ISSN 0034-835X
e-ISSN 2596-0466

Revista de Informação Legislativa

volume 62
246
abril a junho de 2025

SENADO FEDERAL

Por que exigir das plataformas digitais transparência na moderação de conteúdo?

Why Require Digital Platforms to Be Transparent in Content Moderation?

João Pedro Favaretto Salvador[1]

Alexandre Pacheco da Silva[2]

Resumo

Com base em mapeamento da literatura e de proposições legislativas, o artigo teoriza sobre quatro objetivos orientadores de uma regulação que exija transparência das plataformas digitais no exercício da moderação de conteúdo: aprimorar a moderação, permitir que os usuários protejam e exerçam seus direitos, permitir a fiscalização do uso das plataformas como "procuradoras" por autoridades governamentais e promover a confiança e o sentimento de legitimidade dos usuários, governos e anunciantes na atividade da plataforma. Os objetivos são construídos, primeiro, mediante um estudo dos potenciais e limites da transparência como ferramenta regulatória e, segundo, com esteio na reflexão sobre as razões que justificam a intervenção do Poder Público na atividade das plataformas. Com esses referenciais teóricos e exemplos práticos, apresentam-se os objetivos das intervenções de transparência e sugere-se o que se pode exigir das plataformas para que eles sejam cumpridos.

Palavras-chave: transparência; liberdade de expressão; moderação de conteúdo; plataformas digitais; Direito Digital.

Abstract

Based on a mapping of the literature and Brazilian legislative proposals, the article theorizes four guiding objectives for a regulation that requires transparency from digital platforms in the exercise of content moderation: improving moderation, enabling users to protect

1 João Pedro Favaretto Salvador é mestre em Direito pela Universidade de São Paulo, São Paulo, SP, Brasil; doutorando em Direito e Desenvolvimento na Fundação Getulio Vargas (FGV) Direito, São Paulo, SP, Brasil; líder de Projetos no Centro de Ensino e Pesquisa em Inovação da FGV Direito, São Paulo, SP, Brasil. E-mail: joao.salvador.pro@gmail.com

2 Alexandre Pacheco da Silva é doutor em Política Científica e Tecnológica pela Unicamp, Campinas, SP, Brasil; professor de Direito da Fundação Getulio Vargas (FGV) Direito, São Paulo, SP, Brasil; coordenador do Centro de Ensino e Pesquisa em Inovação da FGV Direito, São Paulo, SP, Brasil. E-mail: alexandre.silva@fgv.br

and exercise their rights, allowing government authorities to monitor the use of platforms as "proxies", and promoting trust and a sense of legitimacy among users, governments, and advertisers in the platform's activity. The objectives are constructed, first, through a study of the potential and limits of transparency as a regulatory tool and, second, based on a reflection on the reasons that justify government intervention in the activity of platforms. Using these theoretical references and practical examples, the objectives of transparency interventions are presented and what can be required of platforms to achieve these purposes are suggested.

Keywords: transparency; freedom of expression; content moderation; digital platforms; digital Law.

Recebido em 28/6/24
Aprovado em 28/12/24

DOI: https://doi.org/10.70015/ril_v62_n246_p173

Como citar este artigo: ABNT[3] e APA[4]

1 Introdução

O crescimento das plataformas digitais fez com que acumulassem poderes. Ao atingirem bilhões de usuários em todo o mundo, elas têm impactado não apenas o dia a dia de seus clientes, mas também têm influído diretamente no debate público. Por meio da moderação do conteúdo publicado, disputas entre cidadãos comuns, eleições nacionais e até debates internacionais sobre conflitos armados passaram a ser temperados por decisões de atores privados sobre o fluxo de informação em espaços digitais. Desse acúmulo de poder surgiu a demanda por maior responsabilidade.

Inicialmente opacas (Klonick, 2018, p. 1.665), as decisões das plataformas começaram a fundamentar-se em regulamentos abertos ao público e passíveis de crítica. Porém, ainda que o grau de transparência tenha evoluído muito, tanto a literatura especializada[5]

3 SALVADOR, João Pedro Favaretto; SILVA, Alexandre Pacheco da. Por que exigir das plataformas digitais transparência na moderação de conteúdo? *Revista de Informação Legislativa*: RIL, Brasília, DF, v. 62, n. 246, p. 173-202, abr./jun. 2025. DOI: https://doi.org/10.70015/ril_v62_n246_p173. Disponível em: https://www12.senado.leg.br/ril/edicoes/62/246/ril_v62_n246_p173

4 Salvador, J. P. F., & Silva, A. P. da. (2025). Por que exigir das plataformas digitais transparência na moderação de conteúdo? *Revista de Informação Legislativa: RIL*, 62(246), 173-202. https://doi.org/10.70015/ril_v62_n246_p173

5 Ver Suzor (2019); Hovyadinov (2019); Suzor, West, Quodling e York (2019); Leerssen (2020); Karanicolas (2021); Zornetta (2021).

quanto organizações da sociedade civil[6] consideram-no insuficiente, pois têm sido cada vez maiores os impactos das decisões das plataformas no exercício da livre expressão e na forma do debate democrático.

O resultado dessa percepção foi que a exigência de maior transparência voluntária das plataformas acarretou a necessidade de maior intervenção do Poder Público em suas atividades. No exterior, essa demanda converteu-se em iniciativas legislativas como o *Online Safety Act* (United Kingdom, 2023), do Reino Unido, e o *Digital Services Act* (The Digital [...], 2024), da União Europeia. No Brasil, também cresceu a exigência de intervenções legais de transparência: entre janeiro de 2020 e junho de 2022, deputados e senadores apresentaram 20 projetos de lei (PLs) em que se obrigavam as plataformas digitais a divulgarem informações. O primeiro deles foi o PL nº 2.630/2020 (*Projeto de lei das fake news*)[7].

Este artigo é fruto da pesquisa *Uma taxonomia da transparência na moderação de conteúdo*, realizada entre o início de 2022 e o fim de 2023. Ela não só se concentrou nesse contexto, particularmente em virtude da ausência de consenso a respeito de como se deve organizar o debate sobre transparência na moderação de conteúdo, mas também sistematizou as principais razões e modelos de intervenções de transparência encontrados na literatura especializada e na prática legislativa brasileira. Mapeou-se e examinou-se o conteúdo dos 20 PLs que definiam intervenções de transparência na moderação de conteúdo. Alguns deles são referidos ao longo do texto, especialmente como exemplos das categorias teorizadas pelos pesquisadores. Também se investigou o conteúdo de 53 obras (entre artigos, relatórios e livros) que se propõem discutir aspectos da transparência dentro e fora do contexto da moderação de conteúdo; nem todas elas são aqui mencionadas, mas se podem encontrar em outros produtos da pesquisa[8].

O principal resultado desse processo foi a teorização de quatro objetivos que podem informar e justificar a construção e a interpretação de intervenções de transparência na moderação de conteúdo: a) aprimorar a atividade de moderação; b) permitir que usuários protejam e exerçam seus direitos; c) permitir que autoridades governamentais fiscalizem o uso das plataformas como "procuradoras"; e d) promover a confiança e o sentimento de legitimidade dos usuários, governos e anunciantes na atividade da plataforma. Antes de

---

6 Ver Coalizão Direitos na Rede ([2024]); Kurtz, Carmo e Vieira (2021); Ribeiro, D'Agostini, Sarmento e Rachid (2021); Cruz, Lana e Jost (2023).

7 De início, utilizou-se o Sigalei, um sistema de monitoramento de proposições, com o objetivo de coletar o maior número possível de PLs que criassem obrigações (de transparência ou não) direcionadas às plataformas digitais de rede social. Para isso, foram utilizados termos de busca bastante amplos. Com o emprego das expressões *rede social* ou *redes sociais*, detectaram-se 337 projetos de lei ordinária e de lei complementar, propostos entre 1º/1/2020 e 6/4/2022 (data de realização da busca). Chegou-se a uma base de 20 PLs após a filtragem manual que visou encontrar os PLs que "exigiam que as plataformas revelassem informações sobre seus procedimentos internos de análise, detecção, sancionamento e controle de conteúdo para agentes externos". Dado que nenhum deles alcançou a aprovação até a data de redação deste artigo, é plausível que ainda esteja aberto a contribuições o debate sobre as melhores formas de se regular a moderação de conteúdo. Para mais detalhes sobre a metodologia de coleta e para o acesso à base de projetos de lei encontrados, ver Salvador, Guimarães e Galati (2024).

8 Quanto a pormenores da metodologia de seleção da literatura e para o acesso à lista de obras analisadas, ver Salvador, Galati e Guimarães (2023).

apresentar os resultados, o artigo responde a duas perguntas cujas respostas são necessárias para compreender tanto o raciocínio traçado pelos pesquisadores quanto os objetivos da transparência teorizados.

Na seção 2, examinam-se as contribuições da transparência como técnica de regulação, com o fim de encontrar um referencial teórico para se entenderem as funções da transparência no caso específico da moderação de conteúdo. Com fundamento na literatura, conclui-se que, apesar de potencialmente transformadoras a um baixo custo, as intervenções de transparência têm sua efetividade limitada por variáveis contextuais nem sempre controláveis e que hipóteses de sigilo legítimo têm restringido seu alcance. Por isso, o regulador é desafiado a identificar e controlar certas variáveis para ampliar a probabilidade de sucesso da intervenção, sem, com isso, ultrapassar seus limites legais.

Na seção 3, discute-se sobre o que torna a moderação de conteúdo digna de intervenções de ordem pública, em resposta a uma suposta impertinência de intervenções do Estado numa atividade, a princípio, de natureza privada. A resposta, que se torna pressuposto dos resultados apresentados, é que a moderação de conteúdo, embora fundada legalmente numa relação contratual, tem efeitos que transcendem essa relação, e pode apresentar-se tanto como fonte de riscos quanto como ferramenta de proteção para o debate público. Além disso, os erros e a opacidade na moderação podem levar a decisões problemáticas que refreiam injustificadamente o exercício de liberdades fundamentais pelos usuários. Essas características criam a demanda por intervenções do Estado destinadas a proteger o debate público e os direitos dos usuários de decisões problemáticas das plataformas e do conteúdo publicado por outros usuários.

A seção 4 organiza-se em torno dos quatro objetivos da transparência na moderação de conteúdo. Apresentam-se as principais características de cada um deles e sua relação com os fundamentos teóricos da transparência expostos na seção 2 e com os efeitos da moderação descritos na seção 3. Para cada objetivo o artigo sugere categorias de informação – que, se divulgadas, têm maior potencial de atingir resultados efetivos – e apresenta exemplos práticos para se compreender melhor a questão.

A conclusão do artigo aponta temas complexos que merecem pesquisas mais aprofundadas para que a qualidade das intervenções de transparência continue a evoluir. Destaca-se o uso cada vez mais comum de ferramentas de detecção automática de conteúdo, cuja implantação – como resposta para o volume de publicações que precisam ser analisadas pelas plataformas – cria uma série de novas dificuldades para se efetivar o ideal de transparência e de motivação das decisões.

2 O potencial e os limites da transparência

O vocábulo *transparência* costuma ser entendido como o contrário de *opacidade*; remete, pois, às noções de clareza e de visibilidade. Contudo, *transparência* significa também

ausência de sigilo. Assim, uma organização ou uma pessoa são perfeitamente *transparentes* quando não mantêm informações em sigilo: elas podem ser acessadas por outras pessoas ou organizações. Uma intervenção legal que promove transparência, portanto, visa tornar acessível por meio de obrigações legais uma informação anteriormente sigilosa, tais como os assuntos internos de uma organização. Com isso, espera-se que a informação revelada permita ao destinatário tomar decisões informadas ou estimule mudanças socialmente positivas no comportamento do alvo da intervenção.

De início, a ideia de transparência como um caminho para a promoção de mudanças concretas fortaleceu-se como movimento político nos EUA na primeira metade do século XX. Suas primeiras grandes vitórias foram a aprovação do *Freedom of Information Act*[9] em 1966 e sua posterior atualização em 1974, como resposta a escândalos de corrupção como o caso Watergate (Birchall, 2011, p. 11). À época, a ideia de *transparência* ligava-se intimamente à prestação de contas pelo Poder Público. Grupos políticos e organizações da sociedade civil que defendiam o avanço da transparência propagavam um ideal: a publicidade dos atos previne a corrupção, que é quase inevitável em processos sigilosos[10].

Segundo Zuccolotto e Teixeira (2019, p. 13), numa democracia representativa o povo delega poder ao representante, mas conserva o poder de destituí-lo. Esse poder pode ser exercido apenas se houver instrumentos de controle do poder político que permitam aos representados tanto avaliar se os governantes estão agindo como representantes de fato quanto, de algum modo, decidir com base nessa avaliação. A transparência, na forma da publicidade dos atos públicos, desponta então como condição para o devido funcionamento desses instrumentos, já que é impossível avaliar a qualidade de atos sigilosos. Além disso, a correção desses problemas numa democracia teria como pressuposto o debate, a deliberação e a troca de argumentos informados sobre as diversas opções e caminhos para a ação, o que também só seria possível sem graves assimetrias de informações entre governantes e governados (Zapatero Gómez, 2019, p. 110).

Há mérito, pois, na premissa de que a transparência poderia, em algum grau, prevenir o abuso no exercício do Poder Público: os cidadãos estariam munidos de informação para: a) exigir prestação de contas pelo governo, seja por seus gastos, seja por seu desempenho no oferecimento de serviços públicos (o que, em tese, reduziria a corrupção e ineficiência); e b) participar do debate público sobre o futuro de sua nação. A transparência é necessária tanto para a supervisão dos atos administrativos quanto para a tomada coletiva de decisão; ambas são condições para a dissuasão de atos ineficientes e mal-intencionados. Ao longo

---

9 O *Freedom of Information Act* exige que as agências governamentais norte-americanas divulguem diversas categorias de documentos e informações, além de exceções a esse dever de transparência. O texto foi alterado diversas vezes desde sua concepção (United States, 2022).

10 Durante esse período, popularizou-se a frase "a luz do sol é um dos melhores desinfetantes; a luz elétrica, o mais eficiente policial", cunhada em 1914 por Louis Brandeis, então advogado e posteriormente juiz da Suprema Corte norte-americana. A frase está no ensaio "What Publicity Can Do", publicado no livro *Other People's Money and How the Bankers Use It* (Brandeis, 1914, p. 92): "Sunlight is said to be the best of disinfectants; electric light the most efficient policeman".

do tempo, com a popularização da transparência como ferramenta de controle do Poder Público, também se fortaleceram argumentos a favor de intervenções de transparência direcionadas a agentes privados. Aqui também fincou raízes o conceito de que a divulgação de informação é capaz de provocar impactos sociais positivos.

Zapatero Gómez (2019, p. 110) apresenta duas explicações para o surgimento e a popularização dessa abordagem. Em primeiro lugar, a correção de assimetrias de informação pelo Estado seria relevante para garantir a eficiência do mercado e a proteção de interesses dos consumidores; para o autor, esses dois objetivos só podem ser alcançados se o consumidor estiver munido de informações claras, acessíveis e suficientes sobre produtos e serviços. Em diversos casos, há incentivo econômico para que as empresas divulguem informação sobre seus produtos para os consumidores; em outros casos, porém, o custo de revelar a informação pode ser maior que os benefícios, o que justifica a regulação com o fim de garantir a divulgação de dados importantes[11]. Em segundo lugar, se comparado a outras formas de regulação, não é muito alto o custo de estabelecer essas obrigações para o governo e para os governados (Zapatero Gómez, 2019, p. 112). Exigir mais informação é uma boa alternativa para intervir em processos internos ou nas características do produto oferecido aos consumidores, quando o risco do produto é moderado ou baixo ou quando as empresas não são motivadas de modo suficiente para voluntariamente divulgar a informação necessária. Segundo o autor, a obrigação de transparência contribuiria para melhorar a qualidade dos produtos e proteger o consumidor sem impactos profundos sobre a concorrência e a inovação, o que lhe permitiria fazer escolhas informadas sobre o serviço que contrata ou sobre o produto que adquire.

Assim, todas as intervenções de transparência, tanto no setor público quanto no privado, fundamentam-se no ideal bastante convincente de que a divulgação de informações sigilosas é capaz de iniciar processos que levam a mudanças positivas no comportamento de seus usuários e alvos. Essas mudanças podem ser tanto a melhoria de produtos e de serviços disponíveis no mercado quanto a inibição de atos administrativos ilícitos e ineficientes. Todavia, não é óbvio o nexo causal entre a divulgação da informação e a ocorrência das mudanças concretas desejadas, nem ocorre efetivamente em toda intervenção de transparência. Muitos dentre os que exigem mais transparência acreditam em que a possibilidade de observar um sistema ou uma instituição é suficiente por si só para haver controle e compreensão, o que legitimaria qualquer intervenção nesse sentido. A verdade, porém, é que a transparência não implica controle por si mesma (Ananny; Crawford, 2018, p. 975). Embora tenha grande potencial, a transparência também apresenta limites, pois existem

11 Um exemplo é a divulgação de informações em rótulos de produtos alimentícios. É comum que as empresas que os produzem e distribuem sejam obrigadas a divulgar nos rótulos informações não particularmente convidativas, mas consideradas essenciais para o consumidor tomar decisões informadas sobre sua saúde; é o caso de quando o rótulo alerta sobre riscos à saúde por adição excessiva de sódio, açúcar ou gorduras saturadas. Não se espera que os produtores divulguem essas informações sem que tenham um incentivo externo, pois elas podem afastar o consumidor, o que representa um custo contrário aos interesses empresariais.

variáveis que influenciam no sucesso de uma intervenção de transparência que promova mudanças no comportamento dos agentes obrigados a divulgar informação.

Apesar do baixo custo, se concebidas sem atenção às variáveis adequadas, intervenções de transparência também se podem tornar empecilhos para a atividade de seu alvo ou até prejudicar o sigilo legitimo de indivíduos. Nesses casos, por mais convenientes que pareçam, elas podem falhar e causar mais problemas.

2.1 Nem toda intervenção de transparência é efetiva

Tornar disponíveis grandes quantidades de informação pode ser pouco útil e até prejudicial para se atingirem as finalidades de uma intervenção de transparência. Quando ausentes determinadas condições sociais e técnicas ideais, pode ser ineficaz a visibilidade de uma intervenção. Uma intervenção de transparência é efetiva apenas quando realmente tem o potencial de gerar as mudanças pretendidas, algo que os estudiosos costumam chamar de transferência significativa (*meaningful transparency*) (Hovyadinov, 2019; Suzor; West; Quodling; York, 2019; Karanicolas, 2021). Esse potencial depende tanto de características da intervenção – *o quê*, *para quem* e *de que forma* é divulgado – quanto de condições contextuais, como a propensão dos usuários da transparência (os que recebem a informação) a serem motivados a agir pela informação revelada e a predisposição dos alvos da transparência (os que são obrigados a divulgar a informação) para mudarem seu comportamento.

Com o fim de sistematizar essas variáveis, Fung, Graham e Weil (2007, p. 51) concebem o chamado *ciclo da ação*, em que a informação difundida por uma intervenção de transparência gera efeitos positivos sobre o comportamento de seu alvo – normalmente um provedor de um serviço, público ou privado. Por meio do mapeamento de intervenções de transparência bem-sucedidas, os autores identificaram que: a) a informação tornada disponível é útil e acessível a pelo menos um grupo de usuários da informação insatisfeito com a prestação do serviço; b) essa informação leva os usuários a mudarem suas decisões e ações, seja para buscarem prestadores de serviços concorrentes, seja para se mobilizarem por mudanças; c) essas novas ações acarretam consequências significativas para o prestador; e d) os prestadores respondem de forma construtiva a essas consequências e buscam a resolução do problema.

Para ilustrar o ciclo da ação, Kosack e Fung (2014, p. 70-71) dão o exemplo da informação divulgada por prestadores de serviço de saúde pública. Os autores apontam que cidadãos com acesso a informações sobre a prestação de serviços de saúde são capazes de comparar a qualidade de prestadores, como os hospitais e as unidades de pronto atendimento, e exigir mudanças com base nisso. Dados úteis para esse tipo de comparação são: a) a disponibilidade de recursos e medicamentos; b) a duração das filas de espera; e c) a abrangência dos procedimentos oferecidos. Com esses dados, os destinatários da informação podem identificar o prestador que mais se adéqua a seus interesses, avaliar a eficiência da alocação de recursos públicos ou mesmo identificar as tecnologias e modalidades de

prestação de serviços disponíveis. Ao encontrarem problemas ou pontos de discordância nessas informações, podem demandar melhorias. Contudo, para que essas demandas resultem em mudanças concretas, os cidadãos devem pressionar os prestadores que têm agido de forma ineficiente – seja optando apenas por prestadores de alta qualidade, seja engajando-se em protesto, seja utilizando canais de sugestão ou crítica pré-estabelecidos. Se estiverem propensos a responder de forma construtiva, os prestadores ineficientes ou os órgãos públicos responsáveis por sua administração podem utilizar esse feedback para buscar melhorias em seu desempenho, ao realocar o uso de recursos para reduzir filas, garantir a disponibilidade de medicamentos ou ampliar os procedimentos oferecidos.

Na justificação das intervenções de transparência, o regulador pode influenciar as variáveis do ciclo, como a tendência dos usuários de agirem baseados na informação recebida e a dos prestadores de serviços de responderem construtivamente a aspectos relacionados à qualidade da informação e à forma de sua divulgação. A predisposição dos usuários para agirem pode ser aumentada se o regulador exigir a criação de canais de reclamação que reduzam os custos da ação. E, com base no estabelecimento de sanções à inação, é possível ampliar a inclinação dos prestadores para responderem de forma positiva.

Porém, trata-se de técnicas regulatórias posteriores à exigência de divulgação de informação. Mais importantes para o enfoque deste artigo são as duas variáveis que promovem a efetividade de uma intervenção de transparência e podem ser influenciadas quando se exige a divulgação dos dados: a visibilidade e a capacidade de inferência da informação. Zuccolotto e Teixeira (2019, p. 52) argumentam que a informação adquire visibilidade ao ser divulgada de forma completa, acessível e com relativa facilidade. O regulador garante visibilidade ao determinar o grau de pormenorização dos dados a serem divulgados e o meio cabível para essa divulgação. Assim, a visibilidade é condição para a efetiva intervenção de transparência, pois, se a informação não estiver facilmente acessível e completa, é improvável que seja capaz de persuadir seus usuários a agirem para fomentar mudanças no comportamento do prestador de serviços.

No caso do exemplo anterior, se apenas parte das unidades de pronto atendimento e hospitais públicos divulga dados sobre a disponibilidade de medicamentos, torna-se impossível comparar a qualidade dos prestadores de serviços de saúde, o que inviabiliza a ação fundamentada dos usuários desses serviços. Da mesma forma, se os dados forem divulgados, mas estiverem acessíveis apenas mediante procedimentos burocráticos complexos, é improvável que motivem os usuários à ação. Às vezes, mesmo completa e totalmente acessível, a informação pode estar disponível de um modo que torna difícil ao usuário chegar a conclusões úteis com base na observação. Nessa situação, diz-se que a informação tem capacidade de inferência limitada (Zuccolotto; Teixeira, 2019, p. 54). Assim, o regulador pode exigir que os prestadores de serviço divulguem informações em formatos variados que influam diretamente na capacidade de inferência dos usuários.

Em vez de tornar disponíveis os dados brutos, frequentemente em formatos distintos para cada instituição, o regulador pode facilitar o exercício de comparar hospitais públicos

e exigir que se divulguem os dados em formato de tabela comparativa elaborada por especialistas externos ao alvo da intervenção. Nesses casos, as possíveis inferências tornam-se mais acessíveis ao público não especializado, o que propicia a tomada de ação[12]. Desse modo, conquanto não tenha controle sobre a propensão dos usuários ou do alvo da transparência à mudança de comportamento, o regulador é capaz de manejar a qualidade e a forma da divulgação dos dados a fim de garantir que sejam visíveis e passíveis de interpretação. Se exige a disponibilidade dos dados corretos, da forma correta e para os usuários corretos, o regulador pode criar um cenário fértil para a proteção dos interesses dos usuários e o controle de seus alvos.

Não obstante, tais decisões só podem ser tomadas de forma adequada se o regulador tiver consciência dos efeitos pretendidos ou, no mínimo, esperados da intervenção. Sem um ciclo de ação pré-concebido, a mera divulgação de informação tem pouca probabilidade de ser efetiva; dessa maneira, a transparência não deve ser tratada como fim, mas como meio.

2.2 Nem toda intervenção de transparência é positiva

O movimento político por transparência ganhou força ao responsabilizar o sigilo pela ineficiência e pela corrupção; em regra, passou-se a considerá-lo problemático. Contudo, existem circunstâncias nos âmbitos privado e público em que o sigilo é desejável e mesmo valorizado numa democracia. É essencial que o regulador leve em conta esses limites, visto que suas intervenções de transparência podem causar mais prejuízos que benefícios ao exigirem a divulgação de informações cujo sigilo é justificado.

O valor do sigilo prepondera com mais frequência no âmbito privado que no público. Por exemplo, a efetivação da privacidade – ideal constitucionalmente protegido – pressupõe ser possível ao cidadão manter em sigilo aspectos de sua vida privada. Sem que algum grau de sigilo seja considerado legítimo, o sujeito da democracia pode ver tolhido aquilo que o torna singular: a sua autonomia (Birchall, 2011, p. 12). Um regime que promove a transparência não deve ir longe demais a ponto de constringir o sigilo legítimo mediante uma política de vigilância das atividades privadas.

Além da privacidade do indivíduo, as atividades do mercado também pressupõem um grau legítimo de sigilo. A livre concorrência depende de decisões estratégicas relacionadas a informações sobre o desenvolvimento de produtos e serviços que devem ser compartilhadas, patenteadas ou mantidas em segredo para a maximização dos lucros (Birchall, 2011, p. 13). Diferentemente do caso da Administração Pública – a respeito da qual existe a expectativa de que seja a mais transparente possível para viabilizar seu controle por vias

12 Em contrapartida, ainda que menos acessíveis *a priori*, dados brutos podem evitar a interferência de vieses de terceiros e a manipulação na interpretação da informação. Além disso, se oferecidas em formato propício e padronizado, grandes quantidades de dados brutos podem ser interpretadas mediante a tecnologia da informação. Existe, portanto, um equilíbrio a ser alcançado entre a autonomia do usuário da informação e a capacidade de inferência dos dados.

democráticas –, as empresas privadas em geral não têm o dever de divulgar informações sobre suas práticas internas; por razões concorrenciais e de privacidade, muitas dessas informações costumam ser protegidas pelo Direito como segredos de negócio[13] e dados pessoais[14]. Nos casos em que se exige transparência das empresas privadas, torna-se mais complexo fundamentar a intervenção: no âmbito público, a transparência é regra e o sigilo deve ser justificado; no âmbito privado, costuma valer o contrário.

No campo da Administração Pública, contudo, também existem situações em que a necessidade de transparência deve ser contrabalançada com o valor do sigilo. É o caso da atividade das autoridades de investigação e segurança pública: elas também devem ser controláveis, mas a transparência completa de suas atividades pode inviabilizar a efetivação de seus fins. Técnicas de investigação como a interceptação telefônica e a infiltração, bem como outras relativamente comuns em inquéritos criminais, pressupõem o desconhecimento do investigado a respeito de sua vigilância. Assim, por mais que o sigilo justifique o uso dessas técnicas pelas autoridades, sua transparência deve ser limitada para não inviabilizar investigações. Nesses casos, deve-se tanto preservar ao máximo o sigilo quanto buscar outras ferramentas de controle, tal como a exigência de autorização judicial prévia e de divulgação *a posteriori* das informações sigilosas[15].

2.3 Elaborar intervenções de transparência não é esforço trivial

Intervenções de transparência podem ser pouco custosas e muito transformadoras; por isso, frequentemente são consideradas positivas e efetivas por si mesmas. No entanto, como se afirmou, esse ideal de transparência encontra limites nas condições práticas para a sua implantação e nas hipóteses em que o sigilo é legítimo e desejável. Reconhecer a existência desses limites não inviabiliza o uso das intervenções de transparência, mas atribui ao regulador um ônus significativo de planejamento e justificação.

---

13 O art. 195 da Lei nº 9.279/1996 (*Lei de propriedade industrial*) reconhece a legitimidade dos segredos de negócio e protege-os ao firmar que comete crime de concorrência desleal quem "divulga, explora ou utiliza-se, sem autorização, de conhecimentos, informações ou dados confidenciais, utilizáveis na indústria, comércio ou prestação de serviços, excluídos aqueles que sejam de conhecimento público ou que sejam evidentes para um técnico no assunto, a que teve acesso mediante relação contratual ou empregatícia, mesmo após o término do contrato" (Brasil, [2024]).

14 De acordo com o art. 6º da Lei nº 13.709/2018 (*Lei geral de proteção de dados*), a empresa tem o dever de proteger e manter em sigilo dados pessoais de seus clientes durante seu tratamento: "As atividades de tratamento de dados pessoais deverão observar a boa-fé e os seguintes princípios: [...] VII – segurança: utilização de medidas técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou difusão" (Brasil, [2022b]).

15 Para o emprego da interceptação telefônica, a Lei nº 9.296/1996 não só exige que a autorização judicial seja prévia e fundamentada, mas também determina que as diligências devam ocorrer em autos apartados até o momento em que seus resultados estejam prontos para o relato. Assim, é possível garantir o sigilo e a efetividade da investigação até ele se tornar desnecessário e voltar a prevalecer a lógica de transparência: "Art. 8º A interceptação de comunicação telefônica, de qualquer natureza, ocorrerá em autos apartados, apensados aos autos do inquérito policial ou do processo criminal, preservando-se o sigilo das diligências, gravações e transcrições respectivas" (Brasil, [2019]).

Para que as intervenções sejam efetivas, o regulador deve ter consciência das mudanças concretas que pretende incentivar, presumir como usuários e alvos da transparência se comportarão dentro do ciclo de ação e, com esses fundamentos, decidir sobre que informações devem ser divulgadas e qual será a forma de sua divulgação. Em outras palavras, a chance de sucesso de uma intervenção de transparência será muito pequena se o regulador não tiver em mente os objetivos dessa intervenção.

Ao mesmo tempo, o regulador deve considerar se a divulgação de determinadas informações pode ser problemática diante das situações em que o sigilo é desejável. A busca de um equilíbrio nesse sentido será mais trabalhosa na elaboração de intervenções de transparência sobre atores privados do que sobre os públicos, já que o âmbito privado tem a autonomia informacional como regra. Desse modo, a criação de obrigações de transparência para organizações privadas não é algo trivial: sempre que se obriga uma empresa a divulgar uma informação sobre suas práticas internas, essa obrigação deve cumprir uma finalidade legítima e não provocar mais prejuízos que benefícios. Só se devem divulgar os dados realmente necessários para que ocorram os efeitos visados pelo regulador.

3 A relevância pública da moderação de conteúdo

A moderação de conteúdo tem recebido atenção particular quando realizada por plataformas digitais. Trata-se de condutas que elas adotam para garantir o cumprimento das regras de comportamento que elas mesmas estabelecem em seus contratos de adesão (como termos de uso e políticas da comunidade) ou que são levadas a aplicar em razão de legislação e de ordens judiciais. Assume-se neste artigo uma concepção ampla da *moderação de conteúdo*; conforme define Grimmelmann (2015, p. 47, tradução nossa), ela compreende todos "os mecanismos de governança que estruturam a participação em comunidade para facilitar a cooperação e prevenir abuso". Desse modo, aqui se considera que a moderação de conteúdo transcende as sanções explicitamente direcionadas ao conteúdo que viola as regras das plataformas, como a remoção de uma publicação ou a suspensão de uma conta; ela inclui a ampliação e a redução do alcance de uma publicação com base num juízo sobre sua qualidade.

A moderação de conteúdo é uma atividade que tem fundamento legal num contrato entre a plataforma e o usuário, no qual este reconhece e aceita que sua livre expressão se sujeite a restrições firmadas por aquele. Com efeito, há uma inquietação sobre a pertinência e a legitimidade de se interferir nessa atividade eminentemente privada. Em outras palavras: por que o Estado deve interferir no que se apresenta, a princípio, como manifestação de vontade dos envolvidos (Nunziato, 2018)? A resposta a essa pergunta condiciona os fins da intervenção, que devem ser legítimos, justificados e discutidos antes de reflexões mais aprofundadas.

No caso particular da moderação de conteúdo, a justificativa para a intervenção assenta-se em particular nos efeitos dessa atividade sobre o debate público[16] e sobre os usuários diretamente afetados. As plataformas digitais propõem-se organizar e fazer escolhas sobre como o conteúdo de seus usuários é transportado e tornado disponível a outros (Gillespie, 2018, p. 81). Elas o fazem para oferecer-lhes uma audiência que não teriam se utilizassem tecnologias de comunicação tradicionais. As plataformas, pois, não são meras distribuidoras de informação, como as operadoras de telefonia: elas moldam o espaço de comunicação ao observarem o comportamento dos usuários com o fim de garantir seu engajamento e, consequentemente, seu contato com anúncios – fonte de parcela significativa da sua receita[17]. A moderação de conteúdo é justamente a ferramenta utilizada para delinear esse espaço de comunicação – e, como afirma Grimmelmann (2015, p. 47, tradução nossa), "facilitar a cooperação e prevenir o abuso".

Entretanto, a enorme quantidade de usuários significa que as plataformas, ao modelarem ativamente o espaço, também têm exercido influência considerável sobre a opinião pública. Ao tomarem decisões sobre que tipo de conteúdo terá mais ou menos alcance em vista das regras que criam, elas têm intervindo sobre o que pode e o que não pode ser dito também em temas sensíveis, como eleições, discriminação e conflitos culturais. Além disso, o fato de atores mal-intencionados propagarem nas plataformas conteúdo prejudicial ao debate público – como a desinformação e as mensagens de ódio – faz com que a moderação de conteúdo seja tratada como a primeira linha de defesa contra tais discursos.

Essa configuração cria um forte interesse das autoridades públicas na forma como é exercida essa moderação. Pode ser uma poderosa aliada do Estado na efetivação de direitos fundamentais e dos pressupostos do debate democrático uma plataforma que modere de forma eficiente e com o mínimo possível de erros com o fim de proteger o debate público de conteúdos danosos. Por outro lado, uma plataforma que modere conteúdo de forma inconsistente e opaca – apenas para garantir o engajamento dos usuários, independentemente de potenciais malefícios ao debate público – pode tornar-se alvo de regulações que tencionam evitar a proliferação desses danos.

Assim, certos efeitos da moderação sobre os afetados diretamente por ela também se podem tornar um fundamento para intervenções de ordem pública. Não se trata necessariamente do caso do usuário que age em contrariedade com os regulamentos das plataformas e é sancionado conforme preveem esses regulamentos. Isso poderia ser defendido como uma dinâmica típica de uma relação contratual, apesar de ser questionável a discricionariedade das plataformas para a elaboração das regras– como apontam Mendes e Fernandes (2022, p. 43). Fala-se, sobretudo, dos casos em que, devido a uma assimetria de informação, o usuário não entende por que foi sancionado ou, pior ainda, nem sequer compreende as

16 Na mesma linha, o Tribunal Superior Eleitoral costuma utilizar-se do argumento da preservação da saúde do "ambiente informacional" para justificar intervenções na atividade de moderação de conteúdo. Ver Osorio, Alvim, Siqueira, Barcelos, Vargas, Rodrigues e Rondon (2022).

17 Em 2020, praticamente toda a receita do Facebook proveio da publicidade de outras empresas. Ver Statista (2024).

regras com que, em tese, concordou. A isso se acrescem os casos problemáticos em que plataforma simplesmente erra ao sancionar usuários que não agiram irregularmente.

A relevância desses casos (abordados mais detidamente na subseção 4.2) está no impacto que a moderação pode ter sobre os direitos dos usuários afetados por decisões injustas. Embora as plataformas só detenham poder sobre as ações do usuário nos espaços digitais que elas controlam, a restrição do acesso a eles pode acarretar danos que alcançam outros aspectos da vida privada (West, 2018). Assim, quando atingidos pelas sanções mais graves aplicadas pelas plataformas, muitos usuários lamentam a perda de fotos e de outros documentos de valor sentimental, bem como a privação da capacidade de comunicar-se com amigos e familiares. Diversos usuários que dependem de formas digitais de remuneração (como propaganda) alegam dano à sua vida profissional quando são considerados infratores das regras da plataforma. Sem acesso às plataformas, eles também perdem o acesso à sua comunidade familiar, profissional, econômica e cultural.

A participação nesses espaços virtuais de comunicação tem sido um instrumento notável para o exercício pleno das liberdades individuais e políticas, já que condensam grande parte do debate público e cumprem papel significativo, especialmente em períodos eleitorais (Mendes; Fernandes, 2022, p. 43). A exclusão injustificada de usuários desses espaços, então, significa também uma restrição ao acesso a informações, opiniões e dados essenciais à plena participação política. Quando são consequência de uma relação contratual sem assimetria de informação entre usuário e plataforma, é até plausível defender que esses efeitos danosos não justificam uma intervenção de ordem pública. Pelo menos em tese, as plataformas têm uma margem de discricionariedade para tomar decisões sobre que conteúdo deve ou não circular em seus espaços de comunicação, o que não é necessariamente problemático quando são transparentes essas regras. No entanto, decisões injustificadas, erradas e opacas das plataformas podem afetar as liberdades fundamentais dos usuários, o que leva ao interesse das autoridades públicas em protegê-los dessas decisões, seja ao corrigirem assimetrias de informação, seja ao exigirem das plataformas maior precisão na moderação de conteúdo.

Desse modo, apesar de seu fundamento numa relação contratual entre a plataforma e o usuário, a moderação de conteúdo: a) tem efeitos que excedem o que é estabelecido entre as partes, de forma que seu direcionamento afeta o debate público democrático; e b) pode ser exercida de maneira viciada e acarretar danos injustificados aos direitos individuais e públicos dos alvos de decisões problemáticas.

4 Objetivos da transparência na moderação de conteúdo

Realizam-se intervenções de transparência com o fim de mudar o comportamento de seu alvo e dos usuários da informação divulgada, dada a pressuposição de que existe alguma assimetria de informação entre eles. O regulador busca, com isso, fornecer ao

público informações para tomadas de decisão socialmente benéficas. Problemas comuns a serem resolvidos mediante obrigações de transparência, por exemplo, são: a) os riscos resultantes da assimetria de informação; b) a baixa qualidade da prestação de serviços críticos providos por uma organização; c) a perpetuação de discriminação e desigualdades sociais devida à assimetria de informação; e d) a prevalência da corrupção em instituições que servem ao interesse público (Zapatero Gómez, 2019, p. 110).

Os objetivos de uma intervenção de transparência devem ser o mais claros possível, tanto no início quanto no final do processo regulatório. Isso ocorre porque: a) não será possível definir adequadamente as informações a serem divulgadas, nem como ou para quem elas devem ser divulgadas, se não houver um objetivo claro traçado no início da elaboração de uma intervenção de transparência; b) no objetivo de uma intervenção de transparência está a fonte de sua justificação, e é essencial haver uma finalidade explícita para avaliar se se justifica uma intervenção; e c) o objetivo de uma intervenção de transparência pode subsidiar a interpretação quando a legislação não é suficientemente clara a respeito dos limites dessa intervenção nas hipóteses de sigilo legítimo.

A primeira dessas razões foi comentada na subseção 2.1: os objetivos da transparência determinam as variáveis que condicionam sua efetividade. A segunda razão pode ser ilustrada com o caso das plataformas; se intentam corrigir um desequilíbrio que causa problemas específicos, as intervenções devem ter como alvos as organizações tidas como responsáveis por criarem um risco ao interesse público. Como se observou, é cada vez mais comum entidades do setor privado estarem sujeitas a exigências de transparência semelhantes às do setor público, o que ocorre justamente quando a atividade dessas entidades privadas se intersecciona com interesses públicos (Karanicolas, 2021, p. 62). Plataformas de redes sociais são administradas por empresas privadas e, por isso, não se lhes aplica a noção de que a transparência é regra e o sigilo, exceção. Qualquer intervenção de transparência direcionada a essas plataformas deve fundamentar-se na ideia de que a divulgação de informações anteriormente sigilosas tem o potencial de evitar dano ou risco ao interesse público relacionado à atividade dessa plataforma. Além disso, em razão das hipóteses de sigilo legítimo comuns à atividade empresarial privada, essas intervenções devem limitar-se a divulgar informação necessária para cumprir esse objetivo de proteger o interesse público. Assim, dois exercícios de justificação são necessários para se legitimarem intervenções de transparência: o regulador deve demonstrar não só que a moderação é prejudicial ao interesse público mas também que, de alguma forma, a intervenção proposta limita esse prejuízo. Na ausência de clareza sobre os objetivos de uma intervenção, é impossível avaliar se o regulador considerou esses exercícios. Em outras palavras: não se pode verificar se essa intervenção é legítima em face dos limites estabelecidos pela autonomia privada.

E a terceira razão decorre das dificuldades práticas de se proceder a uma intervenção de transparência. Dentre os problemas cruciais na sua elaboração está o de identificar que informação se deve revelar, como e para quem – e eles só podem ser solucionados por uma decisão sobre a finalidade da intervenção. Não é possível decidir sobre *o quê* sem decidir

sobre *para quê* é suficiente. Esse desafio reflete-se em particular na concepção dos limites da intervenção: são as hipóteses em que o alvo pode recusar-se a divulgar algo, seja porque a informação é desnecessária para atingir os objetivos propostos, seja por existirem razões legítimas para a manutenção do sigilo. Porém, dada a dificuldade de se traçarem esses limites para todos os casos específicos, é comum que o legislador utilize linguagem aberta para definir essas hipóteses e atribua ao aplicador da regra a tarefa de estabelecer com precisão o que o alvo da intervenção tem a obrigação de divulgar em decorrência da lei. Um exemplo dessa prática observa-se no PL nº 2.630/2020[18]. Dentre as determinações relacionadas ao tema, o art. 21 do projeto define o conteúdo a ser divulgado pela plataforma em seus termos de uso:

> Art. 21. Os termos de uso dos provedores devem conter os parâmetros utilizados nos seus sistemas de recomendação de conteúdo, *ressalvados os segredos comercial e industrial*, bem como:
>
> I – descrição geral dos algoritmos utilizados;
>
> II – destaque para os principais parâmetros que determinam a recomendação ou direcionamento de conteúdo ao usuário; e
>
> III – opções disponíveis aos usuários para modificar os parâmetros de recomendação ou direcionamento (Brasil, 2023, grifo nosso).

Conquanto tenha proposto um conteúdo mínimo para os termos de uso divulgados pelas plataformas, o legislador estabeleceu que o alvo da intervenção não é obrigado a fornecer essas informações se consideradas segredos comerciais ou industriais. Contudo, não existe definição legal de *segredo comercial e industrial*; desse modo, o limite exato da intervenção não é claro para o alvo, nem para os usuários, nem para o aplicador da norma[19]. A interpretação da norma, nesse caso, não pode torná-la inefetiva; a obrigação de transparência em questão não pode ser ampla demais para não prejudicar os segredos comerciais e industriais da plataforma. E o contrário também é verdade: se a noção de *segredo comercial e industrial* for considerada excessivamente ampla, a norma que impõe a obrigação de transparência perde sua razão de ser, pois não se fornecem informações suficientes para a realização de seu objetivo.

Por conseguinte, a solução dessa incerteza começa por uma questão: que resultado se espera da divulgação dos parâmetros utilizados pelas plataformas em seus sistemas de recomendação de conteúdo? Depois de se responder a essa pergunta, é possível indagar sobre

---

18 A referência aqui é à versão mais avançada do PL nº 2.630/2020. Ele pode ser acessado no site da Câmara dos Deputados e estava em tramitação quando se concluiu este artigo. Trata-se do Parecer Preliminar de Plenário nº 1, publicado em 27/4/2023 pelo deputado Orlando Silva (PCdoB-SP), relator do projeto (Brasil, 2023).

19 O aplicador da norma pode ser um órgão do Judiciário ou uma autoridade administrativa supervisora, a depender da configuração final do projeto.

quão detalhadas devem ser essas informações ou sobre que dados devem ser divulgados para que esse resultado seja atingido. Se for capaz de responder a esses questionamentos, o intérprete pode encontrar soluções que não só promovam a ocorrência de mudanças concretas e um ciclo da ação de sucesso, mas também preservem ao máximo o sigilo legítimo das plataformas.

Por três motivos – um de orientação, um de legitimação e outro de interpretação –, é essencial que o debate sobre como devem ser conduzidas as intervenções de transparência na moderação de conteúdo suponha compreender suas finalidades. A fim de contribuir para esse debate, apresentam-se exemplos de quatro principais objetivos buscados com essas intervenções: a) aprimorar a atividade de moderação, tanto ao permitir colaboração e opinião de *stakeholders* (como pesquisadores e organizações da sociedade civil), quanto ao gerar fiscalização, incentivos e pressão para a mudança de comportamento das plataformas; b) permitir que usuários protejam e exerçam seus direitos de manifestação ao compreenderem as regras da moderação e as ferramentas de recurso e correção de erros; c) fiscalizar o uso das plataformas como "procuradoras" pelas autoridades governamentais, que as utilizam mediante demandas judiciais e administrativas, formais e informais, para controlar conteúdo[20]; e d) promover a confiança e o sentimento de legitimidade dos usuários, governos e anunciantes na atividade da plataforma.

4.1 Transparência para aprimorar a moderação de conteúdo

É comum que obrigações de transparências sejam justificadas pela ideia de que a divulgação de mais informações pelas plataformas sobre seus procedimentos internos poderia aprimorar a atividade de moderação. Isso ocorreria, por exemplo, não só com a colaboração, a pressão e a opinião de *stakeholders*, pesquisadores, imprensa, organizações da sociedade civil, mas também com a regulação informada pelo maior conhecimento sobre o funcionamento da moderação.

A ideia de aprimoramento da moderação de conteúdo pressupõe um juízo sobre o que é uma *moderação de conteúdo positiva*. Neste estudo, considera-se que a qualidade da moderação de conteúdo não se mede apenas por um critério de eficiência na remoção de conteúdo potencialmente lesivo. Uma boa moderação não remove todo conteúdo lesivo o

20 Nas categorias a seguir, é possível identificar um paralelo entre elas e alguns dos tipos gerais de transparência identificados por Kosack e Fung (2014, p. 68-69). Eles concebem quatro categorias de transparência identificadas por seu alvo e pelos usuários da informação divulgada: a) quando o alvo é o governo e os usuários são os cidadãos na posição de participantes do debate público democrático, fala-se em "transparência para a liberdade de informação" – aqui há um paralelo com a fiscalização do uso de plataformas como *proxies* ("procuradoras"), que visa justamente permitir um debate público sobre os limites do poder governamental; b) quando o alvo é o governo, mas os usuários são cidadãos na posição de consumidores ou beneficiários de serviços públicos, configura-se a "transparência para a prestação de contas" – aqui não há paralelo, já que a moderação de conteúdo não trata de serviços públicos; c) quando os alvos são entidades privadas e os usuários são cidadãos na posição de participantes do debate público democrático, realiza-se a "transparência para o comportamento corporativo responsável" – em linha com a categoria de transparência para o aprimoramento da moderação; e d) quando os alvos são entidades privadas e os usuários são consumidores e beneficiários de serviços, caracteriza-se a "transparência regulatória" – categoria similar à transparência para a autoproteção dos usuários.

mais rápido possível, mas o faz porque pode adequadamente distinguir entre conteúdo lesivo e conteúdo permitido. Desse modo, protege o debate público enquanto preserva a liberdade de expressão de quem não está agindo em desacordo com o Direito e com as regras da plataforma. Isso significa buscar a redução da taxa não só de *falsos positivos* (casos em que um conteúdo permitido é detectado como problemático), mas também de *falsos negativos* (situações em que *não* se identifica nem se sanciona conteúdo problemático).

Esse aprimoramento dos sistemas de moderação de conteúdo depende muito da supervisão de terceiros, ao sinalizarem conteúdo problemático e ao sugerirem mudanças estruturais quando o sistema não cumpre as funções esperadas. Aqui existe um paralelismo claro entre essa função da transparência e o ciclo da ação: a transparência é o primeiro passo para as mudanças que resultem da adoção construtiva de feedbacks (Karanicolas, 2021, p. 61). A necessidade de feedbacks cresce quando as plataformas passam a tomar decisões cada vez mais difíceis e cujo impacto sobre o debate público é cada vez mais significativo. Quando, por exemplo, começaram a combater ativamente a desinformação, as plataformas assumiram um compromisso com a garantia de veracidade do conteúdo publicado por usuários – o que é desafiador, visto que a própria definição de *desinformação* é controversa. Inevitavelmente, isso levou ao controle de conteúdo de caráter político, o que envolve autoridades públicas (Karanicolas, 2021, p. 58). Nesses casos, apenas mediante a transparência relativa ao seu processo decisório, as plataformas podem saber se estão tomando as decisões corretas.

Existem pelo menos dois caminhos que, no caso das plataformas, transformam a informação em feedback, e este em mudanças concretas: a) por vias sociais e econômicas (usuários, organizações, anunciantes e a imprensa pressionam as plataformas para mudarem o comportamento); e b) por vias institucionais (o regulador utiliza os dados para informar novas obrigações legais que mudem o comportamento das plataformas).

No primeiro caso, espera-se que os usuários e anunciantes usem as informações divulgadas, normalmente alavancadas pela imprensa, ao tomarem decisões sobre que plataformas utilizarão ou sobre que plataformas terão como parceiras comerciais. Além disso, espera-se que a sociedade civil se organize para demandar mudanças quando há desacordo entre usuários e a atividade das plataformas. Exemplos reais indicam que dados sobre a moderação podem afetar o comportamento de usuários e anunciantes, e gerar movimentos que impulsionam a mudança. Em junho de 2020, diversas empresas parceiras do Facebook ameaçaram abandonar as ferramentas de publicidade da plataforma (que representam grande parte de seus lucros), devido à percepção de que a moderação estava ocorrendo de forma inadequada. O retorno à parceria dependia de uma postura mais rigorosa no combate ao discurso de ódio. Cruciais para o sucesso financeiro da plataforma, as marcas não desejavam ser associadas a esse tipo de conteúdo[21]. Com isso, promoveu-se um grande incentivo econômico para a mudança de comportamento da plataforma. Ao menos

21 Ver Hsu e Isaac (2021).

nesse caso, a ação dos anunciantes iniciou-se com um juízo negativo sobre a capacidade do Facebook de moderar discursos de ódio; essa percepção só podia estar fundamentada em informações (não necessariamente precisas) sobre processos internos da plataforma que alcançaram o público geral. Assim, é razoável esperar que intervenções de transparência que divulguem informações claras sobre a capacidade da moderação tenham efeitos semelhantes ao revelarem inadequações.

No segundo caso, o regulador pode obrigar as plataformas a divulgarem informações com a finalidade de se obterem subsídios para decisões que visem aprimorar a qualidade da moderação, pois a falta de informação sobre como ocorre a moderação torna mais difícil proceder a um debate público informado sobre como regular conteúdo na internet para proteger a liberdade de expressão e outros interesses (Suzor; West; Quodling; York, 2019, p. 1.527). Não é factível propor, por exemplo, normas para a composição das equipes de moderadores de conteúdo, se não se sabe como elas se compõem nem qual é o impacto dessa composição para a qualidade da moderação. Aliás, é provável que diversos defeitos nas práticas internas das plataformas, corrigíveis pela regulação, sejam imperceptíveis se não forem implantadas práticas de transparência.

Essas obrigações devem ser concebidas para preservar ao máximo a autonomia das empresas e a privacidade dos usuários, dado que quase todos os tipos de intervenção de transparência, com diferentes graus de sucesso persuasivo, se podem justificar pela finalidade de melhorar a moderação de conteúdo. Por conseguinte, a fim de controlar essas intervenções, o regulador deve considerar e explicar como os dados que ele pretende revelar influenciarão o comportamento dos usuários da informação e, baseado disso, divulgar a informação necessária para o público necessário. No caso de dados utilizados para informar nova regulação, por exemplo, pode ser mais adequado exigir a divulgação apenas para os órgãos reguladores, por meio de auditorias. Entretanto, quando se espera que o público em geral atue em resposta à informação divulgada, é necessário que a transparência seja mais abrangente.

O que divulgar?

Qualquer informação sobre os processos internos da moderação de conteúdo tem, em maior ou menor grau, o potencial de promover seu aprimoramento, já que sua divulgação abre portas à contribuição externa. Porém, como a transparência tem custos, importa que o regulador eleja tipos de informação com maior utilidade, visibilidade e capacidade de inferência. Destacam-se nesse quesito ao menos três categorias de informação: a) dados quantitativos discriminados sobre a atividade de moderação; b) dados sobre a composição e o treinamento dos times de moderadores humanos; e c) dados sobre o funcionamento e o desenvolvimento das tecnologias de moderação.

Quando divulgados, dados quantitativos sobre a atividade de moderação (como o número de publicações removidas e contas banidas, discriminadas pelo motivo da sanção) permitem

traçar um panorama da atividade das plataformas e de seus desafios[22]. Normalmente revelados em relatórios periódicos de transparência, os dados relativos à detecção de discursos de ódio ou de desinformação permitem – a quem monitora e regula a atividade das plataformas – identificar padrões ao longo do tempo tanto sobre a propagação desse conteúdo quanto sobre o comportamento reativo da plataforma (Zornetta, 2021, p. 8). Esses padrões denotam a fragilidade da moderação de conteúdo: por um lado, viabilizam a construção coletiva de conhecimento a respeito de soluções para essas dificuldades; por outro, regulam a pressão da imprensa, anunciantes e usuários sobre as plataformas para que incorporem tais soluções.

Dados sobre a composição e o treinamento das equipes de moderadores humanos (como critérios de treinamento, idiomas falados pelos moderadores e suas condições de trabalho) revelam fraquezas significativas na organização interna das plataformas. Para Suzor, West, Quodling e York (2019, p. 1.535), elas deveriam divulgar informações agregadas sobre seus quadros de moderadores, inclusive características demográficas e dados sobre os procedimentos de treinamento. Exemplo de como essas informações podem permitir a avaliação da moderação e a construção de feedbacks foi o relatório de transparência publicado pelo X (antigo Twitter) em 2023, como resposta às obrigações estabelecidas pelo *Digital Services Act*, da União Europeia[23]. O relatório revelou que havia 2.294 moderadores para publicações em inglês, mas para postagens em espanhol e em português eram apenas 20 e 41, respectivamente. Para idiomas como o croata e o holandês, havia apenas um moderador em todo o continente. Como consequência, a plataforma foi muito criticada pela falta de comprometimento com a moderação adequada de conteúdo em língua não inglesa.

Essa debilidade, que não foi apontada até surgir obrigação nesse sentido, pode ter efeitos catastróficos para a eficiência e a qualidade da tomada de decisão nesses cenários, pois as ferramentas de detecção automática de conteúdo são incapazes de oferecer soluções satisfatórias para todos os casos. Também pode ser relevante a informação sobre os agentes envolvidos na moderação de conteúdo (Zornetta, 2021, p. 43). Decisões podem ser tomadas por funcionários terceirizados, por empregados da empresa ou até por executivos (mais comum apenas quando os moderados ocupam cargos de grande repercussão[24]). O contato com esses dados permite avaliar como diferentes recursos têm sido destinados a diferentes casos e em diferentes países.

---

22 Exemplo de obrigação nesse sentido ocorre no art. 4º, § 1º, do PL nº 836/2022, proposto pelo deputado Eduardo Bolsonaro (PL-SP), no qual se sugere que os "relatórios [de transparência] devem conter: I – número total de medidas aplicadas a contas e conteúdos, conforme *caput*, adotadas em razão do cumprimento das regras próprias dos provedores e do cumprimento desta Lei, segmentadas por regra aplicada, por metodologia utilizada na detecção da irregularidade, e por tipo de medida adotada" (Brasil, 2022a).

23 Ver Labate (2023).

24 Foi o caso do banimento de Donald Trump pelo Twitter em janeiro de 2021. A decisão resultou de diversas reuniões entre os executivos do Twitter, que em geral não participavam do dia a dia da moderação de conteúdo na plataforma. Ver Dwoskin e Tiku (2021).

Por fim, dados sobre o funcionamento e o desenvolvimento das tecnologias de moderação (por exemplo, como funcionam os algoritmos utilizados para a detecção automática de conteúdo) são relevantes para haver opinião informada sobre seu desempenho. Porém, alguns desses dados são particularmente sensíveis sob o ponto de vista concorrencial, já que as plataformas desenvolvem essas tecnologias para se diferenciarem no mercado. Como apontam Gorwa, Binns e Katzenbach (2020), a funcionalidade desses sistemas e as bases de dados que os alimentam costumam ser intencionalmente escondidas ou apresentadas de forma vaga, e não é simples atingir um consenso sobre a legitimidade desse grau de sigilo[25].

4.2 Transparência para a autoproteção dos usuários

Em diversos casos, intervenções de transparência visam habilitar consumidores para protegerem seus interesses (Kosack; Fung, 2014, p. 68), pois, sem acesso a certos dados, consumidores não são capazes de exercerem seu poder de escolha de forma informada e de se protegerem dos potenciais riscos decorrentes dos serviços e produtos consumidos[26].

No caso da moderação de conteúdo, é comum a assimetria de informação entre o moderador (a plataforma) e o moderado (o usuário da plataforma). Quando o usuário é afetado por uma decisão de moderação, tal como a remoção de uma publicação ou o banimento de sua conta, as plataformas nem sempre fornecem razões suficientemente claras para essa decisão ou mecanismos para seu questionamento (Zornetta, 2021, p. 26). Às vezes, a plataforma limita-se a apontar a regra infringida, mas sem explicar como o conteúdo sancionado se enquadra nessa categoria. Nesses casos, o moderado não tem acesso a dados suficientes sobre seu caso para questionar de forma fundamentada a decisão do moderador, algo essencial para a posterior correção de erros[27].

A assimetria relacionada aos critérios de moderação pode ser considerada problemática antes mesmo de serem tomadas decisões concretas. A moderação encontra fundamento legal justamente no consentimento do usuário no controle de seu conteúdo, explicitado quando ele concorda com os termos e regulamentos das plataformas. Desse modo, poder-se-ia argumentar que a legalidade de sua atividade estaria prejudicada, se a plataforma não é clara sobre os critérios que usa para moderar conteúdo ou se na prática modera de forma diferente do que determinam seus regulamentos. Quando presente, a assimetria

25 Uma discussão mais aprofundada sobre esse tema não cabe neste artigo, mas deve ser fomentada para que esses conflitos sejam esclarecidos.

26 Retoma-se, aqui, o exemplo das informações divulgadas em rótulos de produtos alimentícios; apesar de pouco convidativas para os consumidores, são necessárias para viabilizar decisões de autoproteção.

27 Existem argumentos favoráveis ao sigilo na aplicação de sanções pelas plataformas, mesmo que em hipóteses bastante específicas. Um exemplo é a aplicação de mecanismos de *soft moderation*, em que publicações problemáticas não são removidas, mas seu alcance é reduzido (Zornetta, 2021, p. 26). Para Suzor, West, Quodling e York (2019, p. 1.531), plataformas podem reduzir o alcance de um *spammer* (aquele que utiliza ferramentas para disseminar conteúdo repetitivo e em quantidade enorme para atingir uma grande audiência), mas omitir essa informação de forma a evitar que ele saiba que foi detectado e crie outras contas.

de informação relativa aos fundamentos da moderação torna vulnerável o usuário, o que pode justificar intervenções de transparência.

O que divulgar?

Naturalmente, as intervenções com maior potencial de promover a proteção dos usuários contra decisões infundadas determinam a divulgação dos fundamentos da moderação de conteúdo. Esses dados existem tanto em abstrato (nos regulamentos das plataformas) quanto em concreto (razões para decisões específicas).

Os regulamentos gerais das plataformas precisam ser pormenorizadamente divulgados para que os usuários sejam capazes de consentir nas sanções aplicáveis por elas e adequar seu comportamento a tais regras. Hoje esses regulamentos são bastante minuciosos; mas, quando as redes sociais se tornaram populares, a moderação de conteúdo era conduzida de forma bastante opaca (Klonick, 2018, p. 1.630). Embora os termos de uso já existissem, as regras que guiavam a atividade de moderação não eram públicas ou eram pouco claras. Na prática, os moderadores seguiam protocolos internos, muitas vezes genéricos e permeáveis a valores subjetivos, que tornavam a moderação uma tarefa pouco compreensível ou previsível pelos usuários afetados (Klonick, 2018, p. 1.631).

Apesar de os regulamentos terem evoluído, ainda é comum a disparidade entre o que está disposto e os critérios efetivamente utilizados para moderar conteúdo. Vazamentos publicados por veículos de imprensa e pesquisas (Keller; Leerssen, 2019, p. 28; West, 2018) revelaram que, em algumas plataformas, os moderadores continuam a seguir parâmetros e políticas internas indisponíveis ao público e que contêm regras diferentes dos regulamentos acessíveis. É difícil avaliar quão distintos esses parâmetros são dos regulamentos públicos, ou se isso é uma prática generalizada entre as plataformas; mas essa disparidade é problemática e, por isso, pode motivar intervenções de transparência[28].

As razões que fundamentam decisões específicas são igualmente necessárias para o usuário proteger-se de moderação infundada. Quando o moderador é um revisor humano, é perfeitamente viável exigir que, com base nos regulamentos gerais, a plataforma ao menos informe ao moderado o porquê de seu conteúdo ter sido identificado como infrator – ou seja, como o conteúdo viola uma regra e que regra é essa[29]. As plataformas tendem a concordar em que o usuário deva ter acesso a dados e sistemas que lhe permitam questionar as

---

28 Há um exemplo de obrigação nesse sentido no art. 8º-B do PL nº 2.393/2021, proposto pela deputada Renata Abreu (Podemos-SP); nele se sugere que as redes sociais "com pelo menos um milhão de usuários registrados devem: [...] III – oferecer aos usuários regras claras sobre condutas que possam levar a exclusão de postagens do usuário, do perfil do usuário, ou a limitação de acesso a conteúdo postado pelo usuário" (Brasil, 2021).

29 No art. 21-B do PL nº 2.883/2020, de iniciativa do deputado Filipe Barros (PL-PR), propõe-se que na "hipótese de exclusão de conteúdo ou de conta ou perfil de usuário na aplicação, fica o provedor de conteúdo, sem prejuízo das demais disposições desta lei e do Código de Defesa do Consumidor, a declinar, em linguagem clara, de fácil entendimento e compreensão, os motivos que conduziram à exclusão, garantido ao usuário procedimento que garanta contraditório e ampla defesa, dentro da própria aplicação e por meios intuitivos e de fácil acesso e utilização" (Brasil, 2020b).

decisões das plataformas de acordo com sua própria interpretação de seus regulamentos. Por esse motivo, elas tornam disponíveis processos internos de recurso.

Porém, quando as decisões são tomadas por ferramentas de detecção automática de conteúdo (caso da esmagadora maioria das decisões de moderação a partir de 2020[30]), o acesso às razões da decisão é restrito, o que pode inviabilizar o cumprimento de certas intervenções de transparência. Em razão do modo como são elaboradas e funcionam as ferramentas mais sofisticadas de detecção automática de conteúdo[31], de um ponto de vista técnico é desafiador tornar inteligíveis para o usuário os fundamentos de suas decisões. Nesses casos, a definição do *mínimo* de informação[32] a ser divulgado deve resultar de um debate técnico e jurídico mais aprofundado.

4.3 Transparência para as autoridades públicas fiscalizarem o controle das plataformas

Intervenções de transparência direcionadas às plataformas também podem revelar informações relevantes sobre seu uso como "procuradoras" de autoridades públicas, pois partem da premissa de que estas se utilizam das ferramentas e técnicas de moderação de conteúdo das plataformas para fazerem valer seus interesses; e, com esse fim, utilizam-se de procedimentos formais (como ordens administrativas e judiciais) e de mecanismos informais (como a pressão política) para alinhar o comportamento das plataformas com seus objetivos de limitação da livre expressão.

À primeira vista, é notável que a Administração Pública seja capaz de garantir que a atividade de moderação de conteúdo seja compatível com o interesse público. Isso é especialmente verdade quando essa interferência tem fundamento legal e intenta assegurar que se cumpra a legislação nos espaços digitais de comunicação. É o caso, por exemplo, de decisões judiciais que exigem da plataforma a remoção de conteúdo ilegal de acordo com a legislação de determinado país.

Os problemas surgem quando essa interferência ocorre de forma opaca, desestruturada ou em desacordo com o ordenamento jurídico. Existem casos em que os regulamentos

---

30 De acordo com os relatórios de transparência do Youtube referentes ao período de janeiro a março de 2020, de um total de 6,1 milhões de vídeos removidos (por diversos motivos), 5,7 milhões foram identificados por ferramentas de detecção automática, sem interferência humana (Google, [2020]).

31 Em geral, essas ferramentas são construídas por meio do chamado *aprendizado de máquina*, técnica que permite desenvolver um algoritmo de tomada de decisão mediante o processamento de um grande banco de dados. No caso da moderação de conteúdo, ele "aprende a decidir" ao ler e identificar padrões no enorme conjunto de decisões humanas anteriores contidas no banco de dados. Ao entrar em contato com uma nova publicação, a ferramenta verifica se os padrões que encontrou no banco de dados estão presentes e rotula aquela publicação como infratora ou não. Para uma visão aprofundada e ao mesmo tempo didática sobre o funcionamento de algoritmos e do aprendizado de máquina, ver Desai e Kroll (2017).

32 Zornetta (2021, p. 313) sugere que as plataformas deveriam, no mínimo, revelar o número de falsos positivos e falsos negativos resultantes do uso de algoritmos, ou seja, o número de publicações legítimas sancionadas e o número de publicações problemáticas consideradas legítimas. Esse dado permite que atores externos avaliem não só se a taxa de erros é baixa o suficiente para justificar seu uso na moderação, mas também a evolução desse indicador ao longo do tempo, e demandem melhorias.

das plataformas são alterados e publicações são sancionadas como resultado de pressões informais dos governos, e não de procedimentos formais fiscalizáveis[33]. Isso acontece especialmente quando a legislação do país não prevê de forma clara a ilicitude de certo conteúdo que as autoridades públicas querem tornar inacessível. Esse fenômeno é mais comum em ordenamentos jurídicos como o dos EUA, em que a liberdade de expressão é tratada com alta prioridade[34]; mas ocorre também quando as autoridades buscam soluções para problemas cuja definição ou ilicitude ainda não é consenso, como a proliferação de conteúdo desinformativo.

Pode-se questionar que a transparência, nesse caso, deveria partir das autoridades que empregam ferramentas informais ou sem fundamento legal para influir na atividade das plataformas. De fato, o fundamento para esse tipo de intervenção está no dever de publicidade dos atos do Estado, e não nos efeitos da moderação de conteúdo sobre questões de interesse público. Contudo, é inegável que as plataformas se encontram em posição estratégica para organizar e revelar essas informações essenciais para o controle do Poder Público. Assim, existe também uma razão de conveniência por trás dessa finalidade da transparência.

O que divulgar?

Em geral, as plataformas têm divulgado voluntariamente informações sobre o número de demandas formais de remoção de conteúdo em seus relatórios periódicos de transparência, discriminadas por sua origem[35]. Entretanto, elas divergem quanto ao grau de detalhamento; e nelas prevalece a divulgação de demandas, como ordens judiciais.

Quanto à divulgação dessas informações, a legislação brasileira determina que a plataforma notifique os usuários quando um conteúdo de seus contatos é removido em decorrência de ordem judicial ou administrativa[36]. Porém, a maioria dos relatórios e essas formas de divulgação específica não capturam o uso mais problemático do Poder Público para influenciar o comportamento das plataformas: as demandas por modificações nos regulamentos das plataformas e pela remoção de conteúdo por mecanismos informais e opacos, como notificações extralegais e pressão política. Nesse cenário, quando as autoridades públicas exigem alterações nos regulamentos por vias inacessíveis aos usuários,

---

33 É o caso de quando a moderação ocorre por pressão de governos na forma de e-mails entre autoridades e executivos das empresas. Ver Liang (2023).

34 Citron (2018) atribuiu o nome *censorship creep* ao fenômeno em que as plataformas são utilizadas por governos para restringir discursos de uma forma que, de acordo com o ordenamento jurídico local, seria considerada censura se fosse realizada diretamente pelas autoridades públicas.

35 Ver, por exemplo, a aba *solicitações governamentais de remoção de conteúdo* do relatório periódico de transparência que o Google ([2024]) tornou disponível.

36 O art. 20, parágrafo único, da Lei nº 12.965/2014 (*Marco civil da internet*), por exemplo, determina que, "quando solicitado pelo usuário que disponibilizou o conteúdo tornado indisponível, o provedor de aplicações de internet que exerce essa atividade de forma organizada, profissionalmente e com fins econômicos substituirá o conteúdo tornado indisponível pela motivação ou pela ordem judicial que deu fundamento à indisponibilização" (Brasil, [2018]).

pode parecer que a remoção de certas publicações decorre unicamente de decisões da plataforma, quando, na verdade, elas resultam do exercício do poder político.

A respeito disso, Citron (2018, p. 1.058) defende que as plataformas deveriam divulgar relatórios pormenorizados sobre os esforços de autoridades públicas para restringir conteúdo também por meios informais. Segundo a autora, ao atenderem a esses requisitos, os relatórios de transparência satisfazem ao princípio da transparência ao viabilizarem um debate público sobre os efeitos da censura estatal e ao permitirem que cidadãos e organizações da sociedade civil proponham os limites da influência de autoridades públicas sobre o conteúdo publicado e moderado por plataformas digitais.

É difícil conceber como se daria na prática a exigência de divulgação proposta pela autora. Especialmente nos casos em que, por meio de pressões informais, solicitam a remoção de conteúdo ou a mudança dos regulamentos, as autoridades públicas estão optando por manter essas interlocuções longe do público, talvez por compreenderem que sua legitimidade é questionável. Como a falta de transparência dos agentes públicos é intencional, não parece razoável exigir que as plataformas revelem todos os contatos realizados por atores políticos e arquem com os custos decorrentes da falta de transparência dessas autoridades. É justificável, porém, exigir que as plataformas mantenham a divulgação das informações já compartilhadas em formato visível e de modo que viabilize inferências.

4.4 Transparência para promover o sentimento de confiança e da legitimidade

Além das três finalidades apresentadas, a divulgação de mais informações pelas plataformas tem uma finalidade mais abstrata; ainda assim, pode justificar algum nível de intervenção de ordem pública. A transparência, na forma das obrigações já comentadas, também promoveria a confiança dos usuários nas plataformas, bem como o sentimento de que é legítima a atividade de moderação de conteúdo.

À primeira vista, a confiança dos usuários, governos e investidores nas plataformas parece ser positiva exclusivamente para elas – e, desse modo, um interesse de ordem privada que não justificaria a mobilização de recursos públicos para sua promoção. Afinal, ao confiarem na plataforma, os usuários a utilizam mais, os investidores formam mais parcerias com as empresas e os governos evitam uma regulação que pode ser custosa.

Todavia, a desconfiança dos usuários na atividade de moderação das plataformas pode prejudicar sua efetividade, a qual é desejável pela capacidade de efetivar interesses públicos. Esse prejuízo surge sobretudo quando os usuários estão frustrados com a opacidade dos processos internos de moderação. Como já se mencionou, existem situações em que as plataformas não divulgam de forma clara as razões de suas decisões (ou os regulamentos que de fato põem em prática no dia a dia da moderação de conteúdo), nem se esforçam para explicar ao destinatário os motivos para sua sanção. Nesses casos, o usuário interpreta a ação da plataforma como ilegítima, o que afeta diretamente o potencial de mudança de seu comportamento.

Quando os usuários confiam na plataforma e consideram legítima sua atividade, a moderação de conteúdo de publicações problemáticas pode ter efeitos positivos no debate público, pois uma forma particularmente significativa de promover bons comportamentos é a formação de um sentimento de comunhão numa comunidade (Grimmelmann, 2015, p. 63). Quando as plataformas os articulam e os aplicam de forma adequada, os códigos de conduta são recebidos pelo usuário como indicadores de que o comportamento deve ser adaptado a fim de preservar a comunidade digital. Por consequência, o usuário que recebe a sanção da plataforma – mas acredita que ela é justa – tem uma chance muito menor de reincidir no comportamento problemático.

O contrário disso é bem mais provável quando os usuários não confiam na plataforma ou julgam ilegítima sua atividade. Pesquisa conduzida por West (2018) revelou que, por não compreenderem a razão de sua publicação ou conta ter sido sancionada, muitos usuários forjam narrativas e explicações que atribuem à plataforma motivações ocultas ou, muitas vezes, perversas. É o caso dos que acreditam serem alvos de perseguição, em virtude de sua posição ideológica tornada pública, ou de uma conspiração de que fazem parte as plataformas; e tais percepções podem ser validadas quando a sanção é recorrente (Suzor; West; Quodling; York, 2019, p. 1.533). Com esse raciocínio, tais usuários passam a tratar as plataformas como antagonistas, ameaças à sua liberdade, em vez de ajustarem seu comportamento para participar da comunidade de forma regrada, interpretando as regras da plataforma como forças positivas para a construção de um ambiente saudável. O potencial de mudança de comportamento da moderação de conteúdo – e, por consequência, seu potencial de promoção de interesses públicos – fica prejudicado.

Por isso, o objetivo de promover a confiança dos usuários na moderação de conteúdo pode justificar a exigência de que as plataformas divulguem informações, em especial sobre suas regras e sobre as razões das decisões tomadas. Quanto mais informação for revelada, os sancionados estarão mais bem informados sobre as normas que regem uma comunidade digital para tomarem decisões sobre seu comportamento. Por conseguinte, obrigar as plataformas a serem mais transparentes pode torná-las melhores "empreendedoras de normas" (Fagan, 2018, p. 394), capazes de propagar boas práticas de comunicação e debate.

5 Conclusão

Com fundamento em pesquisa exploratória da literatura e da prática legislativa brasileira, este artigo propôs quatro objetivos que podem dar sentido a intervenções de transparência na moderação de conteúdo e direcionar sua interpretação: a) aprimorar a atividade de moderação, tanto ao permitir a colaboração e a opinião de *stakeholders* (como pesquisadores e organizações da sociedade civil), quanto ao gerar fiscalização, incentivos e pressão para a mudança de comportamento das plataformas; b) permitir que usuários protejam e exerçam seus direitos de manifestação por compreenderem as regras da moderação e das

ferramentas de recurso e correção de erros; c) fiscalizar o emprego das plataformas como "procuradoras" por autoridades governamentais, que as utilizam mediante demandas judiciais e administrativas, formais e informais, para o controle de conteúdo; e d) promover a confiança e o sentimento de legitimidade dos usuários na atividade da plataforma.

Espera-se que a proposta sirva como ponto de partida para investigações mais profundas sobre os desafios inerentes à realização desses quatro objetivos. Este estudo apenas tangencialmente tratou dos dados que poderiam ser divulgados para efetivar esses objetivos, com o propósito de conferir maior concretude à teoria, ao mencionar brevemente certas dificuldades. Isso não significa que elas sejam simples, ou que seja simples a escolha das informações a serem divulgadas por intervenções. Na verdade, embora essa discussão seja facilitada se houver consenso sobre os objetivos da transparência, nela residem as principais controvérsias da temática.

Em particular, os autores acreditam que os conflitos concretos entre os objetivos da transparência e as hipóteses de sigilo legítimo merecem mais atenção da literatura. Se é verdade que a privacidade tem sido tema de aprofundamento significativo com o advento da proteção de dados pessoais como linha de pesquisa, também é verdade que há muito pouco escrito e previsto na legislação sobre quão ampla é no Direito brasileiro a proteção concedida pelos segredos comerciais. A identificação dos limites dessa categoria é tão conveniente quanto essencial para que se superem barreiras teóricas e jurisprudenciais para a concretização dos objetivos da transparência.

Também merecem atenção especial os desafios para que se efetive a transparência resultante da adoção de tecnologias de detecção automática de conteúdo pelas plataformas. Essas ferramentas podem tornar-se cada vez mais eficazes para identificar conteúdo problemático; porém, sem dedicação especial, dificilmente se tornarão melhores para explicar os fundamentos de suas decisões. Naturalmente, se o moderador não é sequer capaz de revelar suas razões de forma compreensível para humanos, existe aqui um empecilho considerável para se efetivarem alguns objetivos da transparência; ela demanda colaboração entre os pesquisadores da transparência e os pesquisadores que desenvolvem e implantam essas tecnologias.

Para se superarem tais desafios, é crucial atingir um grau mínimo de consenso sobre as finalidades que legitimam e dão sentido às intervenções de transparência. Os limites do segredo de negócios dependerão do quanto dificultam a efetivação de outros direitos; e o nível de detalhamento exigido das explicações da detecção automática dependerá do que se considera necessário para dar efetividade a esses objetivos. Em suma: um consenso sobre os objetivos da transparência pode organizar um debate produtivo, caso se coordenem os esforços de pesquisadores que investigam os mais diversos aspectos de sua efetivação.

Referências

ANANNY, Mike; CRAWFORD, Kate. Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, [*s. l.*], v. 20, n. 3, p. 973-989, Mar. 2018. DOI: https://doi.org/10.1177/1461444816676645.

BIRCHALL, Clare. Introduction to 'Secrecy and transparency': the politics of opacity and openness. *Theory, Culture & Society*, [*s. l.*], v. 28, n. 7-8, p. 7-25, Dec. 2011. DOI: https://doi.org/10.1177/0263276411427744.

BRANDEIS, Louis D. *Other people's money*: and how the bankers use it. New York: Frederick A. Stokes Co., 1914.

BRASIL. Câmara dos Deputados. *Parecer proferido em Plenário ao Projeto de Lei nº 2.630, de 2020, e apensados*. Institui a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet. Brasília, DF: Câmara dos Deputados, 2023. Disponível em: https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2265334. Acesso em: 2 dez. 2024.

_____. Câmara dos Deputados. *Projeto de Lei nº 283, de 2020*. Dispõe sobre o rito sumário para a retirada de conteúdos ilegais de redes sociais. Brasília, DF: Câmara dos Deputados, 2020a. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2237042. Acesso em: 2 dez. 2024.

_____. Câmara dos Deputados. *Projeto de Lei nº 836, de 2022*. Institui o marco legal dos provedores de mensageria e redes sociais e estabelece regras para educação midiática. Brasília, DF: Câmara dos Deputados, 2022a. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2319323. Acesso em: 2 dez. 2024.

_____. Câmara dos Deputados. *Projeto de Lei nº 2.393, de 2021*. Altera o Marco Civil da Internet – Lei nº 12.965, de 23 de abril de 2014 – para promover a liberdade de expressão nas redes sociais, e proibir a exclusão de perfis de usuários sem decisão judicial, e dá outras providências. Brasília, DF: Câmara dos Deputados, 2021. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2288858. Acesso em: 2 dez. 2024.

_____. Câmara dos Deputados. *Projeto de Lei nº 2.883, de 2020*. Altera o Marco Civil da Internet – Lei nº 12.965, de 23 de abril de 2014, e a Lei do Sistema Brasileiro de Defesa da Concorrência – Lei nº 12.529/2011. Brasília, DF: Câmara dos Deputados, 2020b. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2253673. Acesso em: 2 dez. 2024.

_____. *Lei nº 9.279, de 14 de maio de 1996*. Regula direitos e obrigações relativos à propriedade industrial. Brasília, DF: Presidência da República, [2024]. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9279.htm. Acesso em: 2 dez. 2024.

_____. *Lei nº 9.296, de 24 de julho de 1996*. Regulamenta o inciso XII, parte final, do art. 5º da Constituição Federal. Brasília, DF: Presidência da República, [2019]. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l9296.htm. Acesso em: 2 dez. 2024.

_____. *Lei nº 12.965, de 23 de abril de 2014*. Estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. Brasília, DF: Presidência da República, [2018]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 2 dez. 2024.

_____. *Lei nº 13.709, de 14 de agosto de 2018*. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2022b]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 2 dez. 2024.

CITRON, Danielle Keats. Extremist speech, compelled conformity, and censorship creep. *Notre Dame Law Review*, Notre Dame, IN, v. 93, n. 3, p. 1.035-1.071, 2018. Disponível em: https://scholarship.law.nd.edu/ndlr/vol93/iss3/3/. Acesso em: 2 dez. 2024.

COALIZÃO DIREITOS NA REDE. *Regulação de plataformas*. [*S. l.*]: Coalizão Direitos na Rede, [2024]. Disponível em: https://direitosnarede.org.br/campanha/pl2630/. Acesso em: 2 dez. 2024.

CRUZ, Francisco Brito (coord.); LANA, Alice de Perdigão; JOST, Iná. Iguais perante as plataformas?: equidade e transparência na moderação de conteúdo em plataformas digitais. *InternetLab*: diagnósticos e recomendações, São Paulo, n. 9, p. 1-23, jul. 2023. Disponível em: https://internetlab.org.br/wp-content/uploads/2023/08/relatorio_internetlab_crosscheck_PORTUGUES_ok2.pdf. Acesso em: 2 dez. 2024.

DESAI, Deven R.; KROLL, Joshua A. Trust but verify: a guide to algorithms and the law. *Harvard Journal of Law & Technology*, [*s. l.*], v. 31, n. 1, p. 1-64, 2017. Disponível em: https://jolt.law.harvard.edu/assets/articlePDFs/v31/31HarvJLTech1.pdf. Acesso em: 2 dez. 2024.

DWOSKIN, Elizabeth; TIKU, Nitasha. How Twitter, on the front lines of history, finally decided to ban Trump. *The Washington Post*, Washington, DC, Jan. 16, 2021. Disponível em: https://www.washingtonpost.com/technology/2021/01/16/how-twitter-banned-trump/. Acesso em: 2 dez. 2024.

FAGAN, Frank. Systemic social media regulation. *Duke Law & Technology Review*, Durham, NC, v. 16, n. 1, p. 393-439, 2018. Disponível em: https://scholarship.law.duke.edu/dltr/vol16/iss1/14/. Acesso em: 2 dez. 2024.

FUNG, Archon; GRAHAM, Mary; WEIL, David. *Full disclosure*: the perils and promise of transparency. New York: Cambridge University Press, 2007.

GILLESPIE, Tarleton. *Custodians of the internet*: platforms, content moderation, and the hidden decisions that shape social media. New Haven: Yale University Press, 2018.

GOOGLE. *Cumprimento das diretrizes da comunidade do YouTube*: relatório de transparência. [*S. l.*]: Gloogle, [2020]. Disponível em: https://transparencyreport.google.com/youtube-policy/removals?hl=pt_BR&total_channels_removed=period:2020Q1&lu=total_channels_removed. Acesso em: 2 dez. 2024.

_____. *Solicitações governamentais de remoção de conteúdo*: relatório de transparência. [*S. l.*]: Gloogle, [2024]. Disponível em: https://transparencyreport.google.com/government-removals/overview. Acesso em: 2 dez. 2024.

GORWA, Robert; BINNS, Reuben; KATZENBACH, Christian. Algorithmic content moderation: technical and political challenges in the automation of platform governance. *Big Data & Society*, [*s. l.*], v. 7, n. 1, Jan./June 2020. DOI: https://doi.org/10.1177/2053951719897945.

GRIMMELMANN, James. The virtues of moderation. *Yale Journal of Law & Technology*, New Haven, v. 17, n. 1, p. 42-109, 2015. Disponível em: https://openyls.law.yale.edu/handle/20.500.13051/7798. Acesso em: 2 dez. 2024.

HOVYADINOV, Sergei. Toward a more meaningful transparency: examining Twitter, Google, and Facebook's transparency reporting and removal practices in Russia. *SSRN*, [*s. l.*], Nov. 30, 2019. Disponível em: https://ssrn.com/abstract=3535671. Acesso em: 2 dez. 2024.

HSU, Tiffany; ISAAC, Mike. Advertiser exodus snowballs as Facebook struggles to ease concerns. *The New York Times*, New York, Oct. 5, 2021. Disponível em: https://www.nytimes.com/2020/06/30/technology/facebook-advertising-boycott.html. Acesso em: 2 dez. 2024.

KARANICOLAS, Michael. A FOIA for Facebook: meaningful transparency for online platforms. *Saint Louis University Law Journal*, Saint Louis, MO, v. 66, n. 1, p. 49-77, 2021. Disponível em: https://scholarship.law.slu.edu/lj/vol66/iss1/4/. Acesso em: 2 dez. 2024.

KELLER, Daphne; LEERSSEN, Paddy. Facts and where to find them: empirical research on internet platforms and content moderation. *SSRN*, [*s. l.*], Dec. 16, 2019. Disponível em: https://papers.ssrn.com/abstract=3504930. Acesso em: 2 dez. 2024.

KLONICK, Kate. The new governors: the people, rules, and processes governing online speech. *Harvard Law Review*, [*s. l.*], v. 131, n. 6, p. 1.598-1.670, Apr. 2018. Disponível em: https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/. Acesso em: 2 dez. 2024.

KOSACK, Stephen; FUNG, Archon. Does transparency improve governance? *Annual Review of Political Science*, [*s. l.*], v. 17, p. 65-87, May 2014. DOI: https://doi.org/10.1146/annurev-polisci-032210-144356.

KURTZ, Lahis Pasquali; CARMO, Paloma Rocillo Rolim do; VIEIRA, Victor Barbieri Rodrigues. *Transparência na moderação de conteúdo*: tendências regulatórias nacionais. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2021. Disponível em: https://irisbh.com.br/wp-content/uploads/2021/07/Transparencia-na-moderacao-de-conteudo-tendencias-regulatorias-nacionais-IRIS.pdf. Acesso em: 2 dez. 2024.

LABATE, Alice. Relatório de transparência do X revela números baixíssimos de moderadores não falantes de inglês. *Estadão*, São Paulo, 7 nov. 2023. Disponível em: https://www.estadao.com.br/link/empresas/relatorio-de-transparencia-do-x-revela-numeros-baixissimos-de-moderadores-nao-falantes-de-ingles/. Acesso em: 2 dez. 2024.

LEERSSEN, Paddy. The soap box as a black box: regulating transparency in social media recommender systems. *European Journal of Law and Technology*: EJLT, [*s. l.*], v. 11, n. 2, 2020. Disponível em: https://www.ejlt.org/index.php/ejlt/article/download/786/1012/3408. Acesso em: 2 dez. 2024.

LIANG, Annabelle. Biden officials must limit contact with social media firms. *BBC News*, [*s. l.*], 5 July 2023. Disponível em: https://www.bbc.com/news/technology-66106067. Acesso em: 2 dez. 2024.

MENDES, Gilmar Ferreira; FERNANDES, Victor Oliveira. Eficácia dos direitos fundamentais nas relações privadas da internet: o dilema da moderação de conteúdo em redes sociais na perspectiva comparada Brasil-Alemanha. *Revista de Direito Civil Contemporâneo*: RDCC, São Paulo, v. 31, n. 9, p. 33-68, abr./jun. 2022. Disponível em: https://ojs.direitocivilcontemporaneo.com/index.php/rdcc/article/view/1107. Acesso em: 2 dez. 2024.

NUNZIATO, Dawn Carla. From town square to Twittersphere: the public forum doctrine goes digital. *GWU*: legal studies research paper, [*s. l.*], n. 40, p. 1-75, 2018. DOI: https://dx.doi.org/10.2139/ssrn.3249489. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3249489. Acesso em: 2 dez. 2024.

OSORIO, Aline Rezende Peres; ALVIM, Frederico Franco; SIQUEIRA, Giselly; BARCELOS, Julia Rocha de; VARGAS, Marco Antonio Martin; RODRIGUES, Tainah Pereira; RONDON, Thiago. *Programa permanente de enfrentamento à desinformação no âmbito da Justiça Eleitoral*: plano estratégico eleições 2022. Brasília, DF: TSE, 2022. Disponível em: https://www.justicaeleitoral.jus.br/desinformacao/arquivos/programa-permanente-de-enfrentamento-a-desinformacao-novo.pdf. Acesso em: 2 dez. 2024.

RIBEIRO, Gustavo; D'AGOSTINI, Julia; SARMENTO, Paulo; RACHID, Raquel. *Report on algorithmic transparency and disinformation*: a multisectoral approach. Brasilia, DF: Laboratory of Public Policy and Internet, 2021. Disponível em: https://lapin.org.br/wp-content/uploads/2022/05/Transparencia-algoritmica_V6-final.pdf. Acesso em: 2 dez. 2024.

SALVADOR, João Pedro Favaretto; GALATI, Luiz Fernando; GUIMARÃES, Laíssa Maria. *Transparência na moderação de conteúdo*: uma bibliografia. São Paulo: FGV Direito SP, CEPI, 2023. Disponível em: http://bibliotecadigital.fgv.br:80/dspace/handle/10438/33701. Acesso em: 2 dez. 2024.

SALVADOR, João Pedro Favaretto; GUIMARÃES, Tatiane; GALATI, Luiz Fernando. Uma taxonomia da transparência na moderação de conteúdo. *In*: SEMINÁRIO GOVERNANÇA DAS REDES, 4., 2023, Belo Horizonte. *Anais* [...]. Belo Horizonte: Instituto de Referência em Internet e Sociedade, 2024. p. 40-58. Disponível em: https://irisbh.com.br/wp-content/uploads/2024/06/Anais-IVSGDR.pdf. Acesso em: 2 dez. 2024.

STATISTA. *Annual advertising revenue of Meta platforms worldwide from 2009 to 2023*. [*S. l.*]: Statista, Feb. 2024. Disponível em: https://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/. Acesso em: 2 dez. 2024.

SUZOR, Nicolas P. *Lawless*: the secret rules that govern our digital lives. Cambridge, UK: Cambridge University Press, 2019. Disponível em: https://osf.io/preprints/socarxiv/ack26. Acesso em: 2 dez. 2024.

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*: IJoC, [*s. l.*], v. 13, p. 1.526-1.543, 2019. Disponível em: https://ijoc.org/index.php/ijoc/article/view/9736. Acesso em: 2 dez. 2024.

THE DIGITAL Services Act package. *In*: SHAPING Europe's digital future. [*S. l.*]: European Commission, 4 Oct. 2024. Disponível em: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package. Acesso em: 2 dez. 2024.

UNITED KINGDOM. *Online Safety Act 2023*. [London]: legislation.gov.uk, 2023. Disponível em: https://www.legislation.gov.uk/ukpga/2023/50/enacted. Acesso em: 2 dez. 2024.

UNITED STATES. *The Freedom of Information Act, 5 U.S.C. § 552*. Washington, DC: U.S. Department of Justice, Office of Information Policy, Jan. 21, 2022. Disponível em: https://www.justice.gov/oip/freedom-information-act-5-usc-552. Acesso em: 2 dez. 2024.

WEST, Sarah Myers. Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media & Society*, [*s. l.*], v. 20, n. 11, p. 4.366-4.383, Nov. 2018. DOI: https://doi.org/10.1177/1461444818773059.

ZAPATERO GÓMEZ, Virgilio. *The art of legislating*. Translated by Jorge Yetano Roche. Cham: Springer, 2019. (Legisprudence library: studies on the theory and practice of legislation, v. 6).

ZORNETTA, Alessia. *Online misinformation*: improving transparency in content moderation practices of social media companies. 2021. Thesis (Master of Laws) – McGill University, Montreal, 2021. Disponível em: https://escholarship.mcgill.ca/concern/theses/rf55zd782. Acesso em: 2 dez. 2024.

ZUCCOLOTTO, Robson; TEIXEIRA, Marco Antonio Carvalho. *Transparência*: aspectos conceituais e avanços no contexto brasileiro. Brasília, DF: Enap, 2019. (Coleção governo e políticas públicas). Disponível em: https://repositorio.enap.gov.br/handle/1/4161. Acesso em: 2 dez. 2024.

Nota

Este artigo é fruto da pesquisa *Uma taxonomia da transparência na moderação de conteúdo*, conduzida pelo Centro de Ensino e Pesquisa em Inovação (Cepi) da FGV Direito SP entre o início de 2022 e o fim de 2023.

Responsabilidade e licenciamento

Acesse todas as edições da
Revista de Informação Legislativa

www.senado.leg.br/ril