das medidas tomadas por empresas de internet para que busquem terem reconhecidos seus direitos em um segundo momento, como em pleitos que visem reparação de danos causados.

Ao propor essa perspectiva regulatória, partimos de documentos como os princípios de Santa Clara, desenvolvidos por acadêmicos e organizações de direitos digitais nos EUA,[27] e o recente relatório da coalizão Al Sur, que congrega organizações de direitos digitais na América Latina.[28] A partir daí, reunimos as diretrizes abaixo, que buscam criar um processo de diálogo genuíno, baseado em dados confiáveis, entre plataformas, usuários, pesquisadores, sociedade civil e outros interessados.

Informações precisas sobre medidas de moderação e suas justificativas

Usuários devem ter informações detalhadas quanto ao conteúdo sobre o qual a plataforma restringe. Isso quer dizer ao menos (como preconizado no princípio Ciência dos princípios de Santa Clara):

A. apresentação da URL ou um excerto do conteúdo removido;
B. indicação específica da política ou termo violado;
C. forma por meio da qual foi identificada — determinação de autoridades, sistema automatizado, outra exigência legal, sinalização por outros usuários ou sinalização de *trusted flaggers*; e
D. explicação clara do procedimento para recurso.

27 Cf. Santa Clara Principles, santaclaraprinciples.org. Uma atualização do documento é prevista para lançamento em breve, após contribuições de acadêmicos e organizações de outras partes do globo. O InternetLab participou desse processo.

28 Agustina Del Campo et al., "Olhando Al Sur: Rumo a novos consensos regionais em matéria de responsabilidade de intermediários na Internet", 2021, alsur.lat/...

Canal para recorrer da decisão de moderação

Embora algumas plataformas já tenham implementado mecanismos de reconsideração e recurso de decisões de moderação de conteúdo, essa prática ainda é incipiente.

Permitindo que o usuário seja ouvido, a garantia de mecanismo de recurso tem o potencial de subsidiar uma decisão mais informada da plataforma. Assim, ela tem a oportunidade de considerar, por exemplo, contextos, significados, traduções de palavras e acepções específicas de elementos da publicação controversa.

Segundo as diretrizes estabelecidas nos Princípios de Santa Clara, a garantia de mecanismo de recurso deve compreender:

A. revisão humana por alguém não envolvido na decisão original;
B. possibilidade de que o usuário forneça informações adicionais para a apreciação do recurso e
C. declaração por escrito do resultado e razões adotadas na apreciação do recurso.

Informações gerais e compreensíveis sobre políticas e sistemas de moderação aplicáveis

Aperfeiçoar a experiência dos usuários nas plataformas e o *accountability* que elas prestam à sociedade também depende da instituição de critérios para que informem os termos e limites de suas atividades. A exigência de transparência relaciona-se às políticas e aos sistemas de moderação disponíveis. É nesse sentido que se torna imperativo que publiquem suas políticas de conteúdo de forma acessível e clara, em idioma nacional, e notifiquem os usuários em casos de atualização. Isso engloba:

A. tipo de conteúdo e atividades proibidos;
B. providências adotadas para cada violação;
C. critérios utilizados pelos mecanismos de moderação e curadoria, considerando contexto e matizes culturais e idiomáticos;
D. impacto da curadoria de conteúdo em sua visibilidade;
E. critérios para a utilização de moderação humana e automatizada; e
F. quantidade de moderadores alocados para a realização da atividade em nível nacional.[29]

Dados periódicos sobre a aplicação de tais políticas de moderação de conteúdo

Apesar de avanços relativos à transparência a respeito de solicitações de remoção que partem de Estados, as empresas provedoras não publicam relatórios detalhados referentes a providências tomadas em resposta à violação de seus termos de uso. Faltam informações que permitam compreender o cenário em cada país ou região nacional; os dados sobre moderação de conteúdo em geral são apresentados pelas plataformas no agregado mundial.

29 Observacom, *Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta*, 2020, observacom.org/....

Assim, uma outra acepção do chamado por transparência se espelha no princípio Números dos princípios de Santa Clara, que diz respeito aos dados sobre o resultado da aplicação das políticas. Esse princípio se desdobra na imposição de publicação do número total de publicações e contas marcadas (*flagged*) e removidas de forma especificada quanto ao formato - vídeo, áudio, imagem, texto ou *livestream*), à fonte – determinação governamental, algoritmos, outra determinação legal, sinalização por outros usuários ou sinalização de *trusted flaggers* –, e quanto aos locais das remoções.

É crucial que esses relatórios sejam produzidos de acordo com padrões e métricas que permitam compreender as atividades das diferentes plataformas comparativamente.

INDEPENDÊNCIA DE ÓRGÃOS DE FISCALIZAÇÃO

Qualquer que seja o caminho adotado, é crucial que a fiscalização das plataformas seja realizada por órgãos com independência. Isso é fundamental para preservação da liberdade de expressão, considerando a centralidade das plataformas na esfera pública, quando pensamos numa chave de defesa de um ambiente aberto e democrático na internet.

Essa é uma questão institucional intrincada, particularmente dada a falta de estruturas estabelecidas em órgãos públicos que tenham experiência com esses assuntos. Mesmo o rumo adotada no âmbito da União Europeia, que em sua proposta *Digital Services Act* prevê a criação de um novo órgão, pode ser arriscado como modelo. Essa nova entidade, a Coordenadoria de Serviços Digitais de cada país-membro, é concebida nos moldes das autoridades de proteção de dados já existentes. No Brasil, que ainda dá seus primeiros passos na implementação da Lei Geral de Proteção de Dados, essa opção pode ser temerária no momento em que nos encontramos. Basta ter em conta justamente a Autoridade Nacional de Proteção de Dados, concebida como uma autarquia independente, mas criada como um órgão vinculado à Presidência da República — o que suscitou críticas e tem alimentado receios. Assim, é arriscado importar soluções que podem ser apropriadas a outra realidade política institucional, mas não à brasileira — especialmente em um momento tão polarizado e quando o país ainda engatinha na compreensão dos contornos da liberdade de expressão, mesmo em termos de precedentes judiciais. Valeria a pena experimentar com outros modelos, como Códigos de Conduta ou o fortalecimento de espaços multissetoriais, antes que se compreenda a que podem vir, que modelos podem adotar e que funções podem ter autoridades de fiscalização.

CONCLUSÃO

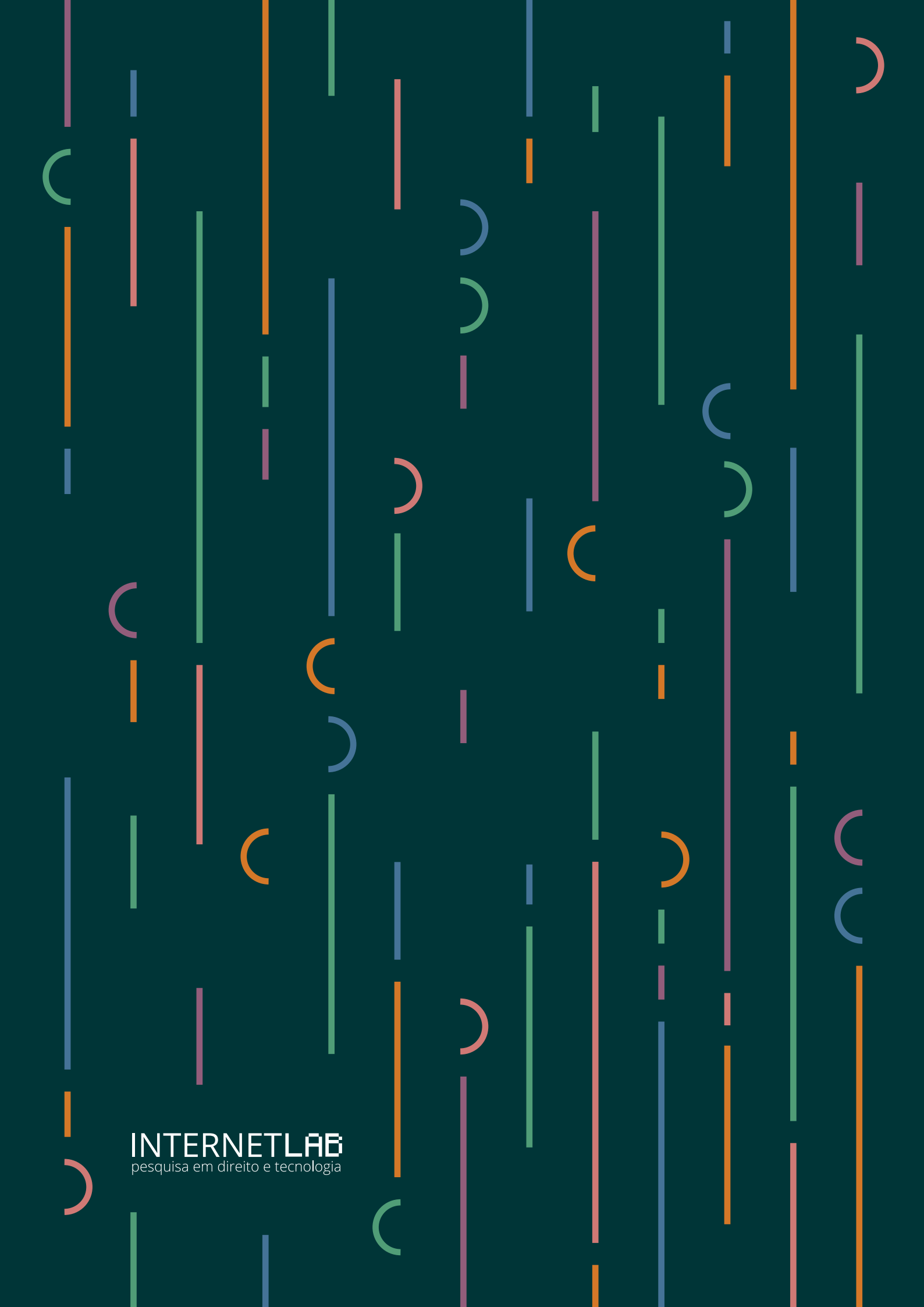
A complexidade do ecossistema de provedores de aplicações de internet joga luz aos gargalos de diversas propostas regulatórias em gestação no Brasil. A internet fez florescer espaços e modelos muito diferentes entre si, tal pluralismo de diferentes espaços deve ser protegido como parte do direito à manifestação do pensamento e do acesso ao conhecimento.

Pensar em uma proposta efetiva e sustentável de regulação a respeito da forma como plataformas digitais gerem a expressão de seus usuários envolve necessariamente uma busca para fazer valer direitos no ambiente digital sem deixar de promover o que nele há de positivo. Inviabilizar sistemas ágeis de moderação de conteúdo vai no sentido contrário dessa busca.

Para ser regulada de maneira protetiva aos direitos fundamentais e compatível com a promoção do pluralismo, a moderação de conteúdo precisa ser considerada em duas dimensões fundamentais. Em primeiro lugar, ela deve ser entendida como atividade fundamental para que provedores de aplicação de internet forneçam serviços e espaços seguros, íntegros e diversos — inclusive na liberdade para que provedores façam escolhas diferentes. Em segundo, devemos entendê-la como atividade que pode importar indevidas restrições a direitos e à livre manifestação do pensamento por parte de usuários, especialmente em casos de erros, abusos e arbítrios.

A partir dessa dupla dimensão, os aspectos mais urgentes da regulação da moderação de conteúdo são os que tornam as políticas e sistemas de grandes plataformas comerciais mais acessíveis e submetidos ao escrutínio público. Duas chaves são centrais para um modelo regulatório atento a isso: direitos dos usuários e defesa de um ambiente aberto e democrático na internet.

Adotar uma abordagem procedimental é seguir nesse caminho evitando a imposição de uma legislação que mais crie problemas do que entregue soluções para as liberdades individuais, a segurança e a autonomia do usuário. Ainda assim, há uma longa caminhada pela frente, que deve levar em conta a realidade social e institucional no Brasil; soluções importadas acriticamente podem se mostrar desajustadas.


INTERNETLAB
pesquisa em direito e tecnologia