UNIVERSIDADE FEDERAL DO RIO DE JANEIRO – UFRJ
ESCOLA DE COMUNICAÇÃO – ECO
INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA – IBICT
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO – PPGCI

BRUNO MAURICIO MATTOS MARTINS

MODERAÇÃO DE CONTEÚDO E TRANSPARÊNCIA: UMA COMPARAÇÃO ENTRE MERCADOS REGULADOS E NÃO REGULADOS

Dissertação de Mestrado
Agosto de 2025

BRUNO MAURICIO MATTOS MARTINS

MODERAÇÃO DE CONTEÚDO E TRANSPARÊNCIA: UMA COMPARAÇÃO ENTRE MERCADOS REGULADOS E NÃO REGULADOS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação do convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e a Escola de Comunicação da Universidade Federal do Rio de Janeiro (Eco-UFRJ), como requisito parcial à obtenção do título de mestre em Ciência da Informação.

Orientadora: Profª. Drª. R. Marie Santini.
Coorientadora: Profª. Drª. Débora Salles.

Rio de Janeiro

2025

CIP - Catalogação na Publicação

M444m

Mattos Martins, Bruno Mauricio
Moderação de conteúdo e transparência: uma comparação entre mercados regulados e não regulados / Bruno Mauricio Mattos Martins. -- Rio de Janeiro, 2025.
185 f.

Orientadora: R. Marie Santini.
Coorientadora: Débora Salles.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Escola da Comunicação, Instituto Brasileiro de Informação em Ciência e Tecnologia, Programa de Pós-Graduação em Ciência da Informação, 2025.

1. Digital Services Act. 2. Governança de plataformas. 3. Moderação de conteúdo. 4. Regulação de plataformas. 5. Relatórios de transparência. I. Santini, R. Marie, orient. II. Salles, Débora, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

BRUNO MAURICIO MATTOS MARTINS

MODERAÇÃO DE CONTEÚDO E TRANSPARÊNCIA: UMA COMPARAÇÃO ENTRE MERCADOS REGULADOS E NÃO REGULADOS

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Ciência da Informação do convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict) e a Escola de Comunicação da Universidade Federal do Rio de Janeiro (Eco-UFRJ), como requisito parcial à obtenção do título de mestre em Ciência da Informação.

Rio de Janeiro, 26 de agosto de 2025

gov.br Documento assinado digitalmente
ROSE MARIE SANTINI DE OLIVEIRA
Data: 18/11/2025 11:01:16-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. R. Marie Santini (Orientadora) PPGCI Ibict/Eco-UFRJ

gov.br Documento assinado digitalmente
DEBORA GOMES SALLES
Data: 17/11/2025 12:21:57-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Débora Salles (Coorientadora) NetLab UFRJ

gov.br Documento assinado digitalmente
GIUSEPPE MARIO COCCO
Data: 17/11/2025 14:42:00-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Giuseppe Mario Cocco (Membro interno) PPGCI Ibict/Eco-UFRJ

gov.br Documento assinado digitalmente
RAQUEL DA CUNHA RECUERO
Data: 18/11/2025 07:51:42-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Raquel da Cunha Recuero (Membro externo) PPGCOM UFRGS

À Helena, que tanto tem iluminado nossas vidas desde que chegou a este mundo.

AGRADECIMENTOS

A Ana e Sergio, meus pais, pelo amor incondicional, sem o qual eu nada seria.

À Larissa, por todo o amor e afeto nesses anos e por segurar minha mão quando parecia que as coisas não poderiam dar certo. A nós, todos os dias dessa vida.

A Breno e Bernardo, meus irmãos, pela amizade de berço.

À Rodmar, minha avó, por todo o carinho desde que eu me entendo por gente.

À Elizabeth, minha tia, pela inspiração para seguir nessa carreira tão desafiadora que é a vida acadêmica.

À Marie e à Débora, por tantas oportunidades e por toda a belíssima jornada que temos trilhado nos últimos anos, da qual este trabalho é apenas uma parte.

Aos professores Giuseppe e Raquel, por todas as contribuições, diretas e indiretas, a este trabalho e por terem aceitado avaliá-lo com tanta atenção e zelo.

Ao Renato, pela amizade extraordinária que cultivamos desde o início de nossa trajetória na Escola de Comunicação e por tantas trocas, na vida e na academia.

À Maria Francisca, pelas conversas brilhantes e reflexões compartilhadas, alternando leveza e dureza, e que contribuíram de forma decisiva para a construção desta pesquisa.

Ao Pedro, pela amizade de longa data e por ter acreditado em mim, mesmo nos momentos mais difíceis, assim como eu acredito em você.

Ao Daniel, pela constante parceria em tantos shows e pela cornetagem futebolística incessante, respiros valiosos nos meses dedicados a esta dissertação.

Aos colegas do NetLab UFRJ, em especial ao João Gabriel, à Daphne, ao Alékis, à Luciane, à Nicole, à Amanda, à Danielle, ao Erick, ao Felipe e ao Bernardo, que muito me ajudou com algumas das ilustrações deste trabalho, por toda a parceria e pelas trocas diárias. Produzir tanto ao lado de vocês me desacostumou ao ritmo solitário de uma dissertação.

Por fim, a todos que apoiam o trabalho desenvolvido pelo NetLab UFRJ e que nos inspiram a seguir em frente.

Welcome, my son, welcome to the machine
Where have you been?
It's alright, we know where you've been

RESUMO

MATTOS MARTINS, Bruno Mauricio. **Moderação de conteúdo e transparência**: uma comparação entre mercados regulados e não regulados. Orientadora: R. Marie Santini. Coorientadora: Débora Salles. 2025. 185 f. Dissertação (Mestrado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia/Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2025.

Controlada pelas próprias empresas, marcada pela opacidade e guiada por interesses comerciais, a moderação de conteúdo é uma das principais maneiras pelas quais as plataformas de redes sociais exercem sua governança, limitando o que seus usuários podem dizer ou fazer. Em um contexto global de crescentes pressões regulatórias, diferentes marcos normativos enfatizam a necessidade de reduzir essa opacidade, tornando obrigatória, entre outras medidas, a produção de relatórios de transparência, documentos que há anos são divulgados voluntariamente por grandes plataformas, sob críticas à superficialidade dos dados apresentados. A principal resposta regulatória no mundo a esse cenário é o *Digital Services Act* (DSA), aprovado na União Europeia em 2022, que estabelece um novo regime de diligência e transparência para as plataformas, inspirando iniciativas similares em outras regiões. Buscando compreender em que medida novos marcos normativos contribuem para aumentar a transparência da moderação de conteúdo, este trabalho analisa e compara os relatórios de transparência publicados voluntariamente em todo o mundo com aqueles exigidos pelo DSA na União Europeia por quatro plataformas: Facebook, Instagram, YouTube e X/Twitter. Com base em um quadro analítico original, constatamos que, apesar de avanços, o DSA ainda permite que as plataformas mantenham controle significativo tanto sobre o que decidem tornar visível quanto sobre a forma como as informações são apresentadas, revelando-se insuficiente para romper a opacidade processual da moderação de conteúdo, condição essencial para garantir o exercício de direitos fundamentais em ambientes online.

Palavras-chave: *Digital Services Act*; Governança de plataformas; Moderação de conteúdo; Regulação de plataformas; Relatórios de transparência

ABSTRACT

MATTOS MARTINS, Bruno Mauricio. **Moderação de conteúdo e transparência**: uma comparação entre mercados regulados e não regulados. Orientadora: R. Marie Santini. Coorientadora: Débora Salles. 2025. 185 f. Dissertação (Mestrado em Ciência da Informação) – Instituto Brasileiro de Informação em Ciência e Tecnologia/Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2025.

Controlled by the companies themselves, characterized by opacity and driven by commercial interests, content moderation is one of the main ways social media platforms exercise their governance, limiting what their users can say or do. In a global context of increasing pressure, various regulatory frameworks emphasize the need to reduce this opacity, making mandatory, among other measures, the production of transparency reports, documents that have been voluntarily published for years by major platforms, but criticized for the superficiality of the data presented. The main regulatory response to this scenario worldwide is the Digital Services Act (DSA), approved by the European Union in 2022, which establishes a new diligence and transparency regime for platforms, inspiring similar initiatives in other regions. Seeking to understand to what extent new regulatory frameworks contribute to increasing transparency in content moderation, this work analyzes and compares transparency reports voluntarily published worldwide with those required by the DSA in the European Union for four platforms: Facebook, Instagram, YouTube, and X/Twitter. Based on an original analytical framework, we find that despite notable advances, the DSA still allows platforms to maintain significant discretion both over what they decide to make visible and over how information is presented, revealing itself insufficient to break the procedural opacity of content moderation, a key condition for guaranteeing the exercise of fundamental rights in online environments.

Keywords: content moderation; Digital Services Act; platform governance; platform regulation; transparency reporting

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

Anvisa – Agência Nacional de Vigilância Sanitária

CDA – *Communications Decency Act*

CSV – *Comma-separated values*

DMA – *Digital Markets Act*

DSA – *Digital Services Act*

ECD – *e-Commerce Directive*

EUA – Estados Unidos da América

GARM – *Global Alliance for Responsible Media*

GDPR – *General Data Protection Regulation*

IA – Inteligência Artificial

LGPD – Lei Geral de Proteção de Dados

MCI – Marco Civil da Internet

NetzDG – *Netzwerkdurchsetzungsgesetz* (*Network Enforcement Act*)

OCDE – Organização para a Cooperação e Desenvolvimento Econômico

PDF – *Portable Document Format*

SCP – *Santa Clara Principles*

STF – Supremo Tribunal Federal

VLOPs – *Very Large Online Platforms*

VLOSEs – *Very Large Online Search Engines*

SUMÁRIO

1 INTRODUÇÃO

> [Nos últimos sete anos], tem havido um intenso debate sobre os perigos em potencial de conteúdos publicados online. Cada vez mais, governos e a mídia de legado têm atuado para nos censurar. Claro que há muita coisa ruim por aí [nas plataformas de redes sociais], [...] muitas coisas que nós levamos muito a sério. Nós chegamos a um ponto em que há muitos erros e muita censura. [...] Então, nós vamos voltar às nossas raízes e focar na redução destes erros, simplificando nossos termos de serviço e restaurando a liberdade de expressão em nossas plataformas.[1]

Em 7 de janeiro de 2025, Mark Zuckerberg, CEO e fundador da Meta, empresa que controla serviços como Facebook, Instagram, WhatsApp e Threads, anunciou uma nova visão para o funcionamento dessas plataformas: garantir que elas sejam espaços voltados para empoderar e fortalecer seus usuários, protegendo-os de "influências externas" que buscam "censurar" suas opiniões sobre tópicos sensíveis, como gênero e imigração, permitindo que possam se engajar livremente em "conteúdo cívico e político" (Hendrix, 2025). Além disso, Zuckerberg falava abertamente em colaborar com Donald Trump, que logo retornaria à presidência dos Estados Unidos da América (EUA) após quatro anos, em sua luta contra "pressões globais por censura" nas plataformas digitais, incluindo o "Estado chinês de vigilância" e o que ele chamou de "tribunais secretos na América Latina" (Hendrix, 2025). Com isso, as diretrizes das plataformas da Meta passaram a permitir conteúdos que atentam contra minorias de gênero, étnicas e religiosas, além de pessoas com deficiência, mesmo em países como o Brasil, onde tais condutas são consideradas criminosas (Aos Fatos, 2025).

A nova postura de Zuckerberg contrastava diametralmente com aquela que adotara apenas quatro anos antes. Após as invasões ao Capitólio em 6 de janeiro de 2021 por apoiadores de Trump, que haviam passado os meses anteriores difundindo internet afora teorias de que as eleições presidenciais de 2020 haviam sido fraudadas, grandes plataformas de redes sociais, incluindo as da Meta, se viram sob forte pressão pública por terem facilitado a erosão da integridade eleitoral nos EUA (Hern, 2020). Por conta própria, estas plataformas anunciaram o banimento dos perfis de Trump para reduzir o alcance de suas falas (Byers, 2021), tendo sido fortemente acusadas de censura por ele e por seus apoiadores (BBC, 2021).

Zuckerberg, porém, omite o fato nada trivial de que todas as grandes plataformas de redes sociais moderam o conteúdo publicado nelas e, de alguma forma, limitam o que seus usuários podem dizer ou fazer (Gillespie, 2018a). A moderação de conteúdo, na verdade, é o principal pilar de sua governança, estruturada com base em regras definidas e aplicadas unilateralmente pelas próprias plataformas, diante da escassez, ou mesmo ausência, de

[1] Aspas retiradas e traduzidas de publicação feita no perfil oficial de Mark Zuckerberg no Instagram em 7 jan. 2025. Disponível em: https://www.instagram.com/p/DEhf2uTJUs0/. Acesso em: 28 mar. 2025.

barreiras normativas e regulatórias em nível global (Klonick, 2017). Comumente, a moderação se traduz em ações como a remoção de publicações consideradas indesejadas ou problemáticas e o banimento dos usuários responsáveis por elas (Fitzgerald; Lokmanoglu, 2023; Klonick, 2017). Recentemente, há tentativas de compreender essas medidas como parte de um sistema mais amplo de distribuição e recomendação de conteúdo, que inclui práticas como a redução ou o aumento do alcance de certas falas em detrimento de outras (Alizadeh *et al.*, 2022; Gillespie, 2022; Santini; Salles; Mattos, 2023).

Não se deve presumir que as plataformas moderam conteúdo por um senso de responsabilidade social ou coletiva, nem apenas em resposta a pressões governamentais voltadas à institucionalização da censura, mas, primordialmente, por motivações comerciais (Bromell, 2022; Cobbe, 2021; Gillespie, 2018a; Roberts, 2016). O modelo de negócios destas plataformas, baseado na veiculação altamente personalizada e segmentada de publicidade, movimenta bilhões de dólares todos os anos, e moderar conteúdo é a maneira encontrada de elas provarem a seus anunciantes e parceiros comerciais que não necessariamente toleram tudo que é publicado nelas (Gillespie, 2018a; Roberts, 2016).

Essa governança, contudo, contrasta radicalmente com a imagem que sempre venderam ao público geral: de que seriam um "mercado de ideias", capaz de promover a inovação e a liberdade de expressão (Bromell, 2022; Cobbe, 2021; Napoli, 2019). Para evitar criar paradoxos na mente dos usuários, que precisavam acreditar que elas representavam arenas de livre expressão irrestrita, e ao mesmo tempo se proteger de pressões regulatórias e normativas, as plataformas de redes sociais desenvolveram uma lógica de funcionamento baseada na *opacidade*. Essa lógica impede a exposição das fragilidades, contradições e distorções sustentadas por sua autogovernança, permitindo que elas possam continuar tocando seus negócios longe das ameaças de responsabilização que o escrutínio público representaria (Common, 2020; Klonick, 2017; Suzor *et al.*, 2019).

No entanto, a autogovernança opaca das plataformas passou a ser cada vez mais contestada mundialmente, e a confiança e esperança depositadas nestas novas tecnologias foram se desfazendo progressivamente. Em seus primeiros anos, as plataformas de redes sociais foram celebradas por suas supostas características disruptivas, capazes de romper fronteiras tradicionais da comunicação e dar mais poder aos usuários (Tucker *et al.*, 2017). Agora, as plataformas são alvo de críticas por denúncias de uso indevido de dados pessoais, práticas econômicas desleais de viés monopolista e por contribuírem para a erosão democrática em diversos países afetados por operações de influência e manipulação de cidadãos em seus ambientes digitais (Flew, 2018; Gorwa, 2024; Guess; Lyons, 2020; Wardle;

Derakhshan, 2017). A opinião pública e autoridades em todo o mundo passaram a reconhecer que a autogovernança das plataformas de redes sociais permitira abusos que comprometem o bem-estar coletivo, indicando que havia chegado o momento de submetê-las a maior escrutínio e fiscalização (Flew, 2018; Schlesinger, 2020; Zittrain, 2019). Regular as plataformas não era mais uma questão de *se*, mas de *como* e *quando*.

Diversas propostas regulatórias em diferentes países passaram a se concentrar na moderação de conteúdo. Afinal, buscava-se entender por que tudo teria dado tão errado, mesmo diante das reiteradas afirmações das plataformas de que atuavam de boa-fé na construção de ambientes saudáveis de discussão, guiadas por diretrizes que diziam priorizar a proteção dos direitos humanos e o interesse público. O objetivo central era tornar a moderação mais transparente, compreensível e acessível ao público em geral, autoridades públicas, reguladores e pesquisadores, já que, até então, o entendimento sobre essas práticas era baseado em suposições difíceis de comprovar (Gorwa, 2024; Popiel, 2022; Roberts, 2018; Suzor *et al.*, 2019; Vergara; Jain; Mehta, 2024). De fato, muitos são os segredos que cercam a moderação de conteúdo. Moderadores humanos e sistemas algorítmicos automatizados seguem critérios e diretrizes que não são públicos, muitas vezes reforçando vieses que prejudicam desproporcionalmente alguns usuários. Esses usuários geralmente não compreendem os reais motivos pelos quais seus conteúdos são moderados, o que gera um sentimento de injustiça, especialmente quando publicações potencialmente mais problemáticas do que as suas permanecem no ar sem grandes questionamentos (Angwin; Grassegger, 2017; Myers West, 2018; Suzor *et al.*, 2019).

As plataformas buscaram implementar iniciativas voluntárias e autogovernadas para tornar a moderação de conteúdo mais transparente, principalmente por meio da publicação de relatórios de transparência. Esses documentos foram concebidos para responder às crescentes demandas por visibilidade, reunindo informações gerais e estatísticas sobre suas ações de moderação, funcionando como uma estratégia para apaziguar as exigências por maior responsabilização e evitar a adoção de regulações mais rigorosas e vinculantes (Dwivedi, 2022; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024; Suzor *et al.*, 2019; Urman; Makhortykh, 2023; Wagner *et al.*, 2020).

Não demorou muito, porém, para que especialistas e tomadores de decisão percebessem que estes relatórios apenas promovem uma visibilidade controlada das ações de moderação de conteúdo das plataformas de redes sociais. Críticos apontam que esses documentos, publicados sem qualquer padronização, costumam apresentar somente dados agregados e pouco detalhados, dificultando uma análise aprofundada de casos específicos e

das motivações por trás da moderação de conteúdo, e tendem a enfatizar decisões tomadas por obrigações governamentais, resultando na subnotificação intencional das ações de moderação realizadas por determinação própria das plataformas (Hovyadinov, 2019; Kosta; Brewczyńska, 2020; Urman; Makhortykh, 2023). Essas iniciativas, então, pouco contribuem para tornar a moderação de conteúdo mais transparente, pois as plataformas continuam tratando seus processos, protocolos e procedimentos centrais como "segredos industriais".

De certa forma, as iniciativas limitadas de transparência de moderação de conteúdo acabaram se voltando contra as plataformas. Projetos regulatórios passaram a usar essas ações como base para, por exemplo, criar obrigações vinculantes que estabelecem critérios mínimos para a divulgação de relatórios de transparência. A intenção é que os documentos tenham valor real e possibilitem a responsabilização das plataformas por falhas, vieses e riscos potencializados por sua atuação (Heldt, 2019; Suzor *et al.*, 2019). O caso mais proeminente disso é o *Digital Services Act* (DSA), aprovado na União Europeia em 2022 e em plena vigência desde o início de 2024. O DSA busca consolidar um novo regime de diligência e transparência para as plataformas de redes sociais, fazendo com que demonstrem atuar em prol do interesse público e da proteção dos direitos fundamentais de seus usuários, responsabilizando-as quando falharem nesse dever e servindo, em última instância, como referência para outras propostas regulatórias em escala global (Dwivedi, 2022; Helberger; Samuelson, 2024; Leerssen, 2024).

Discutir a transparência da moderação de conteúdo pela via regulatória é necessário, pois as plataformas não têm incentivos para aprimorar voluntariamente suas políticas de transparência, que raramente permitem a produção de conhecimento relevante por pesquisadores e pelo poder público (Bossetta, 2020; Rieder; Hofmann, 2020; Vergara; Jain; Mehta, 2024; Zalnieriute, 2021). Em diversos setores econômicos, a transparência é tida como a linha auxiliar para a dita "boa regulação", fortalecendo a legitimidade de deliberações normativas (Vergara; Jain; Mehta, 2024; Wagner *et al.*, 2020). No caso da moderação de conteúdo, uma transparência verdadeira e robusta permite identificar possíveis vieses e garante a responsabilização adequada das plataformas pela exposição dos usuários a riscos. Além disso, é condição indispensável para identificar violações de direitos como a liberdade de expressão e a privacidade cometidas por essas plataformas (Common, 2020; Gorwa; Garton Ash, 2020).

Nesse contexto, as declarações de Zuckerberg apresentadas no início deste capítulo representam uma reação clara à onda regulatória global que vem se impondo sobre as plataformas de redes sociais. Em especial, refletem a resistência a esforços que buscam

subordinar as práticas de moderação de conteúdo ao interesse público, muitas vezes definidos pelas próprias plataformas de forma estratégica e falaciosa como tentativas de censura estatal, em vez de guiadas por interesses puramente comerciais. O DSA tornou-se o principal alvo das críticas de Zuckerberg e de representantes de outras grandes plataformas, que frequentemente se articulam para barrar iniciativas regulatórias, sobretudo nos EUA, onde a maioria dessas empresas tem origem (Lu, 2025; Popiel, 2018).

No entanto, no que se refere à transparência da moderação de conteúdo, há incertezas sobre o quanto o DSA, apesar de suas ambições e das expectativas que o cercam, realmente representa um avanço em relação às limitações das medidas voluntárias adotadas pelas plataformas, que permanecem a regra na maior parte do mundo, onde ainda atuam de forma autogovernada. Diante disso, este trabalho tem como **objetivo** analisar e comparar os relatórios de transparência de moderação de conteúdo publicados voluntariamente e por exigência do DSA por quatro grandes plataformas de redes sociais: Facebook, Instagram, YouTube e X/Twitter. Com isso, buscamos entender de que maneira a transparência regulada e mandatória imposta às plataformas contribui, de fato, para a redução da opacidade da moderação de conteúdo.

Essas plataformas foram selecionadas por diversos motivos: todas possuem uma ampla base global de usuários; estão sujeitas às maiores exigências de transparência previstas no DSA; compartilham arquiteturas e funcionalidades comuns, o que permite uma comparação mais equilibrada entre elas; apresentam grande relevância acadêmica; são amplamente utilizadas como fontes de informação por seus próprios usuários; e exercem forte lobby político contrário à regulação (Leone de Castris, 2024; Popiel, 2018; Warnke; Maier; Gilbert, 2024; Zuckerman, 2021).

Para a condução deste estudo, desenvolvemos um quadro analítico original, composto por 60 critérios de avaliação, destinado à comparação crítica e sistemática dos relatórios de transparência de moderação de conteúdo publicados pelas plataformas selecionadas entre o final de 2024 e o início de 2025. O quadro não se propõe a avaliar as informações divulgadas nos relatórios em si – afinal, não há meios possíveis para auditá-las com precisão –, mas sim a identificar quais dados são efetivamente publicados por cada plataforma, considerando dois contextos distintos: um quase-global, no qual elas têm ampla liberdade decisional para conduzir suas ações de moderação como acharem melhor, e outro mais restrito, marcado por um novo marco normativo que, teoricamente, limita sua capacidade de autogovernança. Para tanto, consideramos dois aspectos principais: a granularidade das informações, que se refere ao grau de detalhamento dos dados, sendo que uma granularidade baixa indica informações

mais gerais e agregadas; e a qualidade das informações, ou seja, sua clareza e utilidade para garantir uma transparência sólida (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024).

A partir disso, pretendemos responder às seguintes questões de pesquisa:

QP1) Qual é o nível de transparência voluntária da moderação de conteúdo das plataformas selecionadas em escala global?

QP2) Quais práticas de moderação de conteúdo são mais opacas nos relatórios de transparência voluntários publicados globalmente pelas plataformas selecionadas?

QP3) Em que medida o *Digital Services Act* contribui para o aprimoramento da transparência da moderação de conteúdo das plataformas selecionadas na União Europeia?

QP4) Quais práticas de moderação de conteúdo permanecem mais opacas nos relatórios de transparência publicados pelas plataformas em cumprimento ao *Digital Services Act* na União Europeia?

QP5) Quais lacunas são identificadas nas exigências de transparência de moderação de conteúdo estabelecidas pelo *Digital Services Act*?

A pesquisa parte do reconhecimento de que o debate regulatório tem ganhado destaque na literatura acadêmica sobre plataformas digitais, especialmente redes sociais, mas ainda há lacunas em estudos empíricos que permitam mensurar concretamente a aplicação de marcos normativos, além de sua elaboração. A União Europeia não esconde a ambição de transformar o DSA em referência regulatória global para plataformas e serviços digitais (European Commission, 2024), enquanto a América Latina apresenta uma tendência histórica de se inspirar em modelos europeus na regulação dos setores de tecnologia e telecomunicações (Bizberge; Mastrini; Gómez, 2023). Há, no entanto, uma necessidade urgente de qualificar o debate regulatório na região, onde as tentativas de avanço frequentemente esbarram no forte lobby das plataformas e de outras empresas de tecnologia, que buscam simplificar, deslegitimar e interromper a discussão para preservar o *status quo* de sua governança (Salles *et al.*, 2024).

Diante desse cenário, buscamos extrair os aspectos mais relevantes e identificar deficiências regulatórias e de transparência que ainda carecem de soluções específicas em outras regiões, oferecendo evidências que contribuam para o avanço desse debate. Tratar da moderação de conteúdo não é tarefa simples: dificilmente haverá consenso entre plataformas, Estados e usuários sobre seus termos, então os critérios adotados e o grau de transparência continuarão a ser questionados e precisarão ser constantemente adaptados (Douek, 2021). Ainda assim, a transparência é essencial para garantir os direitos fundamentais dos usuários

em ambientes digitais, sendo, portanto, imprescindível criar condições que priorizem o interesse público como base para uma nova governança digital centrada na redução da opacidade das plataformas (Gorwa; Garton Ash, 2020; Suzor *et al.*, 2019).

Este trabalho está dividido em três capítulos, sem contar a introdução e as considerações finais. No **Capítulo 2**, realizamos uma revisão sobre as plataformas de redes sociais, situando-as como um subconjunto das plataformas digitais, resultado do processo de plataformização da internet. Em seguida, examinamos a estrutura de seu modelo de negócios e a lógica de suas operações comerciais, para então discutir a governança de conteúdo adotada por essas plataformas, marcada majoritariamente pela autogovernança. Encerramos o capítulo com uma análise das crises de opinião pública e dos impasses regulatórios que emergiram a partir da chamada desordem informacional, colocando as plataformas de redes sociais no centro de pressões por reformas em seus regimes de governança de conteúdo em diferentes partes do mundo. Assim, nosso referencial teórico se ancora nos Estudos de Plataforma (Gillespie, 2010; Helmond, 2015; Poell; Nieborg; van Dijck, 2019; van Dijck, 2020), na Economia Política da Informação e da Comunicação (Bromell, 2022; Dantas, 2014; Dantas; Raulino, 2020; Zuboff, 2019, 2020), e na literatura acadêmica sobre governança de plataformas (Gorwa, 2019a, b; Warnke; Maier; Gilbert, 2024).

O **Capítulo 3** é dedicado à moderação de conteúdo, um dos principais pilares da autogovernança das plataformas de redes sociais (Gillespie, 2018a; Grimmelmann, 2015). Trata-se de um conjunto de práticas de filtragem, organização e curadoria das publicações que circulam nesses ambientes, profundamente entrelaçadas com suas estratégias comerciais e com impacto direto sobre o que é visível – ou invisível – para os usuários no cotidiano das plataformas (Roberts, 2016, 2018). Após detalharmos os processos operacionais da moderação de conteúdo, voltamo-nos aos problemas de transparência que a envolvem e a mantêm como um verdadeiro segredo industrial das grandes plataformas de redes sociais. Essa opacidade impede que as práticas de moderação sejam devidamente escrutinadas e que os riscos a elas associados sejam enfrentados de forma adequada (Gorwa; Garton Ash, 2020; Suzor *et al.*, 2019; Zalnieriute, 2021). Entre as medidas de transparência adotadas voluntariamente pelas plataformas para conferir, em teoria, maior visibilidade às suas práticas de moderação de conteúdo, este estudo foca nos relatórios de transparência (Urman; Makhortykh, 2023). O capítulo se encerra com a apresentação do *Digital Services Act*, considerado o principal marco regulatório para enfrentar os problemas de transparência na moderação de conteúdo das grandes plataformas de redes sociais (Dwivedi, 2022; Eifert *et al.*, 2021; Leerssen, 2024).

Por fim, nossa análise é apresentada no **Capítulo 4**. O capítulo tem início com a seleção e apresentação dos relatórios de transparência de moderação de conteúdo escolhidos para o estudo, abrangendo tanto publicações voluntárias das plataformas selecionadas em âmbito global quanto materiais divulgados especificamente ao público e às autoridades da União Europeia em cumprimento às exigências do DSA. Na sequência, detalhamos a construção de um quadro analítico para a comparação crítica dos relatórios. O quadro, bem como seu embasamento teórico, pode ser conferido na íntegra no **Apêndice**. Na segunda parte do capítulo, apresentamos os resultados de nossa análise, a fim de responder às questões de pesquisa aqui propostas.

2 PLATAFORMAS DE REDES SOCIAIS: CONTROVÉRSIAS E DESAFIOS REGULATÓRIOS GLOBAIS

Neste capítulo, abordamos diferentes discussões em torno das plataformas de redes sociais, desde a definição do termo até as pressões crescentes por uma regulação mais efetiva dessas tecnologias. Na **seção 2.1**, nos dedicamos à conceituação de *plataformas digitais* e de *plataformas de redes sociais*, compreendidas aqui como uma categoria específica das primeiras. Isso nos permitiu explorar debates sobre o uso do termo "plataforma" para designar essas tecnologias e sobre o processo de *plataformização*, que alterou de forma definitiva os padrões de uso da internet e foi decisivo para que essas empresas acumulassem crescente poder sobre as comunicações e interações sociais. Na **seção 2.2**, guiados pela perspectiva da Economia Política da Informação e da Comunicação, examinamos o modelo de negócios dessas plataformas, com destaque para a noção de Capitalismo de Vigilância, entendida como uma lógica de acumulação baseada na extração e modelagem massiva e sistemática dos dados de usuários, visando à geração de lucros por parte dessas empresas.

Na **seção 2.3**, discutimos a governança multissetorial de plataformas digitais e de redes sociais, orientando nossa leitura pela oposição entre a "governança *das* plataformas" e a "governança *pelas* plataformas". A primeira diz respeito à forma como essas empresas são – ou deixam de ser – reguladas por marcos normativos externos; a segunda, à maneira como elas próprias estruturam sua atuação e se autogovernam, frequentemente em contextos de fragilidade ou ausência regulatória. Por fim, na **seção 2.4**, argumentamos que essa forma de autogovernança, caracterizada pela ausência de mecanismos efetivos de responsabilização e sustentada por um modelo de negócios voltado à maximização do engajamento e dos lucros em detrimento do bem-estar coletivo, resultou em um cenário de *desordem informacional*, levando à contestação do status normativo das plataformas e ao fortalecimento das demandas por uma regulação externa mais rigorosa.

2.1 Entendendo as plataformas de redes sociais

No ecossistema informacional contemporâneo, é fácil se perder diante da profusão de termos que, embora muitas vezes usados para expressar ideias semelhantes, carregam conotações semânticas, políticas e normativas bastante distintas. De maneira coloquial, é comum que expressões como "sites", "aplicativos" e "plataformas" sejam usadas como

sinônimos, o que pode levar a equívocos sobre as especificidades de cada tecnologia, subestimando ou superestimando aspectos fundamentais de seu funcionamento. Por essa razão, julgamos necessário iniciar este trabalho delimitando com clareza o que entendemos por *plataformas digitais* e, mais precisamente, por *plataformas de redes sociais*, antes de chegarmos às recentes discussões em torno de sua regulação e governança.

As plataformas digitais são definidas pelas oportunidades em escala inédita de comunicação, interação, vendas e trocas, materiais ou simbólicas, que oferecem (Gillespie, 2010). Para Eifert *et al*. (2021, p. 987–988, tradução do autor), elas seriam infraestruturas tecnológicas que conectam pessoas físicas e/ou jurídicas entre si para que elas engajem em "interações que visam à criação de valor", como a troca de bens, serviços ou informações. Sintetizando essas ideias, adotamos o entendimento de Gorwa (2024, p. 16, tradução do autor), para quem as plataformas são "produtos viabilizados digitalmente que medeiam relações entre duas ou mais partes, geralmente com elementos técnicos que permitem a terceiros interagir com eles ou desenvolver funcionalidades a partir deles". Para o autor, essa definição é relevante por três razões centrais, que serão exploradas no decorrer deste trabalho e atravessarão as discussões aqui desenvolvidas: (i) leva em conta os aspectos técnicos do funcionamento das plataformas, compreendendo-as, ainda assim, como *produtos* desenvolvidos para gerar lucro às suas empresas-mãe; (ii) parte do entendimento de que essas plataformas não atuam como intermediárias neutras nas comunicações e trocas realizadas por seus usuários; e (iii) concebe as plataformas como *mercados multilaterais*, que estruturam relações entre atores diversos (Gorwa, 2024).

À luz de Flew, Martin e Suzor (2019), entendemos que há plataformas digitais de variados tipos, como lojas de aplicativos (iTunes, Google Play), aplicativos de encontros (Tinder, Grindr), ferramentas de hospedagem (Airbnb), buscadores (Google, Bing), ferramentas de transporte (Uber, 99), *marketplaces* online (Mercado Livre, eBay) e serviços de *streaming* (Netflix, Spotify), para citar alguns exemplos. Este trabalho, por sua vez, foca na categoria das plataformas *de redes sociais*[2], que buscamos delinear com precisão. Para isso, recorremos à definição formulada por Gillespie (2018a), que destaca quatro fatores principais que caracterizam as plataformas de redes sociais: (i) elas hospedam, organizam e fazem circular as publicações e as interações de seus usuários; (ii) sem que elas tenham produzido ou

[2] Utilizamos a expressão "plataformas de redes sociais" para nos referirmos ao que a literatura acadêmica anglófona denomina "*social media platforms*" ou, simplesmente, "*social media*". Essa formulação foi uma escolha estilística, aproximando o termo do uso coloquial e amplamente difundido de "redes sociais". Quando nos referirmos a "plataformas digitais" ao longo deste trabalho, estaremos tratando de situações que também se aplicam às plataformas de redes sociais, ainda que não limitadas a elas.

encomendado a maior parte desse conteúdo; (iii) a partir de uma infraestrutura computacional de extração e processamento de dados, voltada à veiculação de publicidade como meio de geração de lucro; (iv) moderando – e devendo moderar – as publicações de seus usuários, segundo logísticas de detecção, revisão e aplicação de regras, ponto crucial a este trabalho e explorado em maior profundidade no Capítulo 3. Mesmo entre as plataformas de redes sociais, é possível identificar diferentes categorias. O X/Twitter, por exemplo, oferece serviços de *microblogging*; YouTube, TikTok e Instagram são voltados ao compartilhamento de conteúdo audiovisual; o Reddit se dedica à construção de fóruns de discussão; e o Facebook ocupa uma posição quase intermediária, reunindo diversas dessas funções (Flew; Martin; Suzor, 2019; Gorwa, 2024).

Mesmo que o conteúdo que circula nas plataformas de redes sociais, em sua maioria, não seja produzido ou encomendado por elas, elas tomam decisões cruciais sobre essa circulação, especialmente no que diz respeito ao que distribuir, para quem, como estruturar as conexões e interações entre usuários, e o que deve ou não ser permitido (Gillespie, 2018b). Além disso, as interações e conexões que ajudam a estabelecer não ocorrem em um vácuo ou ao acaso, mas são, em grande medida, viabilizadas pelo processamento de dados gerados pelos próprios usuários, fazendo com que suas experiências nelas reflitam suas preferências individuais (Eifert *et al.*, 2021; Poell; Nieborg; van Dijck, 2019).

Toda ação nas plataformas de redes sociais é mediada por sua arquitetura digital: os protocolos técnicos que possibilitam, limitam e moldam o comportamento dos usuários em um espaço online (Bossetta, 2018). Para Bossetta (2018), há quatro aspectos definidores da arquitetura de plataformas de redes sociais: (i) estrutura de rede, que diz respeito aos critérios que determinam a formação de conexões entre diferentes usuários; (ii) funcionalidade, que abrange os elementos técnicos responsáveis por como as publicações são feitas, distribuídas e acessadas nas plataformas, como o *feed* de atividades do usuário, que agrega, ranqueia e exibe conteúdos de forma simplificada; (iii) filtragem algorítmica, que compreende as regras de seleção, ranqueamento e visibilidade das publicações que chegam a determinado usuário em seu *feed* personalizado por meio de sistemas de recomendação de conteúdo, ajudando a moldar suas percepções sobre determinados assuntos; e (iv) dataficação (*datafication*), que se refere à quantificação das atividades dos usuários e dos "rastros digitais" que eles deixam nas plataformas, os quais são processados algoritmicamente para uma série de finalidades, sobretudo comerciais, como será aprofundado ainda neste capítulo.

A atividade nas plataformas de redes sociais se concentra, basicamente, em dois eixos: o conteúdo gerado pelos próprios usuários, sendo disseminado para outros usuários por meio

do ranqueamento algorítmico e das conexões estabelecidas entre eles, contribuindo para a democratização do acesso à informação e às ferramentas de comunicação (Tucker *et al.*, 2017); e conteúdo publicitário, distribuído algoritmicamente conforme os perfis comportamentais de cada usuário, como será discutido na segunda seção deste capítulo, e que ocupa parcelas cada vez maiores de seus *feeds* de atividade (Herrman, 2023; Knoll; Proksch, 2015; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Grael; *et al.*, 2024). Os usuários também desempenham um papel ativo na seleção, curadoria e distribuição de conteúdo por meio de ações como curtir, denunciar, avaliar e compartilhar, influenciando diretamente o conteúdo ao qual outros usuários são expostos (Helberger; Pierson; Poell, 2017), ainda que a real importância atribuída a essas interações pelas plataformas permaneça incerta. Para Recuero (2019), isso permite compreender as plataformas de redes sociais como estruturas que emergem das ações individuais e coletivas de seus usuários, os quais influenciam a visibilidade de determinados temas, o silenciamento de outros e decidem o que deve circular em quais espaços de discussão. Todo esse arsenal ferramental sociotécnico leva Gillespie (2018a, p. 23, tradução e grifo do autor) a afirmar que "as plataformas talvez não definam sozinhas o conteúdo do discurso público, mas certamente *moldam a forma que ele assume*".

Contudo, a ideia de "plataforma" não é nem estável nem definitiva. Trata-se, antes, de um esforço discursivo de provedores de serviços digitais e empresas de tecnologia e telecomunicações que passaram a ser compreendidos dessa forma e que, na prática, exercem um papel de "curadoria da opinião pública" (Gillespie, 2010), como exploraremos ao longo deste trabalho. Para Gillespie (2010), responsável por essa linha de raciocínio, referir-se a um serviço digital como "plataforma" não é um ato simples ou vazio, mas uma escolha de grande densidade semântica. Na perspectiva do autor, o conceito de "plataforma" aplicado às plataformas digitais e de redes sociais emerge da convergência de quatro sentidos historicamente atribuídos ao termo: (i) computacional, referindo-se a uma infraestrutura que dá suporte ao desenvolvimento e uso de aplicações específicas, como *hardware*, sistemas operacionais e acessórios; (ii) arquitetural, relacionada a uma superfície elevada sobre a qual pessoas se apoiam; (iii) figurado, enquanto base conceitual a uma determinada ação ou evento; e (iv) político, como o palco a partir do qual um candidato se dirige ao público e articula sua agenda (Gillespie, 2010).

Inspirados pela diversidade semântica do termo, na segunda metade dos anos 2000 serviços como YouTube e Facebook começaram a se autodenominar "plataformas", substituindo gradualmente expressões como "sites", "empresas", "serviços" e "comunidades" em suas comunicações institucionais. Essa mudança funcionou como uma estratégia que

sugere um arranjo progressista e igualitário, prometendo impulsionar todos aqueles que voluntariamente decidissem utilizá-las (Gillespie, 2010). Essa abordagem foi eficaz para atrair novos públicos encantados com suas promessas e seu caráter disruptivo, especialmente em contraste com a mídia de massa tradicional, já que os usuários passariam a ter o direito supostamente irrestrito de expressar-se e de se conectar com vozes de seu interesse (Gillespie, 2010). Hoje, a noção de "plataforma" transcende uma simples classificação semântica e é reivindicada por toda a indústria digital-informacional. Muitos dos serviços que conhecemos como plataformas são o resultado de transformações estruturais e arquitetônicas que afetaram toda a internet como parte do processo conhecido como *plataformização* (*platformisation*) (Gorwa, 2019b; Helmond, 2015; Poell; Nieborg; van Dijck, 2019).

Para Poell, Nieborg e van Dijck (2019), a plataformização – um processo liderado por empresas dos Estados Unidos da América (EUA), mas com variações e exceções regionais e nacionais – seria a penetração das infraestruturas, dos processos econômicos e dos modos de governança de plataformas digitais em diferentes setores da economia e esferas da vida coletiva, assim como a reorganização das práticas culturais em torno dessas plataformas. Helmond (2015) argumenta que esse processo ocorreu, em grande parte, graças à programabilidade das plataformas digitais, que, no auge de sua expansão, disponibilizaram seus dados para exploração por pesquisadores, engenheiros e cientistas, contribuindo para a construção de uma internet essencialmente social e interconectada por aplicações baseadas nesses dados. Logo, como parte de uma estratégia de negócios cuidadosamente delineada para ampliar seu valor econômico e político, as plataformas digitais deixaram de ser apenas sites e ferramentas acessadas isoladamente para se consolidarem como motores de um novo ecossistema informacional – uma espécie de sistema operacional sobre o qual toda a internet passa a funcionar (Kirkpatrick, 2011). Na internet plataformizada, sites não são só mais repositórios para publicação e consumo de informação, mas componentes de *software* dependentes da colaboração e participação visando a expansão deste novo ecossistema para outros espaços de interação online (Helmond, 2015).

No processo de plataformização, plataformas digitais geram valor à medida que atraem mais usuários, graças aos chamados efeitos de rede (Napoli, 2019). Os efeitos de rede estão entre as características mais marcantes das plataformas, que buscam a consolidação de estruturas monopolísticas a partir da atração contínua de novos usuários (Warnke; Maier; Gilbert, 2024). Nesse contexto, algoritmos de ranqueamento, recomendação e distribuição de conteúdo desempenham um papel central, moldando a experiência dos usuários e incentivando-os a permanecer em determinada plataforma mesmo quando existem alternativas

disponíveis (ver Santini, 2020). Essa dinâmica protege as plataformas dominantes de ameaças concorrenciais e as incentiva a escalar rapidamente, de modo a reforçar sua posição no mercado (Napoli, 2019).

Na vasta transferência de dados promovida pelas plataformas digitais no contexto da plataformização, combinada aos intensos efeitos de rede que as sustentam, elas se tornaram os equivalentes contemporâneos aos monopólios das ferrovias, telefonia e eletricidade dos séculos XIX e XX, originando uma nova economia fundamentada em suas lógicas comerciais e técnicas (Poell; Nieborg; van Dijck, 2019). Flew (2018) aponta os impactos mensuráveis da plataformização: em 2016, as chamadas *Big Tech* – o quinteto norte-americano *GAFAM*: Google (da *holding* Alphabet), Amazon, Facebook (hoje, Meta, que também controla plataformas como Instagram e WhatsApp), Apple e Microsoft – haviam se tornado as maiores empresas do mundo em valor de mercado; Google e Facebook controlavam cerca de 70% de todo o tráfego web global; o Google detinha aproximadamente 90% do mercado de buscas online, ao passo que o Facebook concentrava 80% do tráfego social *mobile*.

Como coloca Napoli (2019), as principais plataformas digitais atualmente ocupam uma posição tão dominante quanto a das instituições financeiras norte-americanas consideradas "grandes demais para quebrar" ("*too big to fail*") antes da crise financeira de 2007–2008. Por um lado, elas potencializam movimentos democráticos ao atuar como veículos para a disseminação de informações, revigorando a participação, facilitando a ação coletiva e oferecendo a grupos normalmente silenciados a possibilidade de alcançar um público amplo (Tucker *et al.*, 2017). Por outro, as relações de poder entre as plataformas e usuários são extremamente voláteis e desiguais (Poell; Nieborg; van Dijck, 2019). Também vale destacar que os benefícios de um público global conectado em rede são, por vezes, tão evidentes quanto seus riscos: o mesmo espaço que amplifica mensagens inspiradoras também pode ser usado para disseminar conteúdos repreensíveis ou até ilegais (Gillespie, 2018b).

As empresas que operam as maiores plataformas digitais e de redes sociais promovem ativamente a noção de que sua atuação inaugura uma nova era, rompendo com modelos anteriores de mediação e comunicação. Amparadas por um discurso centrado na promoção do "bem" em escala global, essas empresas defendem que sua atuação deve ocorrer à margem de regulações estatais, sob o argumento de que qualquer limitação comprometeria valores abstratos como "inovação" e "disrupção" (Eisenstat; Gilman, 2022). Essa narrativa legitimou a consolidação de um modelo de negócios inédito, fundado na vigilância contínua e na extração, modelagem e exploração econômica de dados dos usuários, viabilizando lucros

bilionários para um setor cuja rentabilidade, em um primeiro momento, fora amplamente questionada pelos próprios agentes envolvidos.

2.2 O modelo de negócios baseado em vigilância

As plataformas de redes sociais não apenas hospedam e distribuem conteúdos e interações no ambiente online, mas também operam com base em modelos comerciais voltados à monetização do engajamento que essas dinâmicas geram (Flew; Martin; Suzor, 2019). Neste trabalho, interpretamos tais práticas a partir de uma abordagem crítica, fundamentada na Economia Política da Informação e da Comunicação. Para sustentar essa lógica de negócios, as *Big Tech* e outras empresas-mãe de plataformas digitais desenvolveram e consolidaram uma nova forma de acumulação característica do contexto contemporâneo: o Capitalismo de Vigilância. Segundo Shoshana Zuboff (2015), professora e pesquisadora que cunhou o termo, o sucesso do capitalismo ao longo dos últimos séculos sempre dependeu do surgimento de novas dinâmicas de mercado, que promovem a ideia de que melhor atendem às necessidades e demandas em constante transformação de diferentes populações. Sendo assim, o Capitalismo de Vigilância não pode ser limitado a avanços tecnológicos e informacionais, definindo-se, de fato, pelas maneiras como determinadas tecnologias são desenvolvidas e exploradas para induzir ações específicas dos usuários (Zuboff, 2019), ao mesmo tempo em que prometem uma experiência online mais personalizada (Bezerra, 2017; Bromell, 2022).

Com o objetivo de prever e modificar o comportamento das pessoas para gerar lucro, o Capitalismo de Vigilância tem origem na expropriação da experiência humana, materializada na extração de dados dos usuários de novas tecnologias da informação e comunicação como buscadores online e plataformas de redes sociais (Zuboff, 2015, 2019). Aqui, enfatizamos a ideia de extração, pois ela evidencia relações sociais marcadas pela indiferença: trata-se de um processo unilateral, não dialogado (Zuboff, 2015). Para a construção de *Big Data*[3], cada ação desempenhada por um usuário na internet importa, por mais insignificante que ela possa parecer: cada clique, cada busca, cada erro de digitação, cada segundo gasto assistindo a um vídeo ou lendo um texto.

[3] Segundo o próprio Google, a noção de *Big Data* se refere a "coleções extremamente grandes e diversas de dados estruturados, não estruturados e semiestruturados que continuam a crescer exponencialmente com o tempo" e que, devido à sua dimensão e complexidade, não podem ser armazenadas e processadas de forma adequada por sistemas tradicionais de gerenciamento de dados (Google Cloud, [*S. d.*], n.p.). Zuboff (2015) aponta o *Big Data* como o alicerce do Capitalismo de Vigilância, daí a importância de, ao menos, expor como ele é apresentado pelos operadores desse novo regime.

Nenhuma dessas ações é descartada, pelo contrário: Zuboff (2019) aponta como, em 2009, representantes do Google admitiram, pela primeira vez, que a empresa armazenava "indefinidamente" os históricos de busca de seus usuários. Já o Twitter foi alvo de polêmicas por não excluir dados de usuários que apagaram suas contas em definitivo, alegando "não ter a capacidade" de atender a essas solicitações (Collier; Kolodny, 2022). Evidentemente, esta é uma decisão que diz menos sobre a capacidade técnica e mais sobre as intenções comerciais destas plataformas. Em razão disso, Walker, Mercea e Bastos (2019) argumentam que a algoritmização das comunidades online, especialmente com o surgimento das plataformas de redes sociais, marcou um ponto de inflexão ao transformar os públicos em rede em uma fonte inesgotável de dados a serem explorados.

Não são somente os dados aparentemente banais da experiência online destes usuários, porém, que são extraídos pelas plataformas digitais no Capitalismo de Vigilância. Zuboff (2015, 2019) elenca diversas outras fontes possíveis, como sensores digitais, dispositivos inteligentes e câmeras de segurança públicas, além de bases de dados governamentais, de instituições financeiras e educacionais, de planos de saúde e companhias aéreas. A acumulação desse vasto volume de dados é essencial, pois, embora muitos pareçam irrelevantes isoladamente, adquirem significado quando analisados em conjunto, compondo uma ampla engrenagem de vigilância. Nesse sentido, Bezerra (2017, p. 77) afirma que a vigilância digital é distribuída e exercida "de modo descentralizado, não hierárquico e com uma diversidade de propósitos e significações".

Assim, como observa Zuboff (2015), as plataformas digitais passaram a saber muito mais sobre as populações que as utilizam do que essas próprias populações sabem sobre si mesmas. Isso dá origem a relações profundamente assimétricas, em que as plataformas detêm um vasto conhecimento sobre seus usuários, enquanto estes sequer sabem quais dados estão sendo coletados e retidos (Dobber *et al.*, 2023; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024), no que Eifert *et al*. (2021) chamaram de "espelho unidirecional" ("*one-way mirror*"). Os autores argumentam que esse cenário permite às plataformas explorar os usuários em benefício próprio, recorrendo a práticas como a discriminação de preços, a manipulação de vieses cognitivos (Eifert *et al.*, 2021) e, principalmente, a veiculação personalizada de publicidade.

Com o crescente aumento das pressões para que seus serviços se tornassem lucrativos, a expertise do Google no processamento e análise de dados comportamentais foi direcionada ao desenvolvimento de um novo modelo de publicidade microssegmentada (Zuboff, 2015). A empresa passou a investir em processos de modelagem, agregação e análise dos vastos

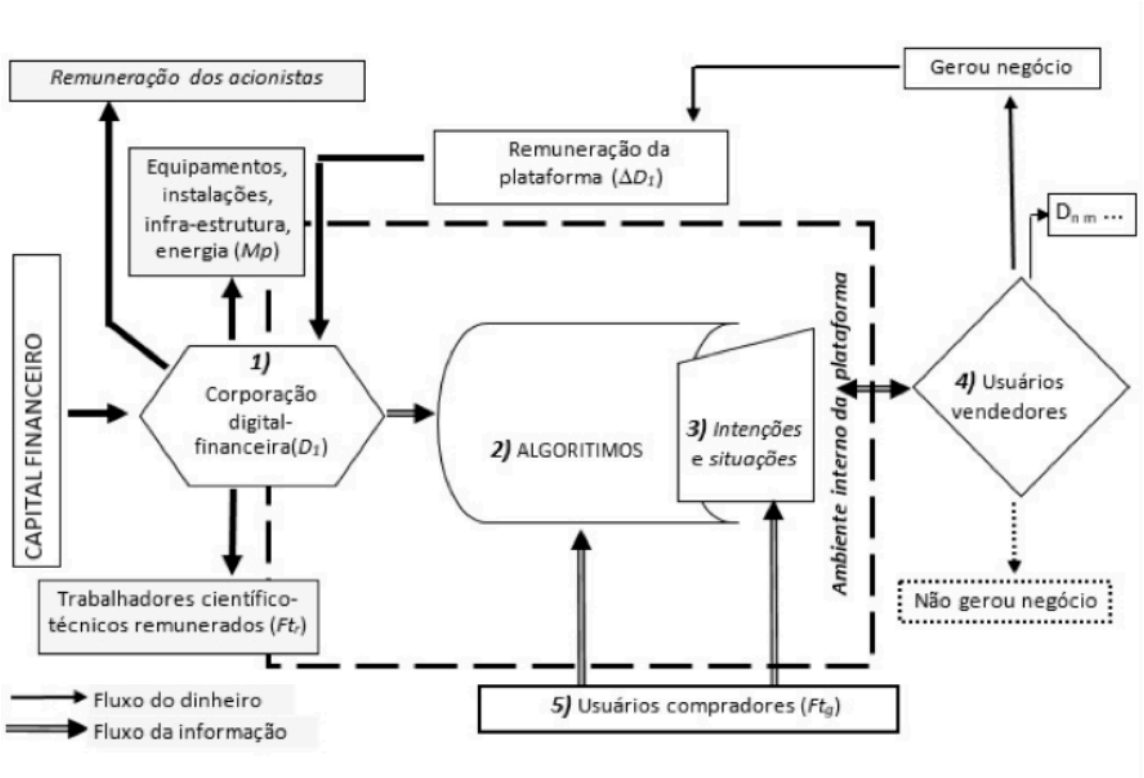
volumes de dados gerados pelas ações de seus usuários, convertendo-os em perfis comportamentais capazes de revelar interesses, ambições, atitudes e outras predisposições, que são então vendidos ao mercado publicitário (Bezerra, 2017; Bromell, 2022; Cobbe, 2021; Zuboff, 2015, 2019, 2020). Com isso, as plataformas se transformaram em "praças de mercado" que interconectam anunciantes e potenciais compradores de bens e serviços (Dantas; Raulino, 2020).

A microssegmentação (*microtargeting*) possibilita o direcionamento automatizado e personalizado de anúncios em plataformas, com base em critérios definidos pelos anunciantes, aos usuários com maior probabilidade de clicar, engajar e comprar ou contratar aquilo que é anunciado (Carah *et al.*, 2024; Papakyriakopoulos *et al.*, 2018; Ribeiro *et al.*, 2019). Para concretizar essa forma de segmentação, espaços publicitários são leiloados em frações de segundo sempre que uma página é carregada ou um aplicativo é aberto, em um processo conduzido pelos sistemas algorítmicos de plataformas de troca de anúncios (*ad exchanges*). Esses sistemas algorítmicos atuam como intermediários entre compradores e vendedores, mas operam com transparência mínima e sem responsabilização pelos seus impactos (Bromell, 2022). No momento em que um usuário acessa uma plataforma, sistemas automatizados identificam quais anunciantes têm interesse em seu perfil e realizam um leilão em tempo real para definir quem exibirá o anúncio em destaque e quanto pagará por isso (Dantas; Raulino, 2020). Conforme Rieder e Hofmann (2020), é isso que torna o Capitalismo de Vigilância uma lógica de acumulação tão única: sistemas algorítmicos de correspondência são mecanismos de ordenação que não seguem a mesma racionalidade dos processos tradicionais de tomada de decisão, gerando incertezas a todos os agentes externos às plataformas quanto a seus funcionamentos internos, performatividades e efeitos sociais mais amplos.

Embora o Capitalismo de Vigilância tenha surgido com o Google, graças a sua "operação de mercado sem precedentes pelos espaços não mapeados da internet, sem quaisquer impedimentos legais ou de competidores" (Zuboff, 2019, p. 9, tradução do autor), seus preceitos foram adotados como *modus operandi* pela maior parte das empresas que operam na economia digital. Praticamente todo serviço ou aparato tecnológico propagandeado e comercializado com as palavras "inteligente" (*smart*) ou "personalizado" depende de fluxos desobstruídos de dados comportamentais (Zuboff, 2019). Não à toa, grande parte da experiência online é sequestrada pela veiculação de publicidade personalizada, mesmo no caso de serviços pagos (Herrman, 2023). Hoje, com o duopólio do mercado publicitário online formado por Google e Meta (Fuchs, 2018; van Dijck; Nieborg; Poell, 2019), essas empresas representam a face mais visível dessa nova lógica de acumulação.

O modelo de negócios das grandes plataformas de redes sociais, fundamentado na veiculação de publicidade microssegmentada, visto na Figura 1, é precisamente esquematizado por Dantas e Raulino (2020): (1) a empresa-mãe da plataforma (por exemplo, Meta, Alphabet) adianta dinheiro (D1) para investir em infraestrutura e força de trabalho contratada e remunerada; (2) esta força de trabalho desenvolve algoritmos para traduzir a experiência online de usuários em dados a serem processados e explorados; (3) usuários vendedores e usuários compradores fornecem, a todo momento, dados sobre suas intenções e interesses, podendo fechar negócio ou não; (4) usuários vendedores são pessoas físicas ou jurídicas que oferecem produtos ou serviços e fornecem aos algoritmos dados gerais sobre seus perfis e suas ofertas para processamento no sistema de leilões; e (5) os usuários compradores também são pessoas físicas ou jurídicas que alimentam os algoritmos com dados gerais sobre seus perfis, tanto quando acessam as plataformas sem a intenção de realizar transações, quanto quando procuram por produtos ou serviços que desejam.

Figura 1 – O esquema do modelo de negócios das plataformas de redes sociais



Fonte: retirado de Dantas e Raulino (2020).

O modelo de negócios baseado na veiculação de publicidade surgiu, dessa forma, menos como fruto de um planejamento deliberado e mais como um desdobramento

“acidental” da capacidade analítica de prever o comportamento humano dessas empresas (Zuboff, 2019). Porém, vender publicidade personalizada pouco adianta se a navegação do usuário for insatisfatória. Para manter os usuários conectados – e, consequentemente, expostos a mais anúncios –, as plataformas oferecem experiências altamente personalizadas a eles em seus “jardins murados” (“*walled gardens*”) de conteúdos “exclusivos” e monetizados, e que cercam o “conhecimento e todas as práticas sociais da humanidade” (Dantas; Raulino, 2020, p. 137; Hill; Shtern, 2024).

A exploração dos dados dos usuários permite às plataformas identificar, com relativa precisão, os interesses de cada indivíduo, distribuindo informações por meio de seus sistemas algorítmicos de recomendação de conteúdo (Bezerra, 2017). No Capitalismo de Vigilância, a “personalização” da experiência online funciona como moeda de troca oferecida pelas plataformas para captar a atenção dos usuários, levando-os a relativizar as implicações da perda de privacidade diante da promessa de entrega sob medida de conteúdos relevantes a seus interesses (Bezerra, 2017; Bromell, 2022; Zuboff, 2015). Como argumentado pelo ex-CEO do Google Eric Schmidt, a privacidade online pouco importa, já que “se você faz algo do qual não gostaria que ninguém soubesse, talvez não devesse estar fazendo isso” (HuffPost, 2010, n.p.).

Como observam Dantas e Raulino (2020), de um lado estão os usuários, organizados em uma “audiência” que busca visibilidade e participação, orientada por valores de uso ligados à satisfação pessoal e ao desejo; de outro, o interesse econômico das plataformas de redes sociais em ampliar continuamente oportunidades de engajamento e os mecanismos de sua monetização. Tanto é que os algoritmos de recomendação de conteúdo das plataformas, protegidos por patentes e outros direitos de propriedade intelectual (Dantas; Raulino, 2020), são constantemente ajustados e refinados para organizar e apresentar conteúdos de forma a “viciar” seus usuários, garantindo assim uma vantagem competitiva frente a concorrentes (Strowel; De Meyere, 2023). Em decorrência disso, as audiências online tornam-se produtos, cuja atenção, cada vez mais escassa diante da sobrecarga informacional, é disputada como recurso estratégico para viabilizar a extração contínua de dados. Essa lógica leva plataformas e produtores de conteúdo a competir por engajamento, em uma dinâmica central à chamada “economia da atenção” (Diaz Ruiz, 2023; Napoli, 2019; Pedersen; Albris; Seaver, 2021).

Longe de ser um sistema estático, o Capitalismo de Vigilância opera por meio do aperfeiçoamento contínuo, no qual cada nova ação do usuário é processada com o objetivo de ampliar o conhecimento sobre seu comportamento e tornar os algoritmos de predição de interesses ainda mais precisos (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Grael; *et*

al., 2024). De certa forma, os usuários das plataformas de redes sociais desempenham, gratuitamente, uma espécie de "trabalho semiótico vivo", que alimenta e aprimora o "trabalho morto" realizado por sistemas algorítmicos (Dantas; Canavarro; Barros, 2014). Trabalho semiótico, pois os perfis comportamentais dos usuários derivam de signos linguísticos como textos, imagens, áudios e vídeos, processados continuamente pelas plataformas que almejam, em cada e qualquer ato, identificar um gesto "monetizável" (Dantas, 2014). Embora empreguem engenheiros e cientistas para desenvolver sistemas que maximizem a atenção dos usuários, as plataformas dependem das publicações de bilhões de pessoas sem vínculo contratual, que, ao "semear" esses territórios digitais, trabalham para gerar continuamente dados que serão explorados para fins econômicos, de maneira que

> [As pessoas alvo das mensagens publicitárias em plataformas de redes sociais], com seus posts, com suas fotos, seus vídeos, elas, pela publicação dos seus atos cotidianos e vulgares, produzem a audiência que se multiplica e multiplica, sempre que a cada ato publicado, algum outro ato será publicado em resposta. Elas substituem os artistas e jornalistas das tradicionais indústrias editoriais ou de onda. Ou seja – e aqui, a nossa hipótese –, elas também trabalham (Dantas, 2014, p. 88).

A concentração de grandes volumes de dados sobre os usuários, aliada à oferta de ferramentas de microssegmentação a baixo custo para anunciantes, confere às plataformas vantagens competitivas significativas e dificulta a entrada de novos concorrentes no mercado de publicidade online (Conselho Administrativo de Defesa Econômica, 2023). Como o uso dessas plataformas é majoritariamente "gratuito" para os usuários, os anunciantes tornam-se os principais pagadores para explorar seu potencial econômico (Bromell, 2022; Dantas; Canavarro; Barros, 2014; van Dijck; Nieborg; Poell, 2019). Por essa razão, Gorwa (2024) caracteriza as plataformas como mercados multilaterais, pois elas não atendem apenas aos usuários, mas também medeiam sua exploração por terceiros em arranjos mercadológicos que influenciam profundamente a distribuição de poder econômico e riqueza, impulsionados por intensos efeitos de rede (Poell; Nieborg; van Dijck, 2019). Embora os usuários aceitem os termos de serviço, o que pode ser interpretado como uma relação jurídica de consumo (Brega, 2023), são os anunciantes e outras empresas que exploram comercialmente as análises de dados das plataformas os verdadeiros clientes, reforçando as assimetrias das relações desenvolvidas no Capitalismo de Vigilância (Bromell, 2022; Zuboff, 2015).

Quando os efeitos de rede que impulsionam o crescimento das plataformas não são compensados por outras dinâmicas de mercado, poucas conseguem se consolidar, dificultando a entrada e a permanência de novos concorrentes, mesmo que estes ofereçam serviços tidos como superiores ou mais inovadores (Eifert *et al*., 2021). Esse crescimento desproporcional posiciona as grandes plataformas como guardiãs do acesso a serviços essenciais, como a

comunicação interpessoal e a circulação de informações de interesse público (Eifert *et al*., 2021). Diante de sua escala inédita, essas empresas pouco demonstram interesse em discutir modelos de negócios alternativos, alegando que qualquer mudança estrutural comprometeria a experiência dos usuários, hoje altamente dependentes de seus serviços e das conveniências proporcionadas pela mediação algorítmica (Culpepper; Thelen, 2019; Flew; Gillett, 2021; Stockmann, 2022). À medida que acumulam um poder econômico sem precedentes, sustentado por relações profundamente assimétricas, torna-se inevitável refletir sobre os contornos da governança dessas plataformas, marcada, em grande parte, por mecanismos autogovernados que acompanham e reforçam a lógica de seu modelo de negócios.

2.3 A governança de plataformas de redes sociais

A aplicação do conceito de *governança* aos estudos de plataformas digitais e de redes sociais, mais especificamente, é um fenômeno relativamente recente na literatura acadêmica, tendo se consolidado durante a última década. Como demonstra Gorwa (2019b), essa mudança se deve à ampliação da própria noção de governança, que deixou de ser vista como uma atribuição exclusiva de atores estatais e passou a incluir também as ações e mecanismos adotados por empresas e entes privados, particularmente no que se refere à sua atuação em nome do interesse público (Napoli, 2019). Outrora, entendia-se a governança apenas como a habilidade de um governo para formular e aplicar leis e fornecer serviços à população, de modo que a chamada "boa governança" se traduzia na capacidade dos Estados de construir instituições funcionais para a manutenção da lei e da ordem pública (Gorwa, 2019b).

A governança de plataformas de redes sociais integra o campo mais amplo da *governança digital*, que se refere a questões específicas relacionadas a tecnologias, aplicações e serviços que utilizam os protocolos e padrões da internet, mas que não constituem a internet em si (Komaitis; Carter, 2023). A governança digital surge da necessidade de gerenciar os elementos que moldam os hábitos online cotidianos dos usuários, por meio de estruturas capazes de responder continuamente aos desafios sociais complexos impostos por essa dinâmica, assegurando padrões mínimos de segurança, qualidade e eficiência (Komaitis; Carter, 2023). Mesmo sendo frequentemente confundidas, a governança digital é um desdobramento da *governança da internet*, não um sinônimo. Esta última diz respeito ao desenvolvimento e à aplicação de princípios, regras e procedimentos de tomada de decisão em um modelo multissetorial, que reúne governos, setor privado e sociedade civil, com o objetivo de moldar a evolução e o uso da internet como um todo (Komaitis; Carter, 2023). Em suma, a

governança da internet antecede a governança digital, por tratar de como a internet é usada, desenvolvida e administrada de forma multissetorial. A governança digital, por sua vez, diz respeito à gestão das tecnologias, aplicações e serviços que operam sobre essa infraestrutura.

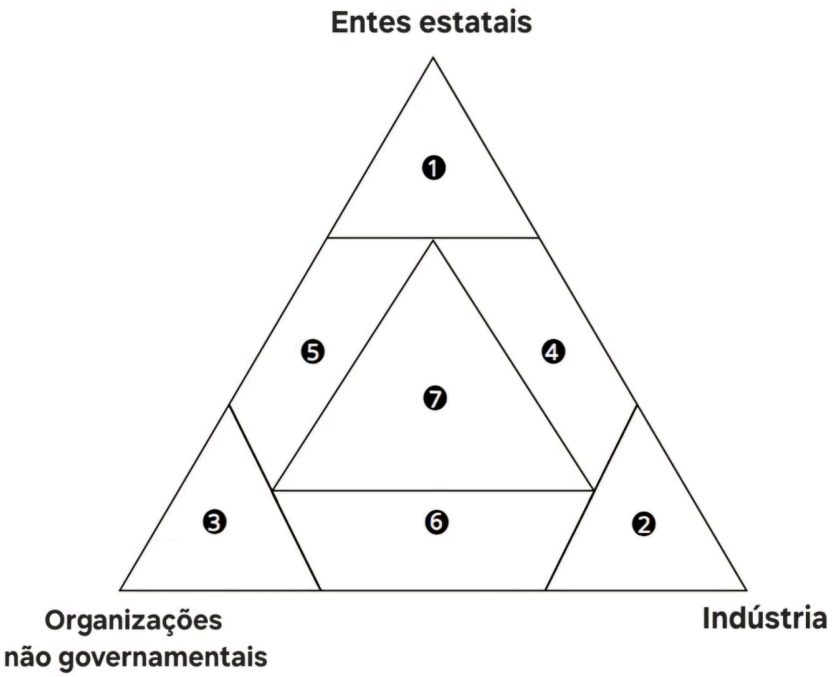
Gorwa (2019b) argumenta que as políticas e diretrizes estabelecidas pelas plataformas, aceitas por seus usuários, juntamente com algoritmos, interfaces e demais elementos de sua arquitetura digital, constituem os principais mecanismos que moldam sua governança. Partindo também da perspectiva de uma governança multissetorial, Nahon (2015) destaca que cada decisão relacionada ao modo como as plataformas de redes sociais são concebidas, bem como à produção, compartilhamento, distribuição e acesso às informações nelas veiculadas, envolve múltiplos atores com interesses diversos, que disputam posições normativas distintas. Assim, a governança de plataformas pode ser compreendida como o conjunto de relações políticas, legais e econômicas que estruturam as interações entre usuários, empresas de tecnologia, governos e outras partes interessadas no ecossistema informacional digital (Gorwa, 2019a, b). Trata-se, portanto, de um processo dinâmico e intrinsecamente político, marcado por disputas entre diferentes grupos que, enquanto persistem, definem o que pode ser feito ou dito nesses ambientes – e em quais circunstâncias (Nahon, 2015).

Para facilitar a compreensão desse panorama, Gorwa (2019a) propõe o "triângulo da governança de plataformas digitais", adaptado do trabalho de Abbott e Snidal (2009). Estes autores propuseram uma nova forma de interpretar a nova "governança transnacional" e multissetorial de atores privados e corporativos, que são governados não só pelo Estado, mas pela própria indústria e atores da sociedade civil, passando a firmar acordos voluntários e não vinculantes para regular a própria atividade (Abbott; Snidal, 2009). Gorwa (2019a, b), então, aplica essa ideia às plataformas digitais, entendendo sua governança como o resultado de uma complexa rede de interações entre diversos atores e interesses, mais do que a habilidade de um agente específico.

Como mostra a Figura 2, esses agentes podem atuar de forma independente ou articulada, buscando equilibrar suas demandas. Os números de 1 a 7 indicam as diversas formas de atuação, pressão e articulação nesse contexto. Os números 1, 2 e 3 correspondem, respectivamente, à atuação individual de entes estatais, da indústria e de organizações não governamentais. Já os números 4, 5 e 6 indicam parcerias entre dois desses três setores: entre entes estatais e indústria; entre entes estatais e organizações não governamentais; e entre organizações não governamentais e indústria. Por fim, o número 7 representa a atuação conjunta das três partes na definição dos rumos da governança de plataformas. É importante reconhecer que a "indústria" não se limita às plataformas e suas empresas-mãe, mas inclui

também os chamados complementadores, como *data brokers*, anunciantes, desenvolvedores e outros agentes essenciais ao pleno funcionamento da internet plataformizada e que contribuem para configurar as plataformas como mercados multilaterais (Gorwa, 2019b; Poell; Nieborg; van Dijck, 2018). Já as "organizações não governamentais" incluem grupos de *advocacy* e de defesa dos direitos digitais, acadêmicos e pesquisadores, e jornalistas investigativos (Gorwa, 2019b).

Figura 2 – "O triângulo da governança de plataformas digitais"



Fonte: retirado e traduzido de Gorwa (2019a).

Nesse contexto, Gillespie (2018b) propõe que as questões relacionadas à governança dos conteúdos que circulam nas plataformas de redes sociais podem ser divididas em duas dimensões: a governança *das* plataformas e a governança *pelas* plataformas. A primeira refere-se à forma como essas plataformas (não) são reguladas por normas e mecanismos externos – historicamente, um emaranhado de medidas informais e formais cuja hierarquia é difícil de discernir (Hill; Shtern, 2024); a segunda, à maneira como elas próprias se governam, diante da ausência ou fragilidade desses marcos normativos.

Desde a popularização da internet, na década de 1990, a circulação online de conteúdos classificados como "violentos" ou "explícitos" tem sido uma questão de preocupação pública (Gillespie, 2018a, b). Esse debate ganhou força, inicialmente, nos EUA,

quando ainda predominavam as primeiras ferramentas de comércio eletrônico e os fóruns de discussão online, antes do surgimento das plataformas de redes sociais (Morrison, 2023). À época, o primeiro grande marco normativo imposto a prestadores de serviços online foi a controversa Seção 230 do *Communications Decency Act* (CDA), aprovado em 1996 nos EUA, chamada por Morrison (2023) de "a espinha dorsal legal da internet". Em linhas gerais, o CDA criminalizava a distribuição de material online considerado "obsceno ou indecente" para menores de 18 anos, prevendo, ainda, punições para casos de ameaças e assédio na internet (Gillespie, 2018a). Mais especificamente, sua Seção 230 determinava que esses prestadores de serviços e outros intermediários online não poderiam ser legalmente responsabilizados pelas falas e ações de seus usuários, ou seja, por sua má utilização por parte de terceiros, atribuindo exclusivamente a estes a responsabilidade pelo conteúdo compartilhado (Klonick, 2017; Morrison, 2023). Mais do que isso, no que ficou conhecido como a "cláusula de porto seguro" ("*safe harbor provision*"), a Seção 230 também concedeu às empresas a possibilidade de intervir nas falas de seus usuários, desde que de boa-fé, sem que esse tipo de curadoria implicasse em novas responsabilidades legais (Gillespie, 2018a).

Dessa forma, havia a *liberdade* para intervir, mas não a *obrigação* – caso essa obrigação existisse, elas seriam forçadas a monitorar todas as falas e ações de seus usuários, uma tarefa extremamente onerosa para seus negócios. Isso era especialmente relevante em um contexto voltado à proteção dessas empresas da nova economia digital e informacional, evitando custos operacionais elevados que poderiam desestimular investimentos em inovação (Eisenstat; Gilman, 2022; Gillespie, 2018b; Zittrain, 2019). Concretamente, a Seção 230 criou um dilema, com repercussões significativas até hoje: ela garantiu imunidade a atores bem-intencionados, que visam proteger a segurança em ambientes online, como também garantiu a proteção destes mesmos atores contra formas colaterais de censura (Klonick, 2017). Menos de um ano depois, a Suprema Corte dos EUA declarou o CDA inconstitucional, por entender que ele ameaçava o direito à liberdade de expressão de cidadãos estadunidenses (Gillespie, 2018a), mas a Seção 230 permaneceu e permanece em vigor.

Já a União Europeia instituiu em 2000 a *e-Commerce Directive* (ECD), com o objetivo de alinhar as diversas legislações nacionais voltadas à então emergente "sociedade da informação", cobrindo serviços como buscadores, hospedagem de sites e provedores de rede (Gorwa, 2021). Entre os pontos regulamentados de forma convergente, destaca-se o regime de responsabilização dos serviços digitais, visando reduzir conflitos jurisdicionais entre os Estados-membros, já que as empresas enfrentavam incertezas quanto às normas nacionais aplicáveis a serviços transfronteiriços (Gorwa, 2021). O objetivo da ECD era estabelecer um

regime equilibrado, que protegesse os direitos dos usuários sem tolher as liberdades operacionais dessas novas empresas, permitindo que florescessem e impulsionassem um ecossistema inovador (Frosio, 2024).

Para tanto, a diretiva instituiu um regime limitado para provedores de serviços digitais, fundado em dois grandes princípios: (i) *notice and takedown* (notificação e derrubada), que estabelecia que os serviços digitais só poderiam ser responsabilizados pela circulação de conteúdo ilegal quando tivessem conhecimento efetivo de sua disseminação, devendo, a partir disso, removê-lo sem a necessidade de ordem judicial; e (ii) *no monitoring obligation* (obrigação de não monitoramento), que impedia os Estados-membros de exigir que os provedores monitorassem ativamente todos os conteúdos transmitidos ou armazenados, sem prejuízo de exigências específicas ou ordens judiciais (Frosio, 2024). Ou seja, garantia-se uma possibilidade de responsabilização em caso de conteúdo ilegal, contrastando radicalmente com a Seção 230 dos EUA, que concede imunidade irrestrita nesses casos. De qualquer forma, esses princípios, em conjunto, incentivaram as plataformas a evitarem a supervisão do que terceiros publicavam, a fim de não assumirem eventual responsabilização (Brega, 2023).

No Brasil, o primeiro regime de responsabilização de plataformas digitais foi definido pela Lei nº 12.965/2014, conhecida como Marco Civil da Internet (MCI), sancionada naquele mesmo ano. Resultado de um longo processo de debate sobre a regulação do uso da internet no Brasil, o MCI é frequentemente descrito como uma espécie de "constituição da internet", ao reconhecer o acesso à rede como instrumento fundamental para o exercício da cidadania (Freitas Aquino, 2015). Com sua promulgação, ele passou a substituir o Código Civil e o Código de Defesa do Consumidor como principal referência legal para a proteção dos direitos dos usuários contra ilegalidades praticadas no ambiente digital (Freitas Aquino, 2015).

Por pouco mais de uma década, seu artigo 19 determinou que

> com o intuito de assegurar a liberdade de expressão e impedir a censura, o provedor de aplicações de internet somente poderá ser responsabilizado civilmente por danos decorrentes de conteúdo gerado por terceiros se, após ordem judicial específica, não tomar as providências para, no âmbito e nos limites técnicos do seu serviço e dentro do prazo assinalado, tornar indisponível o conteúdo apontado como infringente, ressalvadas as disposições legais em contrário (Brasil, 2014).

Nesse sentido, o artigo 19 do MCI se encontrava no meio do caminho entre a Seção 230 do CDA, inspiração determinante para a sua redação (Nemer, 2025; Stroppa *et al.*, 2022), e o princípio de *notice and takedown* da ECD. Por um lado, não concedia às plataformas digitais, incluindo as de redes sociais, uma imunidade ampla e irrestrita em relação ao que terceiros dizem ou fazem em seus espaços. Por outro, tampouco estabelecia sua responsabilização a partir do mero conhecimento da ocorrência de ilegalidades e/ou outros

conteúdos violadores de direitos individuais e coletivos, prevendo-a apenas em caso de descumprimento de ordem judicial específica que determinasse a retirada de uma publicação ou de um conjunto de publicações – uma posição amplamente defendida pelas próprias plataformas (BBC News Brasil, 2025). Com isso, não havia a estipulação de um prazo para a remoção de conteúdo ilegal, sendo este estabelecido por cada decisão judicial.

Em resumo, o artigo 19 do MCI determinava quatro condições para a responsabilização civil de uma plataforma de rede social por conteúdo gerado por terceiros no Brasil: (i) pedido de notificação judicial formulado por pessoa que alegava ter tido seus direitos violados; (ii) avaliação judicial do potencial lesivo do conteúdo; (iii) decisão judicial que exigia a retirada do conteúdo da plataforma, com indicação do conteúdo e prazo para cumprimento; e, por fim, (iv) o descumprimento da decisão (Brega, 2023). Somente dois tipos de conteúdo escapavam a essas determinações: aqueles que violavam a Lei de Direitos Autorais vigente no Brasil[4] e conteúdos de caráter íntimo divulgados sem o consentimento das partes envolvidas (Brega, 2023).

No entanto, para Schreiber (2022), o artigo 19 era um mecanismo extremamente rígido, que protegia muito mais as plataformas do que os cidadãos brasileiros. Para o autor, a necessidade de recorrer a ordens judiciais para a remoção de conteúdos ilegais, ao invés da adoção de um mecanismo de *notice and takedown*, burocratizava a responsabilização das plataformas e acarretava disputas judiciais potencialmente intermináveis e desgastantes (Schreiber, 2022). Além disso, a possibilidade de recorrer judicialmente para a remoção de conteúdos online já estava prevista no país por meio do Código Civil e do Código de Defesa do Consumidor. Caso as plataformas descumprissem a decisão judicial, poderiam ser responsabilizadas pelo crime de desobediência, conforme previsto no Código Penal Brasileiro (Freitas Aquino, 2015; Schreiber, 2022). Assim,

> ao condicionar a responsabilização civil ao descumprimento de "ordem judicial específica", o artigo 19 promove uma proteção gravemente limitada aos direitos dos usuários da internet – frequentemente, direitos fundamentais expressamente garantidos pela Constituição brasileira, como a honra, a imagem e a privacidade (Schreiber, 2022, p. 264, tradução do autor).

Todo esse cenário começou a mudar quando, no final de 2024, o Supremo Tribunal Federal (STF) iniciou o julgamento sobre a constitucionalidade do artigo 19 do MCI. O julgamento foi concluído durante a redação deste trabalho, com o placar de 8 a 3 pela sua inconstitucionalidade *parcial*, abrindo caminho para a adoção de um sistema de *notice and*

[4] Disputas em torno do compartilhamento de conteúdos protegidos por direitos autorais tiveram um impacto considerável sobre a governança de plataformas digitais. No entanto, este é um ponto que não será aprofundado neste trabalho. Para mais sobre essa discussão, ver Gillespie (2018a).

takedown no Brasil (Vivas, 2025). A corte determinou que as plataformas estarão sujeitas à responsabilização civil se for comprovado que deixaram de adotar, em tempo hábil, medidas de prevenção ou remoção de conteúdos que configurem crimes graves, tais como publicações que promovam tentativas de golpe de Estado, terrorismo, racismo, homofobia e crimes contra mulheres e crianças (Supremo Tribunal Federal, 2025). No caso de crimes em geral ou outros atos ilícitos, as plataformas serão responsabilizadas civilmente apenas se, após o recebimento de um pedido de retirada, deixarem de remover o conteúdo. Já no que diz respeito a crimes contra a honra, permanece o entendimento de que a responsabilização civil das plataformas só ocorrerá se estas descumprirem uma ordem judicial para a remoção do conteúdo (Supremo Tribunal Federal, 2025).

Diante do exposto, Gillespie (2018b) aponta para três grandes tipos de regimes de responsabilização que ditam a governança *das* plataformas de redes sociais em todo o mundo: (i) o regime de imunidade ampla, abordagem estadunidense garantida pela Seção 230 e calcada em um entendimento particular do direito à liberdade de expressão; (ii) o regime de responsabilidade condicional, vigente em partes da Europa e da América do Sul, segundo o qual as plataformas de redes sociais não podem ser legalmente responsabilizadas pelo que seus usuários dizem ou fazem, contanto que elas não tenham conhecimento factual do que eles dizem ou fazem; e (iii) o regime de responsabilidade objetiva, que requer que plataformas de redes sociais e outros intermediários online atuem proativamente para restringir a circulação de conteúdo considerado ilegal, adotado em regiões do Oriente Médio e em países como a China.

Embora hoje atendam a bilhões de usuários globalmente, as plataformas de redes sociais, em sua maioria sediadas nos EUA, buscam operar com o máximo de liberdade e o mínimo de responsabilidade possível (Gillespie, 2018a). Mesmo não tendo sido concebidas no contexto da Seção 230, essas empresas a utilizam como escudo legal para evitar responsabilizações pelos riscos e problemas que ajudam a perpetuar (ver Cramer, 2020). Essa legislação permite que regulem o conteúdo segundo suas próprias políticas internas, exportando a todo o mundo uma estrutura aparentemente “sem leis”, na qual atuam como verdadeiros governantes do discurso online (Hill; Shtern, 2024; Klonick, 2017; Suzor, 2019). Com exceção do regime de responsabilidade objetiva, tanto o modelo de imunidade ampla quanto o de responsabilidade condicional permitem que as plataformas operem com relativa independência, mantendo-se a uma distância considerável de pressões externas consideradas indesejadas por seus executivos. Em última análise, para as plataformas, em diversos aspectos, é mais vantajoso submeter-se a certas exigências legais e governamentais de

remoção de conteúdo, tema que será aprofundado no próximo capítulo, do que assumir responsabilidade legal por todas as ações de seus usuários (Gillespie, 2018b).

É essa relativa liberdade para operar, garantida por dispositivos como a Seção 230, a ECD e o MCI (Tomaz, 2023), que dá origem à governança *pelas* plataformas, caracterizada por um elevado grau de autogovernança. São os limites difusos ou inconsistentes das políticas e regulações locais que abrem espaço para que essas empresas definam e imponham seus próprios mecanismos de governança (Gorwa, 2019b). Neste trabalho, entendemos a autogovernança como o conjunto de regras e diretrizes que uma ou mais empresas definem e aplicam por conta própria, com o objetivo de gerir e controlar suas próprias atividades – um modelo que, ao longo do tempo, demonstrou ser eficaz para viabilizar o crescimento econômico sem precedentes das plataformas de redes sociais (Klonick, 2017; Leone de Castris, 2024; Strowel; De Meyere, 2023; Warnke; Maier; Gilbert, 2024). Na prática, a autogovernança afasta o risco de interferência externa e permite que as plataformas, em grande parte dos casos, evitem responsabilizar-se legalmente pelas ações de seus usuários, tomando decisões importantes sobre elas sem obrigações vinculantes de prestação de contas (Gorwa, 2019b; Suzor, 2019). Assim, as plataformas desenvolveram um sistema "voluntário" de autogovernança, centrado em práticas não obrigatórias de moderação de conteúdo, com o intuito de criar um ambiente seguro e acolhedor que maximize o engajamento dos usuários, que serão examinadas a fundo no Capítulo 3 (Klonick, 2017).

A despeito da imagem de neutralidade frequentemente associada às plataformas digitais, seus distintos regimes de autogovernança, baseados em regras de conduta e princípios próprios, evidenciam o papel ativo que desempenham na mediação e distribuição de conteúdo, muitas vezes com impactos mais profundos sobre os usuários do que marcos regulatórios formais (Gillespie, 2018b). Por muito tempo, as plataformas de redes sociais venderam a ideia de que elas seriam entes abertos, imparciais e não intervencionistas, voltados ao empoderamento de cidadãos, principalmente como uma maneira de evitar interferências sobre sua governança (Gillespie, 2018a, b; Nahon, 2015).

Assim, tais quais empresas telefônicas, que apenas ajudam a transmitir informações sem interferência ou controle sobre o que é dito por seus clientes, seriam "meras intermediárias" ("*mere conduits*") (Gillespie, 2018a; Gorwa, 2024; Napoli, 2019; Popiel, 2022). O argumento central é que essas plataformas meramente hospedam a expressão pública, orientadas por um ideal de liberdade de expressão irrestrita, ao mesmo tempo em que delegam aos seus algoritmos a definição do que se torna popular – e, portanto, visível – com base em critérios apresentados como técnicos e neutros, mas cuja lógica permanece opaca e

inacessível à maior parte do público (Cobbe, 2021; Gillespie, 2018a, 2010). Esse entendimento é adotado por marcos normativos como a Seção 230 e o MCI (Brega, 2023; Schreiber, 2022).

No entanto, as plataformas não apenas hospedam o conteúdo de seus usuários: elas também o incentivam, facilitam e amplificam (Gillespie, 2018c), simultaneamente limitando-o de diversas maneiras, vide Capítulo 3. O serviço oferecido pelas plataformas vai além da simples hospedagem de conteúdo, abrangendo também sua organização e entrega, tornando-o buscável e permitindo que algoritmos definam o que aparece nas páginas iniciais, nos *feeds* e nas recomendações personalizadas de cada usuário (Gillespie, 2018c). Isso posto, por mais que prometam imparcialidade e neutralidade,

> a partir do momento em que passaram a introduzir funcionalidades como perfis, comentários em tópicos, etiquetas para categorizar, ordenar e buscar publicações, e, sobretudo, quando deixaram de exibir o conteúdo apenas em ordem cronológica reversa para destacar aquilo que era considerado popular, [as plataformas de redes sociais] deixaram de simplesmente entregar conteúdo aos usuários, e passaram a construí-lo (Gillespie, 2018a, p. 42, tradução do autor).

Tendo isso em vista, Nahon (2015) defende que o envolvimento político e os vieses são partes constitutivas das plataformas de redes sociais, de modo que a "neutralidade" seria sua exceção, e não a regra. Mesmo nos raros casos em que plataformas menores adotam uma postura de não intervenção absoluta, essa decisão revela uma orientação ideológica nítida sobre como pretendem se posicionar e atuar (Åkerlund, 2023). Toda a arquitetura digital das plataformas incorpora valores embutidos em seus códigos, desenvolvidos por profissionais que seguem diretrizes definidas pelas empresas, orientadas por objetivos voltados à regulação do comportamento dos usuários (Nahon, 2015). Assim,

> a arquitetura das plataformas de redes sociais é um ato consciente, deliberado e não neutro, realizado por diversos agentes – geralmente *designers* e desenvolvedores, mas também usuários. A disputa sobre quem tem o poder de arquitetar as plataformas e suas funcionalidades, e como isso é feito, constitui uma das principais manifestações das lutas por poder e dos arranjos políticos na internet. Tecnocratas tendem a argumentar que, por ser baseada em algoritmos supostamente livres de interferência humana, a tecnologia seria capaz de construir processos neutros e não discriminatórios de forma consistente. No entanto, justamente por ser projetada por seres humanos, toda tecnologia é, por definição, intrinsecamente política, refletindo os valores e interesses de seus arquitetos – e, posteriormente, sendo moldada por seus usuários (Nahon, 2015, p. 43, tradução do autor).

A forma como a ideia de "plataforma" foi discursivamente construída tem efeitos concretos sobre sua governança. Ao se posicionarem como intermediárias, as plataformas não buscam apenas atender aos interesses de seus usuários, mas também conquistar a confiança de parceiros comerciais, complementadores e, sobretudo, formuladores de políticas públicas (Gillespie, 2010). A ideia de "plataforma" atua como um conceito flexível, que tensiona e

concilia diferentes modelos de serviço: entre o conteúdo gerado por usuários e o produzido comercialmente; entre o cultivo de comunidades e a veiculação de publicidade; entre a intervenção na entrega de conteúdo e a alegação de neutralidade (Gillespie, 2010). Essa ambiguidade sustenta a narrativa de que as plataformas constituem um fenômeno inédito, que não pode e tampouco deve ser enquadrado por estruturas normativas já existentes, o que lhes assegura maior liberdade para definir os rumos de sua própria governança. Gorwa (2024) aponta que, nos primórdios do Facebook, Mark Zuckerberg apresentava a ferramenta como uma "utilidade pública", tendo sido aconselhado por seus advogados a adotar a terminologia de "plataforma", já que "utilidades públicas" são amplamente reguladas em todo o mundo.

Para alguns autores, os desequilíbrios da governança de plataformas de redes sociais poderiam ser atenuados com a aplicação de marcos regulatórios de outras indústrias, como a midiática. Isso porque sua postura não neutra e o exercício crescente de funções editoriais e de curadoria de conteúdo, especialmente por meio dos opacos processos de moderação, como discutiremos no próximo capítulo, aproximam-nas, em teoria, do papel historicamente desempenhado pelas empresas de mídia (Napoli; Caplan, 2017). Nessa perspectiva, os executivos dessas plataformas estariam cientes de sua natureza midiática, mas rejeitariam essa classificação como uma estratégia para manter vantagens competitivas, uma vez que o setor de tecnologia é significativamente menos regulado do que o setor de mídia (Eisenstat; Gilman, 2022; Napoli; Caplan, 2017). Outros elementos reforçam esse argumento, como o fato de o próprio Capitalismo de Vigilância derivar do modelo de negócios dos tradicionais conglomerados de mídia, ao prometer exatamente o que essas empresas sempre buscaram: identificar e satisfazer os desejos da audiência, sustentadas por receitas publicitárias extremamente lucrativas (Napoli; Caplan, 2017; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Grael; *et al*., 2024).

Ainda que consideremos válida a provocação de enquadrar as plataformas como empresas de mídia, entendemos, neste trabalho, que elas constituem formações híbridas, resultantes da interseção entre mídia e tecnologia da informação, cuja natureza não foi devidamente antecipada pelos debates normativos internacionais (Gillespie, 2018a). A transposição de enquadramentos legais e de governança de empresas de mídia às plataformas digitais pode gerar diversos desafios, potencialmente onerando seus usuários, além de deixar várias questões específicas sobre seu funcionamento em aberto (Gorwa, 2019b). Logo, mesmo não sendo meras intermediárias como gostariam de convencer, elas ocupam uma *posição intermediária* no imaginário popular e regulatório: entre usuários com valores distintos; entre formuladores de políticas públicas e as pessoas que buscam regular; entre a

função de meros intermediários e a de curadores; entre o direito de intervir e a ausência de responsabilidade e fiscalização sobre como essa intervenção é realizada (Gillespie, 2018c).

Esse cenário torna ainda mais urgente o debate sobre a reforma da governança de plataformas, cuja expansão ocorreu em ritmo muito mais acelerado do que outras tecnologias da informação e comunicação, mas sem a correspondente carga regulatória ou de responsabilidade. Hoje, elas concentram e medeiam um volume de interações sociais muito além do que se imaginava possível, consolidando-se como monopólios com enorme poder sobre a comunicação pública e privada (Cobbe, 2021; Langvardt, 2017). Trata-se de uma configuração inédita, em que "empresas privadas estão fazendo o que, até então, só esperávamos de autoridades constituídas do poder soberano" (Solon, 2017, n.p., tradução do autor). Todavia, esse vácuo regulatório passou a ser amplamente contestado nos últimos anos, impulsionado por desafios comunicacionais graves e fenômenos sociais extremos que abalaram a esfera pública, muitos deles enraizados na combinação entre uma governança de conteúdo marcada pela autogovernança negligente e um modelo de negócios orientado à maximização do engajamento e do lucro, em detrimento do interesse público.

2.4 O cenário da desinformação e a reação regulatória

Além da Seção 230 do CDA, as grandes plataformas de redes sociais estruturaram sua governança com base em uma interpretação particular da Primeira Emenda à Constituição dos EUA, que sustenta uma defesa irrestrita da liberdade de expressão. Sob essa perspectiva, a resposta considerada mais adequada à circulação de discursos falsos ou prejudiciais não seria a censura ou a remoção compulsória de conteúdo, mas o fortalecimento da livre expressão e a promoção do debate público. Essa lógica fundamenta-se no princípio do *contradiscurso* (*counterspeech*), conforme discutido por Napoli (2019), segundo o qual a melhor forma de combater ideias problemáticas é com mais discurso, e não com sua supressão. Essa concepção deriva da metáfora do "mercado de ideias" ("*marketplace of ideas*"), aplicada às comunicações online desde os anos 1990. Nessa visão, sociedades democráticas devem incentivar a livre competição entre ideias e opiniões, assim como os mercados incentivam a concorrência entre bens e serviços (Diaz Ruiz, 2023). Parte-se, assim, do pressuposto de que todo discurso, por mais controverso que seja, deve ter espaço, cabendo à sociedade rejeitá-lo não por meio da proibição, mas por sua irrelevância ou falta de adesão (Gillespie, 2018a).

Para muitos, os primeiros anos de efervescência das plataformas de redes sociais foram carregados de um intenso otimismo, mostrando que elas poderiam atuar como ditas

"tecnologias de libertação", capazes de empoderar o "usuário comum" e impulsionar transformações sociais profundas (Dantas, 2014; van Dijck, 2020). Plataformas como Facebook e Twitter serviram como pontos de organização tanto da oposição a regimes autoritários em países como Síria, Egito e Líbia quanto de movimentos contra a austeridade e a desigualdade social em nações ocidentais como Grécia, Espanha e os próprios EUA (Tucker *et al.*, 2017). Essas movimentações eram apresentadas como o triunfo de forças "pró-democracia" heroicas, em confronto com atores vistos majoritariamente como favoráveis à censura das plataformas de redes sociais – percebidas, por sua vez, como uma ameaça ao poder desses grupos (Tucker *et al.*, 2017).

Esse otimismo, porém, teve vida curta. Cada vez mais, as plataformas foram sendo tomadas pelo que Wardle e Derakhshan (2017) batizaram de "desordem informacional" ("*information disorder*"). A desinformação, composta por mensagens falsas, cuidadosamente elaboradas para favorecer interesses políticos e financeiros específicos, passou a se espalhar em larga escala. Essa circulação impulsionou teorias da conspiração, discursos de ódio, ataques orquestrados contra indivíduos e, nos casos mais extremos, deu fôlego ao terrorismo doméstico e internacional e a movimentos como o neonazismo e a supremacia branca (Benkler; Faris; Roberts, 2018; Lewis; Marwick, 2017; Massanari, 2015; McKeown, 2017; Persily; Tucker, 2020; Tucker *et al.*, 2018). Contrariando os defensores do contradiscurso, essas ideias – compartilhadas por usuários comuns, agentes automatizados e, em muitos casos, financiadas por estruturas estatais – passaram a mobilizar multidões ao redor do mundo (Benkler; Faris; Roberts, 2018; Guess; Lyons, 2020; Persily; Tucker, 2020; Woolley, 2020).

Nesse processo, 2016 marcou um ponto de inflexão importante para compreender quando as coisas começaram a sair do controle e como o otimismo em relação às plataformas de redes sociais se esvaziou diante desse cenário de desordem. Naquele ano, a empresa britânica de consultoria política *Cambridge Analytica* esteve no centro de um escândalo por usar massivamente dados de usuários para manipulá-los com anúncios microssegmentados, elaborados para apelar a seus sentimentos e influenciar os resultados tanto das eleições nos EUA quanto do referendo sobre a saída do Reino Unido da União Europeia (Benkler; Faris; Roberts, 2018; Guess; Lyons, 2020; Strowel; De Meyere, 2023; Walker; Mercea; Bastos, 2019). A Agência Russa de Pesquisa em Internet (*Internet Research Agency*) também utilizou plataformas de redes sociais para disseminar uma grande quantidade de desinformação política durante as eleições estadunidenses de 2016, levantando suspeitas sobre os algoritmos de recomendação e distribuição de conteúdo dessas plataformas (Benkler; Faris; Roberts, 2018; Guess; Lyons, 2020). Desde então, diversos processos eleitorais mundo afora foram

marcados pela desordem informacional, incluindo as eleições presidenciais no Brasil em 2018 e 2022 (Bastos *et al.*, 2025; Recuero; Soares; Gruzd, 2020; Santini *et al.*, 2021), além de crises como a pandemia de Covid-19 (Calvo-Gutiérrez; Marín-Lladó, 2023; Kikerpill; Siibak, 2021; Tokojima Machado *et al.*, 2022).

O fenômeno da desordem informacional é alavancado por uma combinação complexa de fatores sociais (Jiang; Fang, 2019), tecnológicos (Santini *et al.*, 2018), cognitivos (Wang *et al.*, 2020), políticos (Bruns, 2021) e culturais (Arrese, 2022), que interagem dinamicamente para favorecer a produção, disseminação e recepção de conteúdos enganosos, imprecisos ou manipulativos. Napoli (2019) cita quatro grandes fatores para a consolidação deste cenário: (i) redução das barreiras de controle sobre a circulação de informações falsas, com a eliminação de antigos obstáculos econômicos e técnicos para a produção e disseminação de conteúdo no ambiente digital, a partir do "empoderamento" de usuários; (ii) maior capacidade de segmentação por parte dos produtores de informações falsas, que conseguem, com facilidade, alcançar públicos específicos e mais receptivos a suas mensagens nas plataformas de redes sociais; (iii) aceleração da propagação das informações falsas, impulsionada principalmente pela atuação coordenada de agentes voltados à manipulação da opinião pública; e (iv) dificuldade crescente de distinguir fontes de informação legítimas de falsas, fenômeno agravado por campanhas de propaganda e manipulação que criam aparências convincentes de credibilidade e autoridade. Walker, Mercea e Bastos (2019) acrescentam que a transferência da governança comunitária, antes exercida pelas próprias pessoas, para os algoritmos eliminou a base para a confiança mútua, abrindo caminho para a disseminação em larga escala de desinformação e campanhas tóxicas.

Neste trabalho, porém, mais do que apresentar e definir os principais conceitos associados ao cenário de desordem informacional, voltamo-nos a uma de suas dimensões específicas: a comercial. A desordem informacional, antes vista apenas como um fenômeno de motivação política, tornou-se uma indústria altamente lucrativa, beneficiando tanto agentes de mercado quanto as próprias plataformas de redes sociais, sustentada por princípios operacionais e pelo modelo de negócios que estruturam seu funcionamento (Diaz Ruiz, 2023; Ghosh, 2020; Warnke; Maier; Gilbert, 2024). A literatura acadêmica demonstra que conteúdos baseados em desinformação atraem mais os usuários em plataformas de redes sociais do que aquilo que pode ser considerado "comum", no que é chamado de *hackeamento da atenção*, e, por isso, são priorizados pelos seus algoritmos de recomendação e ranqueamento (Diaz Ruiz, 2023; Vosoughi; Roy; Aral, 2018; Warnke; Maier; Gilbert, 2024).

O cálculo é simples: mais viralidade e engajamento se traduzem, para as plataformas, em maiores receitas publicitárias. Na teoria, as plataformas estabelecem regras que limitam os tipos de conteúdo que podem gerar receita publicitária (Diaz Ruiz, 2023), mas a aplicação destas diretrizes é limitada e facilmente contornável (Santini; Salles; Belin; *et al.*, 2024). Nesse sentido, Jalli (2024, n.p., tradução do autor) argumenta que emergiu um ecossistema complacente e de tolerância excessiva, "no qual agentes mal-intencionados conseguem disseminar desinformação e discurso de ódio com relativa impunidade, cientes de que as plataformas de redes sociais priorizam a receita publicitária em detrimento da integridade do debate público". Portanto, se há uma relação legal de consumo estabelecida entre os usuários e as plataformas de redes sociais, é evidente que essas plataformas deixaram de protegê-los adequadamente (Santini *et al.*, 2025).

A desordem informacional, inicialmente tratada como um acidente de percurso na história das plataformas de redes sociais, acabou se consolidando como um componente central de sua arquitetura e de suas dinâmicas comerciais e financeiras. Prova disso são as denúncias feitas por *whistleblowers* da Meta, que revelaram que, entre minimizar os danos aos usuários ou maximizar o engajamento online, a empresa deliberadamente optou pelo segundo caminho em suas plataformas, devido ao alto potencial de viralização e lucro comercial que a desinformação costuma gerar (Bradshaw; Stacey, 2021). Leal, Felsberger e Neff (2024) argumentam que sistemas de recomendação baseados na viralidade não são uma imposição inevitável, mas resultado de escolhas feitas pelas plataformas – escolhas que poderiam, em vez disso, priorizar e recompensar o interesse público.

O fato de a desordem informacional ter se convertido em uma indústria lucrativa é um grande indicativo de que as plataformas de redes sociais, apesar de seus intensos esforços para serem percebidos como tais, não são entidades neutras. Afinal, trata-se de sistemas sociotécnicos baseados em algoritmos e não há neutralidade ou passividade algorítmica no ecossistema informacional online (Chander; Krishnamurthy, 2018; Reviglio; Agosti, 2020; Stinson, 2022). Se atualmente o engajamento movido pelo ódio, pela toxicidade e pela desinformação é central aos algoritmos de recomendação e distribuição das plataformas de redes sociais (Nemer, 2025), seus efeitos não podem ser subdimensionados como "produtos neutros", devendo ser compreendidos como consequência de uma combinação de práticas que, em um setor amplamente desregulado, têm gerado lucros extraordinários (Ghosh, 2020).

Com suas escolhas, conscientes ou não, que contribuíram para a desordem informacional, as plataformas alimentaram um movimento internacional de contestação às bases de sua governança. Em 2017, a revista *The Economist* cunhou um dos diagnósticos mais

citados sobre a conjuntura das *Big Tech*, ao introduzir o termo "*techlash*", uma combinação de "*technology*" e "*backlash*" (retaliação). Mesmo sem enfrentar uma crise financeira, as plataformas de redes sociais passaram a lidar com uma crise generalizada de confiança (Culpepper; Thelen, 2019; Frederick, 2021). Rapidamente, passaram a ser vistas como berços de projetos autoritários e nocivos, marcados pela interferência na soberania nacional, riscos à integridade eleitoral e pela escalada do ódio, fatores agravados por recorrentes vazamentos de dados e violações de privacidade (Flew, 2018; van Dijck, 2020). As plataformas foram submetidas a um julgamento simbólico, e seus diretores, convocados a prestar contas em audiências públicas televisionadas para todo o mundo (Flew, 2018).

Em uma das várias ocasiões em que representantes das *Big Tech* foram convocados a testemunhar no Congresso dos EUA, a então senadora Dianne Feinstein fez uma declaração que sinalizava uma virada na percepção pública e política sobre o papel dessas empresas: "Devo dizer que acho que vocês ainda não entenderam. Estamos falando de uma mudança cataclísmica... Vocês criaram essas plataformas, e elas estão sendo mal utilizadas. E vocês têm que fazer algo a respeito – ou nós faremos." (retirado de Flew, 2018, tradução do autor). No campo acadêmico, o *techlash* foi visto como uma oportunidade para analisar a reconfiguração das estruturas de regulação da comunicação digital, marcada por disputas entre plataformas, autoridades nacionais e internacionais e parceiros comerciais sobre os rumos de uma governança mais responsável (Hill; Shtern, 2024). Entre 2017 e 2021, conforme mostram Hill e Shtern (2024), 13 países publicaram mais de 80 investigações e relatórios analisando as plataformas de redes sociais, enquanto autoridades nacionais discutiam a viabilidade de possíveis intervenções regulatórias. Esses movimentos indicaram o início de uma onda regulatória global voltada às plataformas digitais, especialmente às de redes sociais, refletindo uma preocupação comum com a limitação de abusos, a proteção de direitos fundamentais e a contenção dos impactos dessas infraestruturas na esfera pública (Schlesinger, 2020).

De acordo com Ananny e Gillespie (2017), a contestação à governança das plataformas compreende a etapa inicial de uma dinâmica cíclica baseada em "crises públicas e respostas excepcionais" ("*public shocks and platform exceptions*"). Trata-se, respectivamente, da ruptura na lógica de governança promovida pelas plataformas, que revela problemas infraestruturais fundamentais ao seu funcionamento, e da indignação pública acompanhada de pressões regulatórias (Ananny; Gillespie, 2017). Esses movimentos costumam vir acompanhados de "desculpas profusas, promessas de melhorar e pequenas mudanças na experiência do usuário em suas plataformas" (Flew, 2018, p. 27, tradução do autor). Como atores comerciais privados, majoritariamente sediados nos EUA, as plataformas apelam aos

princípios do livre mercado e da liberdade de expressão para reforçar a narrativa de que podem não apenas se autogovernar, mas também corrigir as falhas desse processo com mais autogovernança, evitando, assim, intervenções estatais (Ananny; Gillespie, 2017).

As respostas das plataformas podem se dar de diferentes maneiras, visando diferentes objetivos. Notoriamente opacas e marcadas por sérios déficits de prestação de contas, como será debatido no próximo capítulo (Ananny; Crawford, 2016; Dommett, 2020; Edelson *et al.*, 2021; Geng, 2023; Klonick, 2017; Rieder; Hofmann, 2020; Suzor, 2019; Suzor *et al.*, 2019), algumas plataformas passam a introduzir mecanismos e ferramentas públicas de transparência, enquanto outras procuram se distanciar da origem de determinadas crises, como a radicalização política e ataques terroristas. Ao mesmo tempo, prometem mudanças de design e atualizações de diretrizes, em uma tentativa ambígua de reconhecer sua responsabilidade sem assumi-la plenamente (Ananny; Gillespie, 2017; Bossetta, 2020).

Após os escândalos envolvendo a publicidade microssegmentada em 2016, grandes plataformas, como as da Meta e o Twitter, lançaram suas chamadas "bibliotecas de anúncios", ferramentas de transparência que, em tese, permitiriam a usuários e formuladores de políticas públicas inspecionar e avaliar a distribuição e a circulação de anúncios políticos, identificando eventuais irregularidades e riscos (Edelson; Lauinger; McCoy, 2020; Jamison *et al.*, 2020; Kulkarni, 2018; Leerssen *et al.*, 2019). Paralelamente, as plataformas passaram a restringir o acesso a dados que antes estavam disponíveis para pesquisadores conduzirem estudos de interesse público, sob o argumento de que tal uso colocaria em risco a privacidade dos usuários; na prática, muitos dos achados dessas pesquisas serviram para expor práticas problemáticas e provocar crises públicas (Bruns, 2019; Leone de Castris, 2024; Santini *et al.*, 2025; Walker; Mercea; Bastos, 2019).

Diversas iniciativas foram adotadas pelas plataformas em parceria com organizações da sociedade civil e autoridades públicas – que, segundo Gorwa (2019a), conferem autoridade e credibilidade – com o objetivo de torná-las mais seguras e afastá-las de fontes de crise e pressão. Em 2018, a Meta criou o *Oversight Board*, um conselho independente formado por pesquisadores e representantes da sociedade civil, para supervisionar a veiculação de conteúdos problemáticos em suas plataformas. Esse órgão permite que usuários contestem decisões de remoção de conteúdo, buscando garantir a liberdade de expressão e protegê-los contra possíveis vieses ou erros de julgamento (Gorwa, 2019a; San Martin, 2023). Já em 2019, após o ataque a duas mesquitas na cidade de Christchurch, Nova Zelândia, ser transmitido ao vivo pelo Facebook sem qualquer contenção da plataforma (ver Common, 2020; Gorwa; Binns; Katzenbach, 2020; Popiel; Vasudevan, 2024; Suzor; Gillett, 2022), a

então primeira-ministra Jacinda Ardern e o presidente francês Emmanuel Macron lançaram o *Christchurch Call*, um conjunto não vinculante de compromissos para combater conteúdos terroristas online, assinado por dezoito governos e oito grandes empresas de tecnologia (Gorwa, 2019a, 2024). Por último, a União Europeia apresentou em 2018, com uma revisão em 2022, o Código de Conduta sobre Desinformação (*Code of Practice on Disinformation*), outro conjunto de diretrizes não vinculantes para endereçar a disseminação de desinformação em Estados-membros do bloco (Warnke; Maier; Gilbert, 2024).

Especialistas, tomadores de decisão, legisladores e diversos outros setores sociais, porém, demonstram cada vez menos confiança na capacidade das plataformas de resolver seus próprios problemas com ainda mais autogovernança (Brega, 2023; Gorwa, 2024). Medidas voluntárias e campanhas de relações públicas, embora bem-intencionadas, têm se mostrado flagrantemente insuficientes para enfrentar as questões sistêmicas que florescem nos domínios das plataformas (Jalli, 2024). Muitas dessas iniciativas apresentam escopo limitado e carecem de definições claras sobre as práticas e os padrões a serem seguidos pelas plataformas signatárias, o que dificulta sua implementação de forma efetiva e uniforme (Dwivedi, 2022).

Apesar dos efeitos problemáticos da lógica econômica das plataformas, as respostas iniciais concentraram-se apenas na mitigação de danos imediatos, sem questionar sua estrutura. Como resultado, a desordem informacional continuou a se agravar, uma vez que enfrentar suas causas implicaria confrontar diretamente o modelo de negócios das plataformas (Ghosh, 2020; Warnke; Maier; Gilbert, 2024). Essa condição impede que as empresas atuem efetivamente em prol do interesse público, gerando evidentes conflitos de interesse (Flew, 2018; Flew; Gillett, 2021; Gorwa, 2019a). Diante dos limites da autogovernança, a única alternativa viável passa a ser o endurecimento da governança *das* plataformas, por meio da introdução de novos mecanismos regulatórios formais (Ananny; Gillespie, 2017), em um processo dinâmico no qual diversos atores desempenham papéis e dispõem de capacidades e meios de intervenção distintos (Schlesinger, 2020).

A mudança de ênfase de um discurso positivo, centrado na natureza disruptiva, conectiva e inovadora das novas tecnologias da informação e comunicação, para uma abordagem negativa, que destaca seus danos, ameaças e a vigilância constante dos cidadãos, levou à reflexão sobre quais intervenções sistêmicas poderiam conter seus aparentes excessos; para Zittrain (2019), esse processo marca o início da terceira grande era da governança digital. A primeira, denominada “a era dos direitos”, teria sido marcada pela rejeição a qualquer tentativa de controle sobre a internet e as tecnologias decorrentes de sua expansão, especialmente por parte de governos. Muitas das garantias dessa fase foram viabilizadas pela

Seção 230 do CDA, cuja lógica foi exportada a diversos países (Zittrain, 2019). Na sequência, tivemos a "era da saúde pública", um momento intermediário em que se reconheceu que as promessas do libertarianismo e do otimismo digital, na verdade, prepararam o terreno para uma cultura tóxica, nociva e abusiva, cujo auge foi a desordem informacional, criando uma polarização entre aqueles que defendem maior controle sobre os ambientes digitais e aqueles que veem essa cultura como o "preço a pagar pela liberdade". Enfim, entramos na "era do processo e da legitimidade", quando o debate não é mais sobre *se* as novas tecnologias da informação e comunicação – em especial, as plataformas – devem ser reguladas, mas sim sobre *como* essa regulação deve ser implementada.

No entanto, regular as plataformas digitais, por mais promissor que pareça, está longe de ser tarefa simples – e tampouco há um único caminho a ser seguido ou consenso em torno disso (ver Hill; Shtern, 2024; Schlesinger, 2020). Gorwa (2024) sistematiza quatro categorias principais de propostas regulatórias, ilustradas por iniciativas com diferentes graus de implementação ao redor do mundo. A primeira diz respeito à proteção de dados pessoais, um campo em constante aperfeiçoamento desde os anos 1990 e impulsionado por escândalos como o da *Cambridge Analytica*. A *General Data Protection Regulation* (Regulamento Geral sobre a Proteção de Dados; GDPR), na União Europeia, e a Lei Geral de Proteção de Dados (LGPD), no Brasil, são marcos centrais desse esforço. A segunda categoria regulatória abrange leis de concorrência e antitruste, voltadas a conter práticas predatórias e estimular a entrada de novos agentes econômicos em um mercado altamente concentrado. O *Digital Markets Act* (DMA), da União Europeia, que será apresentado no próximo capítulo, é a principal referência desse tipo. Em terceiro lugar, destacam-se novas legislações trabalhistas voltadas à proteção de trabalhadores informais que atuam sob plataformas como Uber e Airbnb, como o Real Decreto-ley 9/2021, da Espanha, que busca garantir direitos mínimos a entregadores vinculados a aplicativos.

Mais diretamente relevante para este trabalho, contudo, é a quarta e última categoria destacada por Gorwa (2024): a regulação de conteúdo em plataformas de redes sociais. Essa abordagem envolve tanto a atualização de marcos já existentes sobre a responsabilidade das plataformas pelo conteúdo que hospedam quanto a formulação de novas ferramentas regulatórias que ampliem essa responsabilização, sobretudo onde ela ainda é frágil ou inexistente. A crescente responsabilização das plataformas de redes sociais por meio da regulação decorre, em última instância, do reconhecimento de que elas funcionam como infraestruturas que não apenas criaram novas condições para a circulação do discurso público e para o surgimento de novos atores de mercado, mas também geraram uma dinâmica de

dependência que torna difícil para as pessoas simplesmente deixarem de utilizá-las (Ananny; Gillespie, 2017).

Em geral, a regulação de conteúdo se concretiza por meio de regras processuais que orientam como as plataformas devem lidar com conteúdos politicamente ou socialmente sensíveis, incluindo exigências de mecanismos para tratamento de denúncias e garantias de devido processo e transparência (Eifert *et al.*, 2021; Gorwa, 2024). Em especial, mecanismos regulatórios voltados à transparência das plataformas passaram, aos poucos, a integrar essa discussão, replicando movimentos já observados em setores como o financeiro, automobilístico e de saúde, pressionados a mitigar riscos aos quais consumidores frequentemente se expunham (Vergara; Jain; Mehta, 2024). Estes mecanismos obrigam as plataformas a ampliar as medidas de transparência e abertura que, de forma autogovernada, adotaram em resposta à dinâmica de "crises públicas e respostas excepcionais", com a formalização de padrões mínimos para a fiscalização de suas práticas de governança (Dwivedi, 2022), desconstruindo a "incerteza" que caracteriza seu comportamento (Eifert *et al.*, 2021) e submetendo-as ao escrutínio público (Flew, 2018; Gorwa, 2019b). A importância das regulações de transparência reside no fato de que grandes plataformas funcionam como monopólios de dados de difícil acesso, o que impede o exame minucioso de suas práticas e deixa diversas questões ainda sem a devida documentação (Gorwa, 2019a), como aprofundaremos na sequência deste trabalho. A adoção de exigências de transparência acarreta custos operacionais para as plataformas, mas esses investimentos são modestos em comparação a eventuais multas aplicadas (Warnke; Maier; Gilbert, 2024).

Os intensos esforços do lobby pró-plataformas, sobretudo no Sul Global, onde governos dispõem de menos recursos políticos para enfrentar o poder das *Big Tech*, são um grande obstáculo ao debate regulatório. Nos EUA e na União Europeia, inclusive, o setor de tecnologia já ocupa o posto de principal agente de lobby, investindo mais recursos do que outros historicamente influentes, como os de combustíveis fósseis e farmacêutico (Gorwa; Lechowski; Schneiß, 2024). Além de estratégias direcionadas a atores políticos e tomadores de decisão, as plataformas também investem fortemente em campanhas voltadas ao público em geral, a fim de estimular a resistência popular a propostas de regulação, frequentemente apresentando os próprios usuários como os maiores prejudicados por tais medidas (Culpepper; Thelen, 2019; Yates, 2023).

No Brasil, isso ficou evidente com a tentativa de aprovação do PL 2630/2020, popularmente conhecido como o "PL das *Fake News*", que instituiria a Lei Brasileira de Liberdade, Responsabilidade e Transparência na Internet (Bueno; Canaan, 2024). Proposto

em 2020, o projeto destacava a responsabilidade das plataformas digitais, serviços de mensagens privadas e outros provedores de informação no combate à desinformação e na promoção da transparência online, partindo da premissa de que, diante dos desafios impostos pela crescente utilização dessas plataformas, seria necessário suprir as lacunas do MCI (Câmara dos Deputados, 2020). Para impedir a aprovação do projeto, as *Big Tech* alinharam-se discursiva e taticamente à extrema direita brasileira, os mesmos atores responsáveis por fomentar a desordem informacional no país nos últimos anos, valendo-se de estratégias de desinformação para manipular a opinião pública (Salles *et al.*, 2024). Por meio de ações coordenadas, buscaram consolidar a ideia de que a regulação das plataformas representaria uma ameaça à liberdade de expressão online, associando a regulação de conteúdo à imposição de limites sobre o que os usuários podem dizer e à adoção obrigatória de regras de controle de caráter autoritário pelas próprias plataformas (Salles *et al.*, 2024).

Em todo o mundo, o temor da censura permanece uma barreira significativa à aceitação pública de propostas de regulação das plataformas digitais, mesmo quando tais projetos não preveem sanções aos usuários nem a criação de novas regras de conteúdo (Bechmann, 2020; Bromell, 2022; Costa, 2023; Gorwa, 2021; Hill; Shtern, 2024; Popiel, 2018; Schlesinger, 2020; Vese, 2022; Wagner *et al.*, 2020; Zipursky, 2019). Na América Latina, em particular, as acusações de "censura" ganham força diante das tentativas de regulação das plataformas, algo que García Silva e Chanduvi (2024) atribuem ao histórico recente de censura institucionalizada durante as ditaduras militares na região, contribuindo para que legislações voltadas à reforma da governança digital permaneçam à deriva (Bizberge; Mastrini; Gómez, 2023). Califano (2023) observa que as poucas regulações do ecossistema digital que prosperaram na região tratam de temas menos controversos, como a proteção da privacidade e dos dados pessoais, a exemplo da já mencionada LGPD.

Se o STF avançou na revisão do regime de responsabilização das plataformas digitais no Brasil, isso se deve ao vácuo legislativo deixado pela tentativa frustrada de aprovação do PL 2630/2020 (Neitsch, 2024). O próprio tribunal reconhece que as recentes decisões sobre o MCI podem ser revogadas caso o Congresso Nacional decida se dedicar a essa questão (Supremo Tribunal Federal, 2025). Toda a discussão em torno da introdução de um novo marco regulatório de conteúdo para plataformas digitais no Brasil acompanha um movimento encabeçado por países europeus, em um cenário no qual os EUA indicam que não avançarão nessa questão (BBC News Brasil, 2025; Gorwa, 2024), principalmente devido ao retorno de Donald Trump à presidência (Lu, 2025). Inicialmente, países como Alemanha, França e Reino Unido lideraram o debate regulatório (Helberger, 2020), até que a União Europeia, como

bloco, passou a discutir a revisão da ECD e assumiu a dianteira na regulação das plataformas digitais. Esse tema será retomado ao final do próximo capítulo, no intuito de analisar como as discussões sobre regulação de conteúdo impactam um aspecto central da governança *pelas* plataformas de redes sociais: a moderação de conteúdo.

3 A GOVERNANÇA *PELAS* PLATAFORMAS: COMO ELAS MODERAM CONTEÚDO E OS ESFORÇOS PARA TORNÁ-LAS MAIS TRANSPARENTES

Neste capítulo, nos dedicamos à moderação de conteúdo das plataformas, buscando não apenas entender como ela ocorre, mas também revelar o que esconde: problemas crônicos de transparência que, cada vez mais, vêm sendo enfrentados pela via regulatória. Iniciamos com a **seção 3.1**, dedicada a destrinchar os processos, lógicas operacionais e estruturas que sustentam a moderação de conteúdo nas grandes plataformas de redes sociais. Na **subseção 3.1.1**, explicamos como entendemos a moderação, como ela se concretiza na prática e de que forma se insere no modelo de negócios e na estrutura de governança dessas plataformas – mais especificamente, na lógica de governança *pelas* plataformas, como discutido no capítulo anterior. Em seguida, a **subseção 3.1.2** trata do trabalho humano que sustenta esse processo, geralmente invisibilizado e terceirizado para países do Sul Global, onde uma força de trabalho precária aplica, de forma distorcida, diretrizes impostas de cima para baixo. Encerrando a seção, a **subseção 3.1.3** aborda a crescente automatização e algoritmização da moderação, discutindo os desafios e riscos associados a essa tendência, em última análise, inevitável.

Já a **seção 3.2** é dedicada à questão da transparência da moderação de conteúdo, com ênfase em sua (não) aplicação como ferramenta de governança, seja ela autogovernada ou imposta por marcos regulatórios. Na **subseção 3.2.1**, discutimos como a opacidade se consolidou como norma na governança de grandes plataformas de redes sociais e de que maneira essa opacidade se manifesta na moderação: por meio de regras pouco claras, da ausência de detalhes sobre processos decisórios e da divulgação seletiva de estatísticas, restrita a documentos voluntários. Por fim, a **subseção 3.2.2** analisa a centralidade crescente da transparência em novos projetos de regulação de conteúdo, que vêm estabelecendo critérios vinculantes para as políticas adotadas pelas plataformas. Em especial, destacamos as exigências do *Digital Services Act* (DSA) no contexto da União Europeia, que serão determinantes para a análise apresentada no capítulo seguinte.

3.1 A moderação comercial de conteúdo

3.1.1 O que é e por que moderar conteúdo?

A moderação de conteúdo está indissociavelmente ligada às estratégias de gestão de marca e reputação não apenas das grandes plataformas de redes sociais, mas também de seus parceiros comerciais. As plataformas moderam conteúdo por necessidade: primeiramente, é necessário proteger seus usuários uns dos outros, de forma a evitar um êxodo destes para outros espaços considerados "mais seguros"; em segundo lugar, é necessário que essas empresas projetem uma imagem socialmente responsável perante seus anunciantes, uma vez que poucas marcas estariam dispostas a se associar a ambientes que favorecem a proliferação descontrolada de fenômenos sociais extremos (Gillespie, 2018a; Roberts, 2016). A própria necessidade de proteger usuários uns dos outros não decorre, necessariamente, de um apreço à sua dignidade, mas do impacto que sua saída pode ter na receita publicitária (ver Bromell, 2022). Assim, por mais que a moderação de conteúdo possa ser apresentada como um processo voltado ao pleno exercício dos direitos humanos, na prática, ela é movida sobretudo por interesses financeiros, pelo receio de regulações governamentais e pela pressão pública (ver Cobbe, 2021; Díaz; Hecht-Felella, 2021; Klonick, 2017). Por esses motivos, todas as grandes plataformas de redes sociais precisam investir na moderação de conteúdo, ainda que isso implique na desconstrução da ilusão de uma internet livre e aberta (Gillespie, 2018a).

Grimmelmann (2015, p. 47, tradução e grifo do autor) define a moderação de conteúdo como um conjunto de "mecanismos de *governança* que estruturam a participação em uma comunidade de forma a facilitar a cooperação e prevenir o abuso online". No caso das plataformas de redes sociais, Gillespie *et al*. (2020) apontam que esses mecanismos são orientados pela definição de regras e normas comportamentais estabelecidas pelas próprias plataformas, que agem contra o que consideram inaceitável – a expressão máxima da "governança *pelas* plataformas" apresentada no Capítulo 2 (Gillespie, 2018b, c). Sob essa perspectiva, de acordo com Fitzgerald e Lokmanoglu (2023), a moderação de conteúdo é o meio pelo qual as plataformas de redes sociais constroem suas identidades, refletindo a cultura política na qual se originam e moldando a experiência de todos os seus usuários.

Sarah T. Roberts (2016, 2018, 2019) cunha o termo "moderação *comercial* de conteúdo" para diferenciar a moderação das grandes plataformas de redes sociais da moderação de conteúdo voluntária, característica dos primeiros fóruns e espaços de discussão online, levada a cabo por usuários que buscavam manter a "civilidade" destes ambientes (ver

também Badouard; Bellon, 2025; Gorwa, 2018; Myers West, 2018). Para a autora, a moderação comercial de conteúdo seria "a prática *organizada* de filtragem [e curadoria] de conteúdo gerado por usuários publicado em sites, plataformas de redes sociais e outros espaços online" (Roberts, 2019, p. 12, tradução e grifo do autor). Como discutiremos nas seções seguintes, essa prática pode ser realizada pelo trabalho humano, por sistemas automatizados ou por uma combinação de ambos (Fitzgerald; Lokmanoglu, 2023).

Primordialmente, a moderação comercial de conteúdo é uma *commodity* que agrega às plataformas tanto valor simbólico quanto econômico, ao lhes permitir impor coesão e organização ao caos informacional da internet não plataformizada e restringir determinados discursos, com o objetivo de aprimorar a experiência dos usuários e atender às demandas de seus anunciantes (Gillespie, 2018a). Como afirmou uma funcionária de uma grande plataforma de rede social a Moran *et al*. (2025, p. 10, tradução do autor), "é preciso moderar conteúdo; a Coca-Cola não vai querer veicular um anúncio ao lado de uma postagem supremacista branca". Flew (2018) observa que a plataformização levou a um cenário em que o uso da internet é cada vez mais mediado por buscadores e plataformas de redes sociais, gerando uma tensão entre a liberdade dessas empresas para conduzirem seus negócios de forma lucrativa e a necessidade de que atuem em conformidade com o interesse público.

A moderação de conteúdo, de responsabilidade das equipes de *Trust & Safety* (Confiança e Segurança) das plataformas de redes sociais, convencionalmente se traduz em ações como a remoção de publicações, bem como o banimento de perfis, páginas e grupos específicos – medidas tomadas por iniciativa das próprias plataformas ou em resposta a solicitações de entes governamentais, em razão do descumprimento de legislações locais (Fitzgerald; Lokmanoglu, 2023; Gillespie, 2022; Kaushal *et al.*, 2024; Klonick, 2017; Moran *et al.*, 2025). Em determinadas regiões, especialmente sob regimes autoritários, plataformas que não atendem a pedidos governamentais de remoção de conteúdo e/ou suspensão de usuários são comumente banidas (Fitzgerald; Lokmanoglu, 2023; Klonick, 2017). A cobertura midiática também influencia as plataformas a adotarem uma postura mais rigorosa em relação ao conteúdo que circula nelas, em face de denúncias sobre moderação injusta ou da falta de moderação quando necessária (Klonick, 2017). Outras práticas semelhantes de moderação incluem restringir o acesso de usuários de determinadas localidades ou faixas etárias a certos conteúdos, "etiquetar" publicações para alertar os usuários sobre seu caráter potencialmente problemático e desmonetizar os responsáveis pela publicação de conteúdos nocivos (Caplan; Gillespie, 2020; Gillespie, 2022; Kaushal *et al.*, 2024).

Ao analisar mais de 500 contestações de usuários que tiveram suas publicações e perfis removidos de plataformas de redes sociais, Myers West (2018) indica que, da esquerda à direita no espectro político, entre homens e mulheres, a percepção é de que a moderação de conteúdo é uma tentativa ampla de silenciamento e censura. No entanto, a direita – notadamente, a extrema direita – tem demonstrado hostilidade crescente aos esforços de *Trust & Safety*, com reiterados ataques a plataformas por suas iniciativas de combate à desinformação em escala global (Alizadeh *et al.*, 2022; Douek, 2021; Moran *et al.*, 2025). Assim, a governança de conteúdo de cima para baixo, por meio de ações de moderação, gerou uma crise de legitimidade para as plataformas, cujas decisões são, por diversas vezes, vistas como corruptas, arbitrárias e irresponsáveis (Moran *et al.*, 2025). Aqui, há um delicado equilíbrio: se as plataformas moderam conteúdo em excesso, os usuários perdem a confiança nelas e, consequentemente, as abandonam; se as plataformas não moderam nenhum conteúdo, os usuários não se sentem seguros o suficiente (Åkerlund, 2023; Klonick, 2017). É interessante notar também que muitos criadores de conteúdo moderados pelas plataformas constroem uma identidade a partir desse fato, apresentando-se como parte de uma resistência organizada contra as "forças da censura" e conquistando a simpatia de novos públicos (Fitzgerald; Lokmanoglu, 2023).

Entretanto, limitar a moderação a um binarismo em que o conteúdo ou permanece no ar ou é removido é uma visão reducionista. Argumentamos que *todas* as formas pelas quais as plataformas de redes sociais moldam e intervêm sobre os fluxos informacionais online para subordiná-los a seus interesses comerciais podem ser classificadas como ações de moderação de conteúdo. Cada publicação que é recomendada a um usuário é recomendada em detrimento de outras, em um processo de seleção e ordenação que constitui, essencialmente, uma prática estratégica de moderação de conteúdo (Alizadeh *et al.*, 2022). Nenhuma publicação surge espontaneamente na tela de um usuário quando ele acessa uma plataforma; isso ocorre somente após uma avaliação, humana ou automatizada, de seus méritos, que determina que ela pode ser recomendada (Roberts, 2016).

Atualmente, discute-se a redução ou amplificação do alcance e da visibilidade de determinados conteúdos, por meio de ação algorítmica ou não, como formas de moderação (ver Gillespie, 2022; Moran *et al.*, 2025; Santini; Salles; Mattos, 2023). Em primeiro lugar, a redução de alcance é uma forma de moderação menos contestada nos âmbitos público e político do que a remoção de conteúdo, pois os usuários muitas vezes sequer percebem quando ela ocorre. Nesses casos, conteúdos que não são problemáticos o suficiente para serem removidos permanecem disponíveis e podem ser encontrados em buscas, mas suas condições

de circulação são limitadas (Gillespie, 2022). Além disso, a redução de alcance não deixa rastros e é mais difícil de detectar do que o banimento de usuários ou a remoção de publicações (Gillespie, 2022). Já a amplificação do alcance de conteúdos segue na direção oposta e ocorre quando as plataformas decidem "recompensar" certos criadores que consideram mais qualificados (Santini; Salles; Mattos, 2023) – uma determinação, assim como no caso da redução de alcance, que se dá de acordo com seus interesses circunstanciais (Gillespie, 2022). Esse entendimento reforça que a moderação de conteúdo e a recomendação de conteúdo são práticas indissociáveis.

No entanto, a redução e a amplificação do alcance de determinados conteúdos são raramente abordadas oficialmente pelas plataformas, que preferem ignorá-las, minimizá-las ou tratá-las como não intencionais. Afinal, são práticas que ameaçam a ideia de que esses espaços são neutros e promovem a participação coletiva em condições igualitárias (ver Gillespie, 2022; Santini; Salles; Mattos, 2023). O YouTube é uma das poucas plataformas que publicamente alega reduzir o alcance de conteúdo problemático e recompensar criadores de conteúdo "confiáveis", mas não enquadra tais ações como moderação de conteúdo e tampouco é transparente quanto aos critérios que as sustentam (Santini; Salles; Mattos, 2023). O Twitter também possui uma lista interna de "usuários VIP", que são recomendados de forma desproporcional para outros usuários da plataforma (Wiggers, 2023).

Para o público geral, a dimensão comercial da moderação de conteúdo ganhou proeminência com episódios recentes que colocaram em evidência os arranjos econômicos que sustentam essas práticas. Em 2022, o bilionário Elon Musk comprou o Twitter (rebatizado como X) prometendo "salvar a liberdade de expressão", que, segundo ele, teria sido comprometida por intervenções governamentais e pelo avanço de políticas internas favoráveis a usuários de esquerda no espectro político-ideológico (Reuters, 2022; Robertson, 2022). Este movimento acarretou o desmonte das políticas de moderação de conteúdo da plataforma e demissões em massa dos funcionários que compunham suas equipes de *Trust & Safety* em todo o mundo (G1, 2023; O'Brien; Ortutay, 2022). Uma das primeiras consequências dessas decisões veio no fim daquele ano, quando a plataforma comunicou que deixaria de aplicar suas regras de moderação a conteúdos que negassem as dimensões da pandemia de Covid-19 ou a efetividade da vacinação contra a doença. Na mesma época, a plataforma também "anistiou" muitos usuários que haviam sido banidos nos anos anteriores, incluindo o ex-presidente dos EUA Donald Trump (Hart, 2022). Em seguida, as ocorrências de conteúdo violento e discurso de ódio cresceram vertiginosamente na plataforma, com incentivo direto do novo dono, que notoriamente engaja com vozes extremistas e dissemina teorias da

conspiração (Counts; Nakano, 2023). Hoje, a União Europeia reconhece o agora X como a plataforma de rede social com a maior proporção de publicações contendo desinformação (Fitzgerald; Lokmanoglu, 2023).

Por conseguinte, grandes anunciantes passaram a boicotar a plataforma, alegando preocupações com sua reputação. Em 2024, uma pesquisa da Kantar realizada com mais de 1.000 profissionais de marketing sênior nos EUA mostrou que 26% deles planejavam cortar significativamente seus investimentos em publicidade no X, dada sua "imprevisibilidade" (Singh, 2024). Por trás do boicote, estava a *Global Alliance for Responsible Media* (Aliança Global para Mídia Responsável, ou GARM), uma iniciativa formada por mais de 100 empresas em 2019 em resposta à veiculação de anúncios ao lado de conteúdo nocivo em plataformas de redes sociais. Em retaliação, o X decidiu processar a GARM, levando a seu fim (Lee, 2024). Se foi difícil competir com o X/Twitter, a situação com a Meta é ainda mais complexa: conforme indicou O'Reilly (2025), embora muitos anunciantes de alto calibre tenham se sentido incomodados pelas novas políticas das plataformas da empresa e pelas declarações públicas de Mark Zuckerberg no início de 2025, eles dificilmente abandonarão o duopólio publicitário Meta-Google.

Portanto, embora não se deva subestimar a importância da moderação de conteúdo para a experiência final dos usuários, trata-se de um conjunto de práticas imperfeitas, cuja razão de ser impede que ela cumpra efetivamente um papel em prol do interesse público. Independentemente da definição adotada, mais restrita ou mais ampla, com a qual, particularmente, preferimos lidar, seus efeitos são profundos e exigem uma análise crítica cuidadosa. Também subordinada às decisões estratégicas e pressões comerciais das plataformas, encontra-se uma vasta força de trabalho encarregada de aplicar suas diretrizes de moderação, atuando sem qualquer autonomia e não raramente em condições precárias.

3.1.2 O trabalho dos moderadores de conteúdo

Common (2020) divide as atribuições de *Trust & Safety* e os processos de moderação de conteúdo das plataformas de redes sociais em três grandes etapas: (i) o desenvolvimento de regras que as plataformas alegam seguir para governar a conduta de seus usuários; (ii) a aplicação destas regras, que passa pelas denúncias de usuários contra conteúdos potencialmente problemáticos, pela revisão destes conteúdos e pela decisão sobre o que fazer com eles; e (iii) as respostas a estas decisões, que envolvem as contestações feitas por usuários moderados e por organizações da sociedade civil. A etapa de aplicação de regras

talvez seja a mais importante de todo o processo, com o maior impacto sobre a disponibilização de conteúdo nas plataformas (Common, 2020). Como discutiremos em maior detalhe adiante, também trata-se, possivelmente, da fase cujo entendimento é mais comprometido pela falta de transparência que caracteriza essas plataformas.

Muitas são as políticas internas das plataformas que regulam as relações que elas firmam com seus usuários: políticas de privacidade, de direitos autorais, de publicidade, de utilização de dados, para desenvolvedores e, as mais relevantes para a moderação de conteúdo, os termos de serviço e as diretrizes da comunidade ("*community guidelines*") (Myers West, 2018). Em conjunto, estes documentos conferem às plataformas poderes sobre seus usuários mais parecidos com os de governos do que com os de empresas tradicionais (Klonick, 2017). Também conhecidos como termos de uso, os termos de serviço funcionam como uma espécie de "jurisprudência própria" das plataformas e esboçam os termos legais firmados entre elas e seus usuários. Eles são apresentados no momento do registro e fazem referência a precedentes relevantes, tanto nos EUA – onde, em sua maioria, foram formulados –, quanto no restante do mundo (Myers West, 2018).

Já as diretrizes da comunidade descrevem, em linguagem simples e de forma genérica, os tipos de conteúdos e publicações proibidos na plataforma, com exemplos de discurso considerado "sadio" e de discurso considerado "problemático", "tóxico" e/ou "nocivo" (Myers West, 2018). Myers West (2018) argumenta que estas diretrizes não são genéricas ao acaso, pelo contrário. Uma vez que elas devem ser aplicadas a uma base global e heterogênea de usuários, que não compartilham dos mesmos valores e tampouco seguem os mesmos códigos de conduta morais e/ou sociais, é o fato de elas serem genéricas que ajuda as plataformas a lidar com contestações de usuários que discordam de suas ações. Assim, as plataformas de redes sociais redefinem continuamente os limites do que é considerado aceitável dentro de sua lógica de autogovernança, sem qualquer tipo de prestação de contas e guiadas por interesses circunstanciais (Cobbe, 2021).

As plataformas não utilizam somente os termos de serviço e as diretrizes da comunidade para determinar o que deve ser moderado. Normalmente, elas mantêm diretrizes internas, mais detalhadas, que estipulam os limites do "aceitável", mas essas orientações não são divulgadas oficialmente, reiterando a opacidade como princípio estruturante da moderação de conteúdo (ver Gorwa, 2018; Klonick, 2017; Myers West, 2018; Roberts, 2016, 2019). O que se sabe sobre elas é graças a reportagens investigativas, como o trabalho de Angwin e Grassegger (2017), que apontam para as inconsistências em sua redação e aplicação, sobretudo no caso da governança de discurso político.

Tradicionalmente, a aplicação da moderação comercial de conteúdo, com base nessas diretrizes, fica a cargo de trabalhadores terceirizados, que atuam como curadores da informação disponível nas plataformas de redes sociais (Roberts, 2016), subordinados a suas equipes de *Trust & Safety* (Moran *et al.*, 2025), e acontece após as publicações serem realizadas – um trabalho mais *reativo* do que *proativo* (Klonick, 2017). Há exceções, já que algumas poucas plataformas colocam determinados conteúdos "sob espera" antes de serem publicados, para que possam ser revisados por moderadores de conteúdo (Gillespie, 2018a; Klonick, 2017; Roberts, 2016). Em 2016, o Facebook contratou equipes responsáveis por "varrer" a plataforma em busca de atividade terrorista, uma exceção de moderação proativa que se justificou por se tratar de uma questão de segurança nacional (Klonick, 2017). Contudo, por ser impossível submeter todo conteúdo publicado nas grandes plataformas de redes sociais a esta espécie de curadoria editorial, isso significa que praticamente qualquer conteúdo, por "pior" que seja, pode ser publicado, ficando disponível até ser moderado – ou não (Gillespie, 2018a). Ao menos, esse é o paradigma que orienta o funcionamento da internet comercial em países que adotam regimes de imunidade ampla ou de responsabilidade condicional para as plataformas de redes sociais, como é o caso do Brasil, da União Europeia e dos EUA, conforme discutido no Capítulo 2.

Estes moderadores de conteúdo terceirizados compõem uma força de trabalho silenciosa e amplamente invisibilizada (Common, 2020). Geralmente, eles devem assinar acordos de não divulgação e assumir o compromisso de não discutir publicamente o trabalho que eles prestam para as plataformas de redes sociais por diferentes motivos: em primeiro lugar, há a justificativa de que esta é uma forma de proteger os próprios moderadores, já que usuários podem querer se resolver diretamente com eles, no caso de discordarem de uma ação de moderação específica; em segundo, esta seria uma forma de proteger vazamentos de informações pessoais de usuários das plataformas, em um momento em que elas estão sob escrutínio público por conta de violações de privacidade em todo o mundo (Newton, 2019). Estas justificativas oficiais, porém, escondem o fato de que tamanho sigilo existe também para proteger as plataformas e as empresas terceirizadas de possíveis críticas e responsabilizações relacionadas às condições de trabalho dos moderadores de conteúdo (Newton, 2019). Estes trabalhadores terceirizados se concentram em países do Sul Global, a exemplo da Índia e das Filipinas, são mal remunerados, e muitas vezes nem sequer são qualificados para a função que desempenham, graças a treinamentos pouco rigorosos (Common, 2020; Suzor *et al.*, 2019). Portanto, o trabalho dos moderadores de conteúdo é bem

diferente da imagem idealizada dos empregos em grandes empresas de tecnologia, que costumam ser associados a altos salários, estabilidade e benefícios (Gorwa, 2024).

Dia após dia, estes trabalhadores revisam o que há de pior nas plataformas de redes sociais: pornografia infantil, cenas de violência explícita, conteúdo terrorista e, no que parece escapar com mais frequência ao controle das plataformas, desinformação e discurso de ódio – incluindo misoginia, racismo e LGBTfobia (Roberts, 2016). Evidentemente, esse trabalho impõe um custo psicológico significativo que não pode tampouco ser ignorado, com muitos moderadores desenvolvendo traumas decorrentes do trabalho. Common (2020), por exemplo, cita o caso de uma empresa da Flórida que presta serviços terceirizados de moderação de conteúdo para o Facebook, que, sem explicar com antecedência a seus funcionários que o trabalho do qual se encarregariam não era apropriado para indivíduos com histórico de ansiedade e depressão, viu muitos de seus funcionários serem afastados após diagnósticos de transtorno de estresse pós-traumático.

Por meses, Newton (2019) acompanhou de perto a rotina de trabalho dos moderadores de conteúdo em um escritório terceirizado contratado pelo Facebook em Phoenix, nos EUA, e constatou que muitos funcionários desenvolvem quadros severos de ansiedade já durante o período de treinamento. A rotina destes trabalhadores – que recebem, em média 12% do salário anual de um funcionário não terceirizado da Meta – é extremamente vigiada e cronometrada, a ponto de idas ao banheiro precisarem ocorrer dentro de um intervalo de tempo preestabelecido (Newton, 2019). Muitos moderadores de conteúdo comparecem entorpecidos ao trabalho para lidarem com tamanha sobrecarga emocional, enquanto outros passam a concordar com publicações que deveriam remover – neste escritório em específico, moderadores passaram a acreditar que o Holocausto jamais havia acontecido e que a Terra seria plana. Por fim, temendo que ameaças de ex-colegas, que disseram que voltariam ao escritório para “se vingar”, se concretizassem, moderadores passaram a levar armas de fogo ao trabalho, aumentando o clima de tensão e paranoia (Newton, 2019). Cenas como essas provavelmente se repetem em todo o mundo, mas, no Sul Global, os escritórios de moderação de conteúdo terceirizada operam sob ainda mais sigilo.

Os moderadores de conteúdo obedecem a uma estrutura extremamente hierarquizada. De acordo com Klonick (2017), por exemplo, os moderadores de conteúdo do Facebook estão divididos em três escalões: os moderadores de terceiro escalão são os representantes mais numerosos desta força de trabalho invisibilizada, ficando responsáveis pela maior parte do trabalho diário de moderação e revisão de conteúdo; os moderadores de segundo escalão gerenciam e fiscalizam o trabalho dos moderadores de terceiro escalão e ficam responsáveis

pela moderação de conteúdo considerada “prioritária” e mais sensível; e os moderadores de primeiro escalão costumam ser advogados e formuladores de políticas públicas que trabalham nas principais sedes das plataformas de redes sociais. Os moderadores de segundo escalão também ficam responsáveis pela revisão de certas amostras de publicações já avaliadas pelos moderadores de terceiro escalão, como uma forma de garantir que o trabalho destes está sendo desempenhado com a qualidade esperada (Klonick, 2017). Múltiplos moderadores de conteúdo de terceiro escalão avaliam os mesmos conteúdos, e, quando estas avaliações divergem, é dever de um moderador de segundo escalão sacramentar o que há de ser feito (Klonick, 2017). Múltiplas avaliações também ajudam a determinar se os moderadores de conteúdo apresentam um alto nível de concordância entre si (Newton, 2019).

Além disso, é esperado que esses moderadores se transformem em verdadeiras “máquinas humanas”: Common (2020) relata que muitas plataformas exigem que seus moderadores avaliem *ao menos* 2.000 imagens por hora, o que dá apenas 33 segundos para cada avaliação. No caso do Facebook, muitas equipes de moderação devem ser capazes de avaliar uma foto a cada dez segundos. Não há tempo para se ponderar sobre o trabalho que está sendo realizado: os moderadores de conteúdo trabalham somente com base em seus reflexos e não têm como fazer avaliações aprofundadas (Common, 2020). Essas janelas de tempo extremamente curtas tornam mais difícil um trabalho naturalmente desafiador, já que a comunicação humana em rede é complexa e frequentemente envolve ironia, sarcasmo, humor e gírias. Isso pode fazer com que conversas comuns entre amigos soem ofensivas para quem está de fora, incluindo moderadores de conteúdo (Cobbe, 2021).

Para que o conteúdo a ser avaliado pelos moderadores de conteúdo seja mais eficientemente filtrado e selecionado, muitas plataformas se apropriam da força de trabalho gratuita de seus usuários, que denunciam publicações que eles julgam ser inadequadas por uma série de motivos (Buni; Chemaly, 2016; Crawford; Gillespie, 2016; Roberts, 2016). Algumas plataformas solicitam que seus usuários justifiquem os motivos pelos quais denunciaram uma determinada publicação, o que também ajuda no processo de triagem: Crawford e Gillespie (2016) afirmam que publicações denunciadas sob alegações de “conteúdo sexual envolvendo menores de idade” são revisadas imediatamente devido a obrigações legais das plataformas, enquanto publicações denunciadas por outros motivos podem demorar um pouco mais a serem avaliadas.

Além de ajudar os moderadores de conteúdo a navegar por um mar de conteúdo aparentemente infinito, a moderação de conteúdo baseada em denúncias de usuários também serve para legitimar situações em que as plataformas são questionadas por “censurar” outros

usuários – afinal, se a própria comunidade não gostou de um determinado conteúdo, sua vontade deve ser atendida (Klonick, 2017). Porém, como coloca Gillespie (2018a), há um grande problema em empoderar uma "comunidade" que não é realmente uma comunidade, mas um conjunto de usuários heterogêneos, muitas vezes anônimos, espalhados pelo mundo, e frequentemente em conflito entre si. Não é raro que usuários motivados por intolerância abusem dos sistemas de denúncias das plataformas para pressioná-las a remover publicações que não apresentam quaisquer comportamentos violentos, tóxicos e/ou nocivos, um fenômeno conhecido como "*mass reporting*" ("denúncias em massa"), uma modalidade de assédio online orquestrado (Meisner, 2023).

Crawford e Gillespie (2016) citam o caso de usuários homossexuais do Facebook que tiveram suas publicações denunciadas de forma coordenada e removidas sob alegações de "material sexual explícito", mesmo que só contivessem demonstrações de afeto entre estes usuários e seus parceiros. Já Badouard e Bellon (2025) mostram como, em 2015, 50% das denúncias de usuários do Twitter derivaram de estratégias coordenadas de seus usuários. Ainda que muitas plataformas afirmem proibir tais práticas, seus sistemas de denúncias não funcionam como ferramentas verdadeiramente democráticas: além de operarem de forma completamente opaca, podem ser explorados para o silenciamento em massa de usuários (ver Gillespie, 2018a). Não à toa, Evgeny Morozov (2012, p. 104) os classifica como ferramentas de "censura colaborativa" ("*crowd-sourced censorship*"). Preocupadas com usuários que atuam de forma coordenada, algumas plataformas passaram a reconhecer o papel dos "*trusted flaggers*" (ou "*super flaggers*"), algo como "denunciantes de confiança" – usuários oficialmente referendados por apresentarem denúncias bem fundamentadas e que, por isso, têm prioridade na moderação (Badouard; Bellon, 2025). Alguns projetos de regulação das plataformas também reconhecem o papel de tais usuários, como será apresentado adiante.

Conforme Crawford e Gillespie (2016) e Common (2020), como as denúncias codificam o descontentamento dos usuários com as plataformas, publicações denunciadas representam uma ameaça em potencial a suas ambições comerciais, financeiras e políticas. Isso significa, em contrapartida, que cada publicação moderada pelas plataformas é necessariamente acompanhada pela "permissão" concedida à circulação de outras publicações similares, que podem não ter sido denunciadas e, então, encaminhadas aos moderadores de conteúdo. O fato de as plataformas alegarem proibir o discurso de ódio não significa que ele deixe de existir automaticamente: elas intervêm apenas em situações que possam ameaçar seus interesses, evidenciando contradições entre discurso e prática (Roberts, 2016).

De fato, os processos de moderação de conteúdo das plataformas de redes sociais são notoriamente marcados pela aplicação deficiente das diretrizes da comunidade instituídas por elas mesmas (Common, 2020). É importante notar que suas diretrizes da comunidade e as diretrizes reservadas aos moderadores de conteúdo são limitadas por refletirem, em grande medida, a demografia de seus quadros de diretores – segundo Gillespie (2018a, p. 12, tradução do autor), "esmagadoramente brancos, esmagadoramente homens, esmagadoramente escolarizados, esmagadoramente liberais ou libertários e [com uma visão de mundo] esmagadoramente tecnológica". Ou, como similarmente coloca Klonick (2017, p. 1621, tradução do autor), "advogados americanos, treinados e acostumados às normas de liberdade de expressão dos EUA e à Primeira Emenda".

Logo, as perspectivas de minorias sociais são escanteadas, um problema que só tende a piorar à medida que a base de usuários destas plataformas se expande e se diversifica, do Vale do Silício ao resto do mundo, onde leis norte-americanas sobre liberdade de expressão nada mais são do que particularidades regionais (Gillespie, 2018a; Klonick, 2017). Klonick (2017) relata o que talvez seja o primeiro grande caso de conflito entre os valores das plataformas de redes sociais norte-americanas com os valores do "resto do mundo": em 2006, o governo tailandês anunciou que bloquearia o acesso do YouTube no país a não ser que ele removesse vídeos que insultavam o rei do país, um crime cuja pena pode chegar a até 15 anos de prisão. O que mais chocou os funcionários da plataforma foi justamente o quanto o rei da Tailândia era cultuado pela população do país e o quanto a possibilidade de bloquear o YouTube com base nisso – uma medida draconiana, segundo leituras ocidentais – recebeu aclamação popular (Klonick, 2017). No fim, os vídeos foram bloqueados apenas na Tailândia e a plataforma pôde continuar operando no país.

A visão de mundo restrita das grandes plataformas de redes sociais resulta em desigualdades globais na distribuição de recursos humanos e financeiros para a moderação de conteúdo. Aproximadamente 75% dos usuários da internet não têm o inglês como língua nativa, mas as grandes plataformas direcionam a maior parte de seus recursos de moderação a países de língua inglesa, enquanto terceirizam a moderação para outras nações (Shahid, 2024). *Whistleblowers* da Meta revelaram que 87% dos recursos destinados ao combate à desinformação e ao discurso de ódio no Facebook são gastos na moderação de publicações em inglês, mesmo que este seja o idioma principal de só 9% de seus usuários (Elswah, 2024). Apesar da diversidade linguística da União Europeia, apenas 8% dos moderadores de conteúdo do X/Twitter e do Pinterest no bloco são proficientes em outros idiomas que não o inglês (Global Witness, 2023).

Além da falta de preparo linguístico, os moderadores de conteúdo são pouco treinados e carecem de repertório cultural e contextual adequado para avaliar publicações de usuários ao redor do mundo (Newton, 2019). De Gregorio e Stremlau (2022), por exemplo, creditam a limpeza étnica ocorrida em Mianmar, em 2016, à incapacidade dos moderadores de conteúdo do Facebook de gerenciar conteúdo publicado em birmanês. Durante anos, integrantes das Forças Armadas do país, escondidos em perfis anônimos, utilizaram a plataforma para orquestrar uma campanha sistemática de desinformação e propaganda contra os ruaingas, um grupo étnico islâmico do país, que eventualmente culminou em milhares de casos de violência sexual e assassinatos e no maior episódio de migração forçada da história recente (Mozur, 2018). Sem moderadores fluentes no idioma, conteúdos que apoiavam a limpeza étnica dos ruaingas passaram despercebidos pelos moderadores de conteúdo terceirizados (De Gregorio; Stremlau, 2022).

Uma investigação da ProPublica, organização norte-americana de jornalismo independente, expôs diversas das assimetrias dos processos de moderação de conteúdo das plataformas: enquanto um parlamentar dos EUA teve liberdade para afirmar que muçulmanos "radicais" deveriam ser assassinados "em nome de tudo que há de mais sagrado", militantes do movimento *Black Lives Matter* (Vidas Negras Importam) foram banidos das mesmas plataformas após dizerem que "todas as pessoas brancas são racistas" (Angwin; Grassegger, 2017). A mesma investigação revelou que moderadores de conteúdo que trabalham para o Facebook são treinados a remover apenas publicações dirigidas contra grupos pertencentes a "categorias protegidas", definidas com base em critérios como raça, gênero e religiosidade. Essas categorias, contudo, são concebidas de maneira enviesada: Newton (2019) destaca que o Facebook considera que publicações que afirmam que "pessoas autistas deveriam ser esterilizadas" não devem ser removidas, ao contrário daquelas que dizem que "homens deveriam ser esterilizados". Isso ocorre porque, segundo as diretrizes da plataforma, deficiências não recebem o mesmo nível de proteção que raça e gênero, independentemente da opinião que seus moderadores de conteúdo possam ter (Newton, 2019).

Como colocado por Díaz e Hecht-Felella (2021), as demandas e reivindicações de minorias sociais correm o risco de ser moderadas de forma excessiva, ao passo que ataques direcionados a essas mesmas minorias frequentemente passam despercebidos e deixam de ser adequadamente tratados. Essas decisões revelam não só a existência de padrões inconsistentes na moderação, mas também a ausência de um compromisso real com a proteção dos direitos dos usuários, especialmente daqueles que já enfrentam silenciamento sistemático em outros contextos sociais. Embora muitas plataformas tenham ajustado suas políticas internas e

diretrizes da comunidade para melhor atender às demandas de minorias sociais específicas, essas mudanças foram motivadas mais por estratégias de relações públicas e pela crescente pressão sobre suas operações comerciais do que por qualquer consideração genuína pelas causas em questão (Common, 2020).

A aplicação inconsistente das diretrizes da comunidade das plataformas também decorre de outros fatores, como a popularidade de um conteúdo denunciado, fator levado em conta por moderadores de conteúdo (Common, 2020). Quanto mais uma determinada publicação circula, maior a probabilidade de os moderadores optarem por deixá-la no ar, uma vez que conteúdos populares se convertem em maiores receitas publicitárias (Common, 2020; Roberts, 2016). Tal qual a remoção de um conteúdo é um ato de curadoria, Roberts (2016) argumenta que o mesmo se aplica quando um moderador de conteúdo analisa uma publicação racista, homofóbica, sexista ou perturbadora e decide não intervir. Assim,

> Imagens e conteúdos racistas, homofóbicos e misóginos são reafirmados como norma, e as estruturas que os sustentam são invisibilizadas, sugerindo que a existência desse conteúdo é apenas algum tipo de ordem natural das coisas e não, por exemplo, algo potencialmente lucrativo (Roberts, 2016, p. 9, tradução do autor).

3.1.3 A automatização da moderação de conteúdo

O emprego da força de trabalho humana na moderação de conteúdo comumente suscita uma questão: *como tornar a moderação de conteúdo escalável*? (Cobbe, 2021; Flew; Martin; Suzor, 2019; Gillespie, 2020, 2018a; Gomez *et al.*, 2024; Gorwa; Binns; Katzenbach, 2020; Kayyali, 2025; Llansó *et al.*, 2020). A moderação de conteúdo das plataformas está mais do que nunca sob escrutínio público, e mantê-la exclusivamente a cargo de moderadores humanos não permite atender, no tempo esperado, às crescentes demandas sociais e políticas que lhe cercam (ver Gillespie, 2020).

A moderação de conteúdo das plataformas tem origem em outras indústrias de mídia, como o rádio, a televisão e editoras de livros, que sempre precisaram "poupar" suas audiências daquilo que julgavam não ser adequado ou apropriado (Gillespie, 2018a). A principal diferença entre essas indústrias e as plataformas de redes sociais é que estas hospedam e supervisionam uma quantidade inédita de conteúdo – que não produzem – e um número recorde de usuários, ampliando significativamente a magnitude do que precisa ser moderado (Gillespie, 2018a). Como Flew, Martin e Suzor (2019) ressaltam, é impossível que as plataformas de redes sociais, da maneira como elas foram pensadas e concebidas, revisem e moderem previamente todo o conteúdo que é publicado nelas – os autores citam o exemplo do

YouTube, onde 400 horas de vídeo são carregadas a cada minuto, todos os dias. Para que as plataformas possam lidar adequadamente com os desafios que lhes são colocados, a moderação deve assumir um caráter mais *industrial* do que *artesanal* (Moran *et al.*, 2025), substituindo o "trabalho vivo" pelo "trabalho morto" (ver Dantas; Canavarro; Barros, 2014) de sistemas automatizados (Gillespie, 2018a).

Fato é que sistemas automatizados de moderação de conteúdo sempre foram utilizados pelas grandes plataformas de redes sociais, mas com base em regras fixas e pré-programadas, sem espaço para nuances (Gomez *et al.*, 2024). Gillespie (2018a) e Llansó *et al*. (2020) destacam que muitas plataformas costumavam bloquear a publicação de conteúdos com termos proibidos e/ou ofensivos, mas de maneira ineficaz, já que usuários facilmente contornavam essas restrições, escrevendo as palavras de forma errada para evitar a detecção (como *3x3mplo*). Essas soluções não demonstram sensibilidade contextual: nomes de partes íntimas, por exemplo, podem ser usados como xingamentos, mas também aparecem em contextos científicos e educacionais. A mera presença desses termos em uma publicação, por conseguinte, não deveria ser suficiente para justificar sua exclusão sumária (Gillespie, 2018a). Além disso, grandes plataformas recorrem a uma ferramenta chamada PhotoDNA, desenvolvida pela Microsoft, que compara cada nova publicação ao material presente em uma base já conhecida de pornografia infantil, para que esse tipo de violação possa ser moderado preventivamente. O problema é que esta solução se limita apenas à identificação de cópias de conteúdos que já circularam em outras ocasiões (Gillespie, 2018a). Plataformas como o YouTube também conseguem detectar a presença de material protegido por direitos autorais na publicação de novos vídeos e bloqueá-los, caso necessário, tendo desenvolvido ferramentas para isso em resposta a ações jurídicas (Gillespie, 2018a; Gorwa; Binns; Katzenbach, 2020).

Com os avanços tecnológicos recentes, as plataformas, por meio de suas equipes de *Trust & Safety*, têm apostado no desenvolvimento de sistemas algorítmicos baseados em inteligência artificial (IA) e aprendizado de máquina para identificar padrões ligados a conteúdos problemáticos e prever se uma publicação deve ser moderada, classificando-a como discurso de ódio ou *spam* (Gillespie, 2020; Gorwa; Binns; Katzenbach, 2020). Essas tecnologias são, então, apresentadas como a única solução viável para lidar com os desafios de escala, velocidade e precisão na moderação de conteúdo (Gillespie, 2020). Assim, estamos em um momento de transição, em que sistemas voltados à *correspondência* entre conteúdos são paulatinamente substituídos por sistemas *preditivos* e *classificatórios* (Llansó *et al.*, 2020). Com os avanços na moderação algorítmica de conteúdo, as denúncias feitas por

usuários tornam-se mais um mecanismo para medir a satisfação do público do que uma ferramenta efetiva de moderação (Badouard; Bellon, 2025).

Do ponto de vista puramente técnico, os ganhos são significativos: esses sistemas, treinados e aperfeiçoados com base em milhões de avaliações prévias da força de trabalho humana dos moderadores de conteúdo, avaliam e categorizam novos conteúdos, sem que seja necessário haver correspondências anteriores nos bancos de dados das plataformas de redes sociais (Gorwa; Binns; Katzenbach, 2020). Esses sistemas também auxiliam na moderação de conteúdo *proativa* das plataformas, sem necessidade de denúncia, já que podem simplesmente detectar que um conteúdo é proibido antes de ele ir ao ar (Cobbe, 2021).

Em 2016, o Facebook foi a primeira plataforma de rede social a reconhecer publicamente a utilização de sistemas algorítmicos preditivos para realizar a triagem daquilo que seus moderadores deveriam analisar (Gillespie, 2018a). Desde então, outros eventos impulsionaram a adoção desses sistemas, especialmente a pandemia de Covid-19, que levou muitas plataformas a depender exclusivamente deles para a moderação de conteúdo (Gillespie, 2020). Com a impossibilidade de manter o trabalho presencial, os moderadores terceirizados não puderam atuar remotamente, diferentemente dos funcionários em tempo integral das plataformas. À mesma época, o enorme desafio da desinformação sobre a pandemia acelerou a adoção massiva desses sistemas (Gomez *et al.*, 2024). Além da óbvia expansão das medidas de moderação de conteúdo, estes sistemas automatizados são vendidos como uma saída "ética" do problema, por aliviar o fardo psicológico e as condições de trabalho degradantes dos moderadores de conteúdo terceirizados (Gillespie, 2020). O emprego de sistemas algorítmicos também acarreta uma substancial economia de recursos, tornando-se uma alternativa praticamente irresistível às plataformas (Bechmann, 2020; Gillespie, 2018a; Papaevangelou; Votta, 2024).

Apesar dos avanços técnicos, muitos aspectos da moderação de conteúdo automatizada têm de ser discutidos criticamente. Se até mesmo humanos enfrentam dificuldades para definir conceitos ambíguos como "discurso de ódio" e "conteúdo tóxico", a situação se torna exponencialmente mais complexa com o uso de sistemas preditivos para classificação de conteúdo, por mais bem treinados que possam ser (Gorwa; Binns; Katzenbach, 2020). Afinal, um sistema algorítmico não aprende história, política e cultura; ele se baseia em padrões e afinidades semânticas (Llansó *et al.*, 2020). Por exemplo, em 2018, o Facebook afirmou que 99,5% das remoções de conteúdos terroristas, 96% das remoções de conteúdo sexual e nudez e 86% das remoções de conteúdo violento foram automatizadas, contra apenas 38% das instâncias de discurso de ódio (Cobbe, 2021). Conforme Gillespie

(2020), referir-se a algo como “discurso de ódio” não se resume a uma distinção entre o que é certo ou errado, mas constitui uma afirmação social e performativa de que aquilo deveria ser tratado como “discurso de ódio” e que nos leva a refletir sobre o que essa categoria deveria englobar, o que naturalmente suscita discordâncias.

De acordo com Gorwa, Binns e Katzenbach (2020), há três grandes problemas com sistemas algorítmicos de moderação de conteúdo. O primeiro é a falta de transparência decisional, pois esses sistemas não podem ser responsabilizados nem responder pelos erros que cometem. Os critérios que orientam esses sistemas, assim como os padrões de treinamento utilizados, são conhecidos apenas pelas plataformas e seus desenvolvedores, protegidos sob alegações de propriedade intelectual. Essa falta de transparência dificulta não somente a identificação e comprovação de vieses, mas também impede a avaliação dos diversos riscos associados, como discriminação, exclusão e falhas no funcionamento dos sistemas. O segundo é a injustiça, visto que frequentemente cometem erros de forma desproporcional contra minorias sociais, como grupos definidos por gênero, raça, religião ou deficiência (ver Gomez *et al.*, 2024). Equivocadamente, as publicações dessas populações são bloqueadas ou removidas em excesso, após terem sido classificadas como ofensivas. Por fim, há o problema da despolitização: a crença em soluções puramente tecnológicas e técnicas esconde o fato de que a moderação de conteúdo é, essencialmente, uma decisão política e de governança, com profundas consequências sociais, deslocando-a para os domínios da matemática e da computação, com foco na busca por sistemas mais “eficientes”.

Nestes domínios, o desempenho de algoritmos preditivos é determinado a partir de métricas como acurácia, mas o que de fato é uma acurácia suficiente na moderação de conteúdo? Gillespie (2018a) questiona exatamente este ponto: quando lidamos com uma questão política tão delicada e tão disputada quanto a moderação de conteúdo, ter um algoritmo capaz de moderar 90% de publicações tóxicas pode não ser suficiente, a depender dos erros cometidos por ele. Se o problema é a escala, até mesmo 10% de conteúdo tóxico não moderado pode causar um impacto significativo, o que passa a impressão justificável de que os sistemas algorítmicos de moderação de conteúdo talvez não funcionem tão bem, mesmo quando seus números indicam o contrário (Lauer, 2021). Mesmo algoritmos com desempenhos gerais similares entre si podem divergir ao avaliar o mesmo conteúdo, posto que trabalham com base em probabilidades, como demonstram Gomez *et al*. (2024). Argumentam os autores, sistemas algorítmicos de moderação de conteúdo

> fazem previsões estatísticas que não seguem um processo claro baseado em regras. Julgar a liberdade de expressão por meio de modelos estatísticos para controlar o exercício de um direito só é aceitável se esses modelos produzirem resultados

> esperados semelhantes e operarem segundo os mesmos critérios explicáveis e baseados em regras, com as devidas salvaguardas processuais (Gomez *et al.*, 2024, p. 2242, tradução do autor).

A grande promessa da matemática e da computação é o desenvolvimento de sistemas de moderação de conteúdo mais "objetivos", livres dos vieses que afetam moderadores humanos (Common, 2020). No entanto, sistemas algorítmicos são sistemas sociotécnicos, inerentemente normativos, e não puramente técnicos (Badouard; Bellon, 2025; Cobbe, 2021). Como todo sistema algorítmico baseado em IA, eles são "desenvolvidos e testados por pessoas, promulgados e mantidos por pessoas, implementados e aperfeiçoados por pessoas" (Gillespie, 2018a, p. 97, tradução do autor) para atingir os objetivos dessas mesmas ou de outras pessoas (Cobbe, 2021), que imprimem suas visões de mundo nos produtos finais, o que desmonta qualquer pretensão de objetividade (Common, 2020). Isso é especialmente evidente quando sistemas algorítmicos de moderação de conteúdo são treinados com as avaliações arbitrárias de moderadores de conteúdo humanos, como também demonstrado por Gomez *et al*. (2024). Experiências passadas e amplamente documentadas com algoritmos voltados à predição de conteúdo considerado tóxico atestam isso. Gorwa, Binns e Katzenbach (2020) citam o caso de um modelo que classificou 63% de textos com o termo "árabes" como tóxicos, contra apenas 3% daqueles com a frase "eu amo o *fuhrer*". Esse contraste evidencia quais aspectos da desordem informacional seus desenvolvedores consideram problemáticos.

Pode-se argumentar que a crescente dependência de sistemas algorítmicos para a moderação de conteúdo, na verdade, agrava vieses e aumenta a ocorrência de julgamentos errôneos dos usuários por parte das plataformas. Mesmo com seus vieses inerentes, o público tende a perceber suas decisões como meramente técnicas e objetivas, sem contestá-las, o que favorece as ambições comerciais das plataformas (Common, 2020). Embora a moderação automatizada seja, em teoria, mais ágil que o trabalho humano, ela ainda não consegue lidar com certos fenômenos na urgência que eles exigem, como os ataques de Christchurch, na Nova Zelândia, mencionados no Capítulo 2, revelam (ver Bromell, 2022; Common, 2020; Flew, 2024; Gillespie, 2020, 2018a; Gorwa, 2019a, 2024; Gorwa; Binns; Katzenbach, 2020; Popiel; Vasudevan, 2024; Suzor; Gillett, 2022).

A um nível estrutural, Jennifer Cobbe (2021) aponta a um risco mais alarmante da emergência dos sistemas automatizados de moderação de conteúdo: a escalada da censura algorítmica, uma nova forma de autoridade corporativa que pode inserir barreiras comerciais ainda maiores sobre as comunicações cotidianas dos usuários das plataformas de redes sociais. Segundo a autora, a censura algorítmica representa uma evolução das práticas do

Capitalismo de Vigilância, e se baseia na capacidade das plataformas de determinar de forma antecipada e ativa os discursos permitidos e proibidos, alinhando-os a seus interesses comerciais. Dessa maneira, toda forma de comunicação, pública ou privada, fica sujeita às operações das plataformas antes mesmo de ocorrer, podendo ser suprimida de antemão. Isso resulta em um exercício de poder extremamente capilarizado sobre a comunicação privada, e que não poderia ser alcançado com o trabalho humano da moderação de conteúdo ou de outros sistemas não algorítmicos (Cobbe, 2021).

A despeito de os alertas sobre esse novo aparato sociotécnico de vigilância algorítmica poderem parecer exagerados, representantes das plataformas de redes sociais o defendem publicamente. Por exemplo, Jack Dorsey, fundador do Twitter, afirmou que sua ideia era permitir que sistemas algorítmicos varressem todo o conteúdo publicado na plataforma, a fim de recomendar os "mais relevantes" e permitir que os usuários denunciassem os "mais problemáticos" (Gillespie, 2020). Sua fala, aliás, reforça a ideia de que a noção de moderação de conteúdo engloba mais do que somente remover publicações e banir usuários.

A essa altura, a utilização de sistemas algorítmicos para a moderação de conteúdo em plataformas de redes sociais, em conjunto com o trabalho humano ou não, é inevitável. Um dos problemas é que essa utilização não pode ser mandatória – como preveem algumas iniciativas regulatórias para garantir maior "agilidade" –, pois os sistemas atualmente existentes não são 100% confiáveis nem eficazes (Llansó *et al.*, 2020). Uma alternativa possível, advogada por Llansó *et al*. (2020) e Gillespie (2020), é a utilização desses sistemas na moderação prévia de violações como *malware*, pornografia infantil e *spam* – o que, pode-se argumentar, dificilmente encontraria grandes resistências.

Autores como Gillespie (2020) levantam uma questão fundamental: mesmo que *possamos* automatizar a moderação de conteúdo por completo, não está claro se *deveríamos*. Medidas extremas, como o banimento de usuários, talvez devessem ser exclusivamente atribuídas a humanos, visto que "penalizar alguém pela violação de regras é uma das maneiras que nós, enquanto comunidades e sociedades, descobrimos, testamos e reafirmamos valores compartilhados" (Gillespie, 2020, p. 3, tradução do autor). No entanto, o desenvolvimento de uma arquitetura de moderação de conteúdo verdadeiramente responsável encontra barreiras significativas. Entre elas, destaca-se a lógica do "crescimento a qualquer custo" que orienta as plataformas, bem como a manutenção de estruturas opacas e protegidas por segredos comerciais, que dificultam a responsabilização diante de falhas evidentes em sua governança. À luz dessas questões, especialmente a intensificação do uso de sistemas algorítmicos na

moderação, torna-se ainda mais urgente debater os limites, riscos e implicações da falta de transparência que caracteriza esses processos.

3.2 Lacunas e desafios na transparência da moderação de conteúdo

3.2.1 A opacidade estratégica da moderação de conteúdo

Embora as *Big Tech* e as grandes plataformas de redes sociais exaltem seu papel central no ecossistema informacional contemporâneo, elas ainda operam como se estivessem apartadas do interesse público, tomando decisões opacas com base em "regras secretas" para moldar e controlar a expressão pública (Gillespie, 2018a; Suzor, 2019). Como destaca Bromell (2022), essas empresas não demonstram interesse em prestar contas ao público ou em construir uma relação verdadeiramente responsável com ele, priorizando, em vez disso, os interesses de seus investidores, acionistas e parceiros comerciais. Por conseguinte, muitas são as contradições sustentadas por essas corporações na tentativa de preservar seus poderes políticos e econômicos, de forma que

> [o ecossistema das plataformas digitais] parece igualitário, mas é hierárquico; é quase inteiramente corporativo, mas aparenta servir ao interesse público; parece neutro e imparcial, mas sua arquitetura carrega um conjunto específico de valores ideológicos; seus efeitos parecem locais, enquanto seu alcance e impacto são globais; parece substituir o 'governo de cima para baixo' a partir do 'empoderamento de baixo para cima' por parte de seus usuários, mas faz isso por meio de uma *estrutura altamente centralizada que permanece opaca para eles* (van Dijck; Poell; De Waal, 2018, p. 12, tradução e grifo do autor).

Um dos principais obstáculos enfrentados por formuladores de políticas públicas e pesquisadores ao tentar estudar as plataformas de redes sociais é a escassez de informações sobre seu funcionamento. A maior parte dos dados substanciais disponíveis provém de vazamentos e investigações independentes, o que torna cada vez mais difícil mensurar seu impacto social (Leone de Castris, 2024). Nesse cenário de opacidade, cresce a urgência de discutir não apenas o grau de transparência adotado pelas plataformas de redes sociais, mas também os parâmetros ideais que deveriam orientar possíveis obrigações de transparência, especialmente no que diz respeito ao controle sobre o que seus usuários dizem ou fazem.

Inicialmente associada ao direito de acesso a informações sobre ações de interesse público conduzidas por governos locais e nacionais, a concepção de transparência foi se ampliando à medida que se reconhecia que informações relevantes ao público também poderiam emergir de organizações e empresas privadas (Leone de Castris, 2024; Urman; Makhortykh, 2023). Ganhou força, então, a ideia de que algum grau de transparência por

parte de atores privados influentes seria necessário para proteger consumidores em mercados autogovernados marcados por assimetrias de informação, como é o caso das plataformas de redes sociais (Leerssen, 2024; Leone de Castris, 2024). Assim, a transparência se consolida como "a prática de fornecer informações internas sobre 'assuntos de interesse público'" (Urman; Makhortykh, 2023, p. 1, tradução do autor). Para Gorwa e Garton Ash (2020), a força dessa noção está justamente em sua flexibilidade e ambiguidade: a transparência possui apelo junto a públicos diversos, mesmo quando sustentam visões contraditórias entre si. Ainda que política e contestada, ela se consolidou como um predicado inegociável, quase de natureza religiosa, da governança corporativa contemporânea, ao prometer tornar organizações privadas mais responsáveis e eficientes por meio do acesso externo a informações internas (Leerssen, 2024).

No âmbito da governança de plataformas, a capacidade de acessar informações sobre suas funcionalidades e diretrizes constitui um pré-requisito para que o público em geral, governos e outras partes interessadas possam avaliar seu funcionamento, desde os modelos de negócios até a forma como os dados dos usuários são processados e suas políticas aplicadas, o que se mostra especialmente relevante diante das consequências sociais de sua atuação (Urman; Makhortykh, 2023). Na moderação de conteúdo, a transparência se manifesta através do fornecimento de informações essenciais para compreender seu funcionamento e permitir a responsabilização das plataformas por eventuais abusos (Suzor *et al*., 2019). Nessa perspectiva, a divulgação de dados não é um fim em si, mas parte de um processo mais amplo de prestação de contas, voltado ao escrutínio público e à responsabilização, quando cabível. Para Suzor *et al*. (2019), essa transparência deve englobar múltiplas dimensões da moderação, incluindo: (i) como e por quem as regras de moderação são criadas; (ii) como potenciais falhas são identificadas pelas plataformas; (iii) como moderadores são treinados para aplicar as diretrizes das plataformas; (iv) quem são esses moderadores e quais são as condições sob as quais eles trabalham; (v) como as plataformas garantem consistência nos processos de moderação de conteúdo; e (vi) como erros são gerenciados.

À luz de Wagner *et al*. (2020), ao discutir a (falta de) transparência das plataformas de redes sociais, é primordial considerar tanto seus alicerces tecnológicos e mediados quanto as dinâmicas das práticas de visibilidade que buscam tornar pessoas, objetos e processos conhecíveis e governáveis. Gorwa (2019a) descreve as plataformas como monopólios de dados que operam como caixas-pretas, cujos riscos não podem ser avaliados satisfatoriamente por pesquisadores e legisladores, e cujo funcionamento passa longe de ser entendido por seus usuários. Em vista disso, o domínio real das plataformas de redes sociais acabou com

quaisquer expectativas que poderiam existir de que elas seriam a força-motriz de uma nova espécie de "transparência radical" (Gorwa; Garton Ash, 2020).

A caixa-preta é uma metáfora comumente utilizada para descrever o conjunto intangível e de difícil compreensão de arranjos sociotécnicos que compõem a governança de plataformas de redes sociais (Roberts, 2018). Por exemplo, o público não tem meios de saber como funcionam os sistemas algorítmicos de recomendação de conteúdo das plataformas e como seus dados são utilizados para refiná-los e torná-los ainda mais poderosos e eficazes (Llansó *et al.*, 2020). No máximo, são fornecidas informações básicas – avisos como "*você está recebendo este anúncio por ter demonstrado interesse prévio em conteúdos sobre viagens*" –, que estrategicamente reduzem as complexas operações técnicas das plataformas a receitas simplificadas e acessíveis, sem, no entanto, revelar os critérios e padrões efetivamente utilizados (Llansó *et al.*, 2020). Essa opacidade contribui para a consolidação da chamada *Sociedade da Caixa-Preta*, termo cunhado por Frank Pasquale (2016), que aponta para a crescente dependência social de sistemas algorítmicos inacessíveis, controlados por grandes plataformas, e que restringem a capacidade da sociedade de influenciar as decisões sobre o que pode ser feito ou dito online.

Como parte dessa caixa-preta, toda a cadeia de decisões da moderação de conteúdo é envolta em opacidade e tratada como um segredo industrial pelas plataformas (Cobbe, 2021; Gillespie, 2018a; Gorwa; Binns; Katzenbach, 2020; Klonick, 2017; Llansó *et al.*, 2020; Roberts, 2019; Suzor *et al.*, 2019). De início, as políticas que realmente regem a atuação humana e algorítmica nos processos de moderação de conteúdo são, em grande parte, inacessíveis ao público (Klonick, 2017). Para Helberger, Pierson e Poell (2017), é problemático que grandes plataformas, guiadas pelo lucro com publicidade, definam o que é um consumo "saudável" e "diversificado" de informações, especialmente quando os critérios que orientam essas decisões permanecem opacos para os próprios usuários. Nesse arranjo, os usuários são submetidos a padrões estabelecidos e aplicados exclusivamente pelas plataformas (Obar, 2020), sem orientação clara sobre como agir e com poucos recursos para contestar decisões que impactam diretamente suas publicações e interações (Klonick, 2017; Myers West, 2018).

A opacidade das plataformas não é fruto do acaso nem apenas da ausência de regulação externa. Trata-se, antes, de uma escolha estratégica e deliberada que reforça os aspectos mais problemáticos de sua autogovernança do ponto de vista do interesse público (Pasquale, 2016). Essa falta de transparência funciona como um mecanismo de proteção de seu modelo de negócios, uma vez que maior abertura poderia revelar com mais clareza as

situações em que o bem-estar dos usuários é sistematicamente negligenciado em favor de decisões controversas. No campo da governança de conteúdo, isso se traduz no receio de que uma transparência "excessiva" exponha práticas de moderação que favorecem determinados tipos de conteúdo em detrimento de outros, ou revele a permissividade diante da circulação de conteúdos nocivos (Kaushal *et al.*, 2024; Roberts 2018; Strowel; De Meyere, 2023). Além disso, essa opacidade se sustenta com relativa facilidade devido à própria intangibilidade da moderação de conteúdo: para a maioria dos usuários, é difícil perceber ou contestar a ausência de algo que, provavelmente, nunca chegaram a ver e que, logo, sequer se torna parte de sua experiência (Roberts, 2018).

Para Roberts (2018), a ofuscação e o sigilo são estratégias combinadas pelas plataformas de redes sociais para construir uma lógica operacional de opacidade nas ações de moderação de conteúdo. Assim como Gillespie (2018a), a autora destaca que essa opacidade decorre da relutância histórica das plataformas em reconhecer publicamente seu envolvimento diário em ações de moderação. Se, para Gillespie (2018a), a moderação de conteúdo é uma *commodity* das plataformas, a opacidade que a envolve também integra essa lógica, ao dificultar a contestação por parte dos usuários e permitir a continuidade da geração de valor sem grandes questionamentos. Manter em segredo medidas impopulares de moderação é essencial para preservar o engajamento e proteger interesses comerciais ligados à publicidade (Roberts, 2018).

Ao longo da última década, o *status quo* da opacidade das plataformas começou a ser questionado, e o silêncio como procedimento padrão para suas ações de moderação passou a ser desconstruído. As plataformas tornaram-se alvo de críticas crescentes, à medida que diferentes partes interessadas – Estados, usuários e outros atores privados – buscavam compreender seus processos de moderação de conteúdo e influenciar seus rumos (Suzor *et al.*, 2019). Graças à pressão pública, por exemplo, as plataformas passaram a publicar suas diretrizes da comunidade como forma de dar maiores satisfações em relação a estes processos (Gorwa; Binns; Katzenbach, 2020). Já em maio de 2017, o Facebook anunciou que empregaria mais 3.000 moderadores de conteúdo à sua força-tarefa global de combate à desinformação, um dos primeiros pronunciamentos de grandes plataformas acerca de suas práticas de governança de conteúdo e dos trabalhadores nela envolvidos (Roberts, 2018).

No contexto da governança corporativa, a "transparência" é invocada sobretudo quando uma organização privada ameaça os interesses e o bem-estar dos cidadãos, sendo, no entanto, mais pregada do que praticada, como aponta Zalnieriute (2021). Conforme as plataformas passaram a se apropriar e vender a ideia de "transparência" para transmitir uma

imagem de responsabilidade social em suas ações de moderação de conteúdo, especialmente em resposta às crises de opinião pública discutidas no Capítulo 2, ficou evidente que alimentavam uma ilusão (Gorwa; Garton Ash, 2020; Leone de Castris, 2024; Wagner *et al*., 2020). Para Zalnieriute (2021), elas recorrem a uma espécie de *transparency washing*[5], promovendo ações que, em vez de oferecerem informações realmente relevantes sobre suas operações, acabam desviando a atenção de observadores externos do que realmente importa e direcionando-a a questões menores. Assim, muitas ações voluntárias de transparência por parte das plataformas de redes sociais podem, na verdade, confundir e desorientar o público geral, pesquisadores e formuladores de políticas públicas, criando a impressão de que não é preciso implementar critérios vinculantes para regulá-las, já que sua autogovernança seria suficiente para mitigar os riscos decorrentes de suas operações comerciais (Zalnieriute, 2021).

Por parte das plataformas, as ações de *transparency washing*, burocráticas e protocolares, visam mais controlar a opinião pública do que promover avanços concretos (ver também Ananny; Crawford, 2016; Kaushal *et al.*, 2024). Um exemplo disso é a própria publicação das diretrizes da comunidade, que, apesar de parecerem um gesto de abertura, omitem as políticas internas que realmente orientam os processos de moderação de conteúdo, sendo redigidos de forma vaga e com poucos exemplos práticos (Suzor *et al*., 2019). Para Zalnieriute (2021), outros casos de *transparency washing* incluem a publicação de relatórios de transparência de moderação de conteúdo por parte das plataformas de redes sociais. Produzidos de forma voluntária e autogovernada, esses documentos[6] foram concebidos como uma resposta às crescentes pressões por visibilidade pública, compilando informações estatísticas sobre ações de moderação de conteúdo (Dwivedi, 2022; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024). Com isso, funcionam como uma resposta "amigável" às demandas por responsabilização, ao mesmo tempo em que ajudam as plataformas a evitar regulações mais robustas e vinculantes (Suzor *et al*., 2019).

O Google foi pioneiro na publicação de relatórios de transparência de moderação de conteúdo, tendo divulgado os primeiros entre as grandes plataformas em 2010, iniciativa seguida pelo Twitter em 2012 (Urman; Makhortykh, 2023). Naquele momento, os relatórios

---

[5] O sufixo "*washing*" é comumente utilizado para descrever estratégias corporativas que visam mascarar práticas predatórias ou irresponsáveis, projetando uma imagem mais responsável do que corresponde à realidade. Um exemplo clássico é o *greenwashing*, em que empresas utilizam campanhas de marketing e ações de relações públicas para se apresentarem como ambientalmente responsáveis, mesmo sem adotar medidas concretas que sustentem esse posicionamento (Freitas Netto *et al*., 2020).

[6] Relatórios de transparência não são publicados pelas plataformas de redes sociais apenas para, supostamente, dar maior visibilidade aos seus processos de moderação de conteúdo, mas também para abordar questões relacionadas à privacidade e a pedidos de acesso a dados de usuários por parte de autoridades públicas, o que foge ao escopo deste trabalho. Para mais, ver Radsch (2022).

de ambas as empresas focavam em pedidos de remoção de conteúdo por parte de entes estatais e governamentais (Urman; Makhortykh, 2023) ou por violações a direitos autorais (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024), reforçando a percepção de que a moderação de conteúdo seria uma imposição externa, e não uma dimensão estrutural de suas operações comerciais. Essa se tornou uma crítica comum aos relatórios de transparência: por focarem excessivamente em pedidos externos de moderação de conteúdo, havia uma subnotificação intencional do volume e da natureza das ações realizadas diariamente por iniciativa das próprias plataformas (Hovyadinov, 2019; Kosta; Brewczyńska, 2020; Leone de Castris, 2024; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024; Urman; Makhortykh, 2023). Isso é particularmente preocupante quando consideramos que a maior parte do conteúdo moderado pelas plataformas de redes sociais se dá em violação a suas políticas e diretrizes internas, como demonstrado empiricamente por Kaushal *et al*. (2024). Foi somente em 2018 que Google, Facebook e Twitter expandiram o escopo de seus relatórios de transparência de moderação de conteúdo e passaram a incluir informações sobre a aplicação de suas diretrizes da comunidade (Leone de Castris, 2024).

Além disso, os relatórios de transparência costumam apresentar apenas dados agregados e pouco granulares – ou seja, informações genéricas, sem o detalhamento necessário – e raramente trazem análises específicas de casos concretos de moderação (Kosta; Brewczyńska, 2020; Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024; Suzor *et al.*, 2019; Urman; Makhortykh, 2023). Gillespie (2018a) indica que os relatórios de transparência também não revelam quantas publicações foram moderadas após denúncias de usuários. Em vista disso, a utilidade dos relatórios passa a ser questionada, uma vez que as estatísticas divulgadas, tal como são apresentadas, pouco contribuem para reduzir a opacidade dos sistemas de moderação de conteúdo, cujos processos, protocolos e procedimentos centrais permanecem ocultos (Gorwa; Garton Ash, 2020).

As plataformas de redes sociais costumam recorrer a preocupações com a privacidade como justificativa para não compartilhar informações e dados sobre suas equipes de *Trust & Safety* e seus processos de moderação de conteúdo com terceiros ou com o público em geral (Dwivedi, 2022; Moran *et al.*, 2025; Suzor *et al.*, 2019; Vergara; Jain; Mehta, 2024). Houve situações em que a transparência "em excesso" das plataformas fez com que o público também se preocupasse com questões de privacidade. Por exemplo, Suzor *et al*. (2019) citam o episódio em que o CloudFlare, plataforma de cibersegurança, revelou as identidades das pessoas que denunciaram o site *The Daily Stormer* pela hospedagem de conteúdo neonazista e discurso de ódio, o que inspirou campanhas de ódio e assédio direcionadas a elas.

Argumenta-se, assim, que os relatórios de transparência oferecem apenas uma visibilidade controlada sobre as ações de moderação de conteúdo das plataformas, posto que, sem mecanismos de auditoria externa, não há como verificar se as informações divulgadas são enquadradas de maneira seletiva para favorecer a imagem dessas empresas (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024; Wagner *et al.*, 2020). Em última análise, a transparência autogovernada e voluntária das ações de moderação de conteúdo das plataformas de redes sociais estará sempre limitada por mecanismos de escolha, direcionamento e interpretação (Wagner *et al*., 2020), acarretando riscos de aplicação insuficiente das regras, ausência de responsabilização comparável à que ocorreria sob supervisão de um órgão regulador independente e sanções inadequadas (Flew; Gillett, 2021).

Diante disso, é necessário discutir criticamente as limitações do ideal de transparência, em qualquer uma de suas formas, quando aplicado à governança de plataformas de redes sociais. A transparência está longe de ser uma solução mágica para os múltiplos problemas de opacidade das plataformas, como a moderação de conteúdo, a publicidade abusiva e a recomendação de conteúdos nocivos. Ainda que as reivindicações por maior abertura sejam cada vez mais frequentes, muitas delas são formuladas de forma vaga e superficial (Ananny; Crawford, 2016; Geng, 2023; Gillespie, 2018a; Obar, 2020; Rieder; Hofmann, 2020). Embora a transparência seja uma ferramenta de governança amplamente atrativa, isso se deve justamente ao fato de que ela pode assumir significados distintos, variando conforme o tempo, o contexto e os atores envolvidos (Geng, 2023).

Para Ananny e Crawford (2016), por exemplo, a transparência constitui um quadro de governança ineficaz quando aplicada às plataformas de redes sociais: com frequência, medidas de "transparência" podem ocultar intencionalmente informações relevantes, soterrando-as em grandes volumes de dados supérfluos. Além disso, essas medidas tendem a criar falsos binarismos entre "sigilo absoluto" e "abertura total", ao mesmo tempo em que enfrentam limitações técnicas e temporais que comprometem sua efetividade (Ananny; Crawford, 2016). Já Rieder e Hofmann (2020), em oposição à transparência, propõem o paradigma da *observabilidade* como alternativa capaz de superar as limitações do que historicamente se entende como transparência. A principal diferença é que a observabilidade assume um caráter mais pragmático do que a transparência: enquanto esta evoca uma qualidade física de certos materiais, a observabilidade foca na prática ativa de observar, evidenciando as perspectivas dos observadores, sua subjetividade e o processo comunicativo de mediação e interpretação, em contraste com a pretensa objetividade e passividade associadas à transparência (Leerssen, 2024; Rieder; Hofmann, 2020).

Neste trabalho, adotamos, no entanto, o mesmo entendimento proposto por Gorwa e Garton Ash (2020): é importante refletir criticamente sobre as limitações de iniciativas de transparência, mas uma transparência minimamente robusta contribui significativamente para a compreensão pública dos desafios impostos pelas plataformas de redes sociais globalmente. Ainda que seja interessante ampliar o entendimento para novas categorias, como a de observabilidade, essa noção parte do mesmo território semântico e conceitual da transparência, sugerindo mediações de determinada natureza – afinal, algo sempre vai precisar *ser feito observável* por terceiros – e, em última instância, limitando-se à proposta de uma transparência sólida e melhor aplicada. Nesse contexto, a transparência pode gerar importantes ganhos sociais e políticos, tornando-se "condição indispensável para o exercício do controle democrático" das plataformas de redes sociais (Leerssen, 2024, p. 5, tradução do autor), o que pode acontecer com a introdução de novos marcos normativos, como será debatido na sequência. Se, para MacCarthy (2020), a transparência na moderação de conteúdo é essencial para que as plataformas conquistem a confiança de usuários, legisladores e reguladores, a sua ausência tende a reforçar preocupações já existentes e a intensificar a pressão por sua implementação.

Uma transparência minimamente sólida da moderação de conteúdo tem início quando plataformas elencam aos usuários que foram moderados as razões para tanto. Os usuários, por exemplo, muitas vezes não sabem como seus conteúdos moderados foram identificados, o que gera desconfiança em relação aos moderadores humanos, aos sistemas algorítmicos das plataformas, a outros usuários que podem tê-los denunciado e até mesmo a autoridades públicas, suspeitas de influenciarem tais decisões (Suzor *et al.*, 2019). Mais do que uma transparência limitada a notificações individuais, é essencial que as plataformas tornem públicos os detalhes da aplicação de suas regras, permitindo a identificação de possíveis vieses e a responsabilização adequada (Common, 2020). Essa transparência é indispensável para identificar quando direitos fundamentais, como a liberdade de expressão ou a privacidade, são direta ou indiretamente violados pelas plataformas (Gorwa; Garton Ash, 2020), o que se mostra especialmente relevante quando consideramos seu uso na comunicação política tanto em contextos democráticos quanto autoritários (Urman; Makhortykh, 2023). Para isso, é necessário que elas ofereçam uma compreensão clara e precisa sobre seu funcionamento e os critérios que orientam a priorização e a distribuição de informações (Strowel; De Meyere, 2023).

Hoje, as principais preocupações sobre transparência envolvem a crescente dependência das plataformas de sistemas algorítmicos opacos para aplicar suas regras e

diretrizes de forma justa e evitar abusos (Geng, 2023; Suzor *et al.*, 2019). Novos critérios de transparência devem ser capazes de esclarecer de forma útil os mecanismos algorítmicos por trás das decisões de moderação de conteúdo das plataformas, tornando-os mais inteligíveis e revelando como as escolhas comerciais e de design influenciam o compartilhamento e a visibilidade de informações, viabilizando a adoção de medidas frente a eventuais excessos (Strowel; De Meyere, 2023).

O *acesso* público à informação é essencial, mas só se puder ser convertido em *agência* (Obar, 2020); a simples disponibilização dispersa de dados não é suficiente para garantir a responsabilização efetiva das plataformas nem para promover o empoderamento individual no contexto do Capitalismo de Vigilância (Leone de Castris, 2024). Se são as próprias plataformas que determinam os limites e direções de sua transparência, é ilusório esperar que escolham, de forma voluntária, expor falhas que possam comprometer sua imagem perante a opinião pública e as autoridades (Bossetta, 2020). É ainda menos produtivo esperar que indivíduos isolados identifiquem essas falhas apenas navegando por bases de dados ou documentos divulgados voluntariamente pelas plataformas (Pasquale, 2016). Para ser efetiva, a transparência exige a divulgação de informações específicas a públicos específicos, conforme a relevância que essas informações têm para eles (Leone de Castris, 2024). Em termos práticos para este trabalho, a divulgação de números e informações vagas sobre ações pontuais de moderação de conteúdo, em documentos voluntários, pouco contribui para uma transparência efetiva e com impacto real.

O caso recente do X/Twitter ilustra o que, para as plataformas, constitui uma transparência aceitável de sua governança. Pouco após comprar e assumir o comando da plataforma, Elon Musk decidiu seguir aquilo que ele nomeou como uma cartilha de "transparência radical" (Robison, 2023). Em resposta a críticas de usuários que se sentiam "silenciados" diante de um suposto predomínio de vozes alinhadas à esquerda, o novo CEO anunciou a abertura parcial do código-fonte do sistema algorítmico de recomendação da plataforma, numa tentativa de apaziguar preocupações em torno da moderação. Dito e feito: essas porções do código da plataforma foram disponibilizadas publicamente em 31 de março de 2023, sem trechos que, de acordo com comunicado oficial, pudessem comprometer a segurança de seus usuários e seus esforços de combate à manipulação da opinião pública (Wiggers, 2023). Segundo o próprio Musk, essa medida permitiria que as pessoas soubessem que não estavam sendo "manipuladas secretamente" e que pudessem sugerir correções no sistema (Wiggers, 2023). No fim das contas, a divulgação do código revela pouco quando as pessoas não sabem utilizá-lo ou interpretá-lo: pontua Leerssen (2024), mesmo que uma

documentação mais detalhada tivesse sido fornecida, os parâmetros dos algoritmos, por si só, não seriam suficientes para revelar os fatores que realmente impulsionam as tendências de distribuição de conteúdo na plataforma tal qual elas de fato ocorrem.

Fica evidente que, embora as plataformas tenham criado seus modelos de transparência de forma cautelosa para evitar maior exposição, a responsabilidade por garantir segurança jurídica e previsibilidade à transparência recai sobre o Estado, em decorrência do vácuo ilusório que elas próprias geraram (Gorwa, 2024; Gorwa; Garton Ash, 2020; Kosta; Brewczyńska, 2020; Rieder; Hofmann, 2020; Suzor *et al.*, 2019; Wagner *et al.*, 2020; Zalnieriute, 2021). De certo modo, a transparência limitada e pouco efetiva das plataformas acabou se revelando uma armadilha, expondo-as a um escrutínio estatal que provavelmente não teria ocorrido caso tivessem adotado uma postura mais proativa, e inaugurando um novo paradigma voltado à formalização dos mecanismos de transparência.

3.2.2 O *Digital Services Act* e a nova governança de plataformas

A persistente aplicação de diretrizes de moderação baseadas em regras pouco claras e transparentes, associada à resistência em fornecer informações quando solicitadas e à falta de cooperação com autoridades públicas, mesmo em situações de extrema urgência, gerou crescentes pressões públicas, enfim colocando as plataformas sob a mira da regulação externa (Roberts, 2018). Obar (2020) destaca que desconstruir o paradigma da autogovernança é o maior desafio no âmbito da governança de plataformas, pois esse modelo está profundamente enraizado em uma suposta "tradição democrática ocidental". A autogovernança do mercado digital é frequentemente apresentada como uma capacidade de resolver problemas sem comprometer princípios democráticos, direitos dos usuários e a inovação tecnológica. Por esse motivo, o debate de regulação das plataformas enfrenta resistências tão intensas, uma vez que qualquer tentativa de intervenção e maior controle costuma ser percebida como uma ameaça a esses valores (Gorwa, 2019a).

Para a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), a transparência é fundamental para a "boa regulação", pois torna os marcos normativos mais seguros, acessíveis e menos vulneráveis a interesses escusos, além de fortalecer a legitimidade das decisões regulatórias (Wagner *et al*., 2020). Critérios mínimos de transparência são essenciais para que tomadores de decisão possam regular práticas que afetam diretamente o exercício de direitos nas plataformas (Gorwa; Binns; Katzenbach, 2020; Kaushal *et al*., 2024; Wagner et al., 2020). A pesquisa acadêmica desempenha um papel multifacetado nesse

processo, podendo impulsionar ações regulatórias por meio da apresentação de evidências concretas, contribuir para o debate público ou, em casos menos graves, oferecer às plataformas subsídios para aprimorar seus mecanismos de autogovernança (Leerssen, 2024).

O grande desafio, hoje, é transformar as demandas por regulação das plataformas em formas de transparência realmente significativas, que garantam a disponibilização de informações de interesse público de maneira útil, acessível e manejável (Obar, 2020). Inclusive, como aponta Zalnieriute (2021), o investimento das plataformas em uma retórica voltada à "transparência" leva o público, com frequência, a acreditar que elas possuem alguma obrigação legal quanto a isso, quando, na verdade, o que se observa é um acúmulo de iniciativas de *transparency washing*. Como destaca Leone de Castris (2024), discussões sobre tais obrigações ganharam força após o caso *Cambridge Analytica*, que reacendeu preocupações generalizadas sobre o impacto social das plataformas e conferiu um novo senso de urgência ao debate sobre a regulação da transparência.

Nesse contexto, as reformas de transparência tornaram-se centrais para a regulação de conteúdo em plataformas de redes sociais (Leerssen, 2024). Como dito anteriormente, o debate sobre a regulação de conteúdo foi inicialmente liderado por países como Alemanha, França e Reino Unido. O primeiro marco global significativo de transparência obrigatória e regulada para as plataformas foi a aprovação do *Netzwerkdurchsetzungsgesetz*, ou *Network Enforcement Act* (Lei de Fiscalização de Redes; NetzDG), na Alemanha, onde entrou em vigor em 2017, alterando determinados artigos da Lei das Telecomunicações alemã, antes de sua substituição em 2024 (HateAid, 2024). O NetzDG foi a resposta institucional do país ao problema da disseminação em larga escala de desinformação política e discurso de ódio em plataformas de redes sociais, diante da inação das próprias empresas e dos acordos voluntários firmados por elas. Na época, a aprovação da lei marcou um importante avanço, especialmente pelas dificuldades enfrentadas por outros países, incluindo a França, para garantir a constitucionalidade de propostas similares (Dwivedi, 2022).

A lei estabeleceu uma forma mais rigorosa de responsabilização das plataformas pelo conteúdo publicado por seus usuários, tornando-se também a primeira legislação a definir como essas empresas deveriam conduzir a moderação de conteúdo (Brega, 2023; Gorwa, 2024). Gorwa (2024) apresenta que, naquele contexto, as plataformas já não mais realizavam uma moderação de conteúdo "imperfeita, mas aceitável", mas sim uma aplicação "inaceitável" de suas próprias regras e diretrizes. O NetzDG, no entanto, não criava novas categorias de conteúdo a ser removido. Em vez de introduzir definições como "desinformação" ou "discurso de ódio", a lei regulamentava a remoção de conteúdos já

tipificados como crimes no Código Penal alemão (Brega, 2023). Materializava, então, o princípio de que "o que é ilegal offline também deve ser ilegal online", uma resposta direta à recorrente estratégia das plataformas de se esquivar da aplicação de leis nacionais, alegando operar sob jurisdições estrangeiras (Gorwa, 2024). Dessa forma, a desordem informacional na Alemanha não deu origem a novos conceitos jurídicos, configurando-se como um conjunto de condutas criminosas já conhecidas, assim como práticas que não tinham jurisdição previamente estabelecida (Brega, 2023).

A lei determinava que as plataformas de redes sociais[7] com mais de dois milhões de usuários cadastrados no país deveriam instituir mecanismos específicos para que seus usuários pudessem denunciar publicações e outros conteúdos que julgassem ilegais. Na sequência, caberia a elas analisar a denúncia e determinar se o conteúdo era ou não ilegal; em caso afirmativo, ele deveria ser removido permanentemente ou ter seu acesso bloqueado para usuários alemães. Nesse sentido, havia uma distinção fundamental entre dois tipos de conteúdo problemático: os *ilegais*, que deveriam ser removidos ou bloqueados no prazo de até sete dias após a denúncia, e os *manifestamente ilegais*, cuja remoção ou bloqueio deveria ocorrer em até 24 horas (Brega, 2023; Gorwa, 2024).

Como era de se esperar, essa prerrogativa gerou controvérsias entre indivíduos, políticos e organizações nacionais e internacionais de diferentes espectros ideológicos, além, é claro, das próprias plataformas, que viam na lei uma suposta ameaça direta à liberdade de expressão. Não ajudou muito que ela tenha inspirado proposições similares em países tidos como autoritários, incluindo Rússia e Venezuela (Brega, 2023). Argumentava-se que as plataformas poderiam se tornar aparatos de censura generalizada, bloqueando o discurso e o conteúdo publicado por seus usuários da forma que bem entendessem para que não se sujeitassem a sanções previstas na lei. Como consequência, para não serem moderados em massa, os usuários se induziriam à autocensura (Dwivedi, 2022; Heldt, 2019; Maaß; Wortelker; Rott, 2024), no que é conhecido como "efeito inibidor" ("*chilling effect*"). No entanto, essa tese é amplamente contestada: para serem lucrativas, as plataformas precisam justamente manter o engajamento dos usuários, o que torna inviável impedir interações ou aplicar uma moderação sistemática em larga escala (Bedi, 2021; Brega, 2023).

Críticas à parte, o que interessa a este trabalho são as obrigações de transparência de moderação de conteúdo estabelecidas pelo NetzDG. A regulação foi a primeira imposição

[7] Conforme Brega (2023, p. 13), o NetzDG define as plataformas de redes como "provedores de serviços de telecomunicações, que, com a intenção de obter lucro, operam plataformas na Internet, as quais são destinadas ao compartilhamento de conteúdo entre usuários ou sua disponibilização ao público", reconhecendo, portanto, seu claro caráter comercial, mas excluindo aplicativos de troca de mensagens, como WhatsApp e Telegram.

vinculante a obrigar as plataformas de redes sociais a publicar, semestralmente, relatórios de transparência de moderação de conteúdo, detalhando como lidavam com denúncias de usuários acerca de conteúdos potencialmente ilegais (Dwivedi, 2022; Vergara; Jain; Mehta, 2024). Gorwa (2024) aponta que, no auge da ebulição pública por conta da desordem informacional na Alemanha, a opinião pública passou a questionar a opacidade dos processos de moderação de conteúdo de grandes plataformas, em especial o Facebook. Naquele momento, não estava claro como se dava a aplicação das regras previstas em suas diretrizes da comunidade, tampouco se a plataforma contava com equipes de *Trust & Safety* capazes de lidar com as dinâmicas culturais e linguísticas do país (Gorwa, 2024).

Em meio a tantas críticas sobre seus potenciais efeitos negativos sobre a liberdade de expressão dos usuários das plataformas de redes sociais, Brega (2023, p. 14) aponta que suas obrigações de transparência foram, possivelmente, o componente mais elogiado da regulação, com apoio "quase unânime". Para Bassan (2025), isso é sinal de que a transparência poderia funcionar como uma ferramenta regulatória neutra em relação ao conteúdo, capaz de evitar possíveis violações à liberdade de expressão. Na prática, entretanto, os relatórios revelaram-se difíceis de comparar e ficaram aquém das expectativas, com informações apresentadas em métricas muito distintas e níveis variados de granularidade, resultando em um "baixo valor informacional", nos termos de Heldt (2019). Os relatórios de transparência do Facebook, por exemplo, registraram um volume muito baixo de denúncias feitas por usuários. Isso ocorreu porque o mecanismo para reportar conteúdos ilegais havia sido implementado de forma inadequada, sem estar claramente acessível. Devido a brechas na própria legislação, ele se encontrava em páginas separadas e não estava diretamente vinculado às publicações (Dwivedi, 2022; Heldt, 2019).

Embora não tenha provocado, por conta própria, mudanças profundas na governança de plataformas de redes sociais na Alemanha – Brega (2023) e Wagner *et al*. (2020) observam que a maior parte das ações de moderação seguiu sendo pautada pelas diretrizes internas das próprias plataformas, que mantinham primazia sobre a legislação nacional em decisões operacionais –, o NetzDG representou um avanço concreto ao apontar o caminho para o fortalecimento de práticas de transparência que, até então, eram predominantemente voluntárias (Vergara; Jain; Mehta, 2024). Em fevereiro de 2024, porém, o pioneiro NetzDG foi efetivamente substituído (HateAid, 2024) não por uma nova legislação nacional, mas por um marco regulatório significativamente mais ambicioso, tanto em escala quanto em escopo, herdeiro direto de suas proposições de transparência: o *Digital Services Act* (DSA).

Após liderar os debates globais sobre a proteção de dados online com a aprovação da já mencionada GDPR, a União Europeia assumiu também a dianteira na regulação de plataformas digitais. Em 2020, a Comissão Europeia, responsável por aplicar as decisões do Parlamento da União Europeia, divulgou dois projetos de regulações complementares: o DMA, citado ao final do capítulo anterior, voltado a questões de concorrência e competitividade no mercado digital; e o DSA, centrado no conteúdo publicado online e no aumento das responsabilidades das plataformas de redes sociais sobre o que permitem circular. Conjuntamente,

> [...] o DSA e o DMA podem ser conceituados como reações a falhas de mercado amplamente reconhecidas, que, no entanto, assumiram uma forma particular na economia de plataformas e se sobrepõem de maneiras inéditas. Além disso, há uma percepção de um interesse público especial na economia de plataformas, na qual as plataformas são vistas como infraestruturas importantes de um novo tipo (Eifert *et al.*, 2021, p. 1025, tradução do autor).

Mais especificamente, o DSA representa uma virada significativa na governança das plataformas de redes sociais, que deixam de contar com uma imunidade excessiva e passam a ser obrigadas a adotar uma postura mais diligente perante as autoridades e o público geral (Eifert *et al.*, 2021)[8]. Diante disso, surge como substituto direto da ECD, em resposta à sua insuficiência para enfrentar os desafios impostos pela vasta utilização das plataformas de redes sociais (Eifert *et al*., 2021; European Commission, 2024; Frosio, 2024; Schwemer, 2022; Strowel; De Meyere, 2023). A nova regulação adota uma abordagem mais ampla sobre serviços digitais – o que justifica seu nome – e foca, de forma mais específica, nas plataformas digitais, definidas como "serviços de hospedagem online que, a pedido de um usuário, armazenam e disponibilizam informações ao público" (European Parliament, 2022, n.p., tradução do autor). Como um quadro aplicado a toda a União Europeia, o DSA também resolve um problema pré-existente: embora a Comissão Europeia reconhecesse o potencial do NetzDG e de outras iniciativas similares, via com preocupação a fragmentação jurisdicional que incentivavam, ainda que não intencionalmente (Gorwa, 2024).

O DSA, ao contrário do que se poderia esperar, mantém o mesmo regime de responsabilização das plataformas digitais previsto na ECD. Em vez de exigir que monitorem ativamente todo o conteúdo publicado para evitar sanções, reforça que as plataformas podem agir de boa-fé para identificar e remover conteúdos ilegais ou problemáticos, tornando-se

[8] Como mostram Eifert *et al*. (2021), o DMA concentra-se nas relações mercadológicas multilaterais das plataformas. Seu objetivo é regular como essas interações devem ocorrer, combatendo práticas desleais e exigindo maior prestação de contas à Comissão Europeia, especialmente em processos de fusões e aquisições. Embora seus efeitos sejam menos perceptíveis para o público em geral, essas medidas são fundamentais para conter o que pode ser caracterizado como abuso de posição dominante, o que se manifesta por meio de condutas monopolistas e anticompetitivas (Eifert *et al*., 2021).

responsáveis por eles apenas quando são identificados e permanecem no ar (Eifert *et al.*, 2021; Frosio, 2024; Schwemer, 2022; Shattock, 2021). O DSA institui, na verdade, um novo regime de *diligência* e *transparência* na regulação de conteúdo, com o objetivo de permitir que autoridades, pesquisadores e a sociedade como um todo compreendam melhor a arquitetura das plataformas digitais e os riscos que elas podem representar ao interesse público (Strowel; De Meyere, 2023).

Muitas das obrigações de transparência propostas pelo DSA, em diferentes áreas, apenas regulamentam medidas voluntárias e autogovernadas que as plataformas já vinham adotando globalmente, como a publicação de relatórios de transparência de moderação de conteúdo (Leerssen, 2024). Não à toa, a palavra "transparência" e suas variações são citadas cerca de 30 vezes na redação final da regulação (European Parliament, 2022). Neste novo regime, as plataformas passam a ser obrigadas a agir – e a demonstrar que estão agindo – em defesa dos direitos fundamentais de seus usuários, entre os quais estão a dignidade humana, a liberdade de expressão, a proteção de dados pessoais e os direitos dos consumidores, principalmente diante dos chamados riscos sistêmicos (Eifert *et al.*, 2021; European Commission, 2024; Frosio, 2024; Strowel; De Meyere, 2023).

A categoria de riscos sistêmicos não é definida de forma precisa na regulação, que indica que esses riscos derivam do design, do uso inadequado e das funcionalidades das plataformas, como sistemas de recomendação de conteúdo e de distribuição de publicidade. No entanto, a norma destaca quatro possíveis tipos de riscos sistêmicos: (i) disseminação de conteúdo ilegal; (ii) impactos no exercício de direitos fundamentais; (iii) ameaças aos processos democráticos; e (iv) riscos relacionados à violência de gênero, à proteção da saúde pública, à proteção de menores e a consequências graves para o bem-estar físico e mental das pessoas (European Parliament, 2022). Ou seja, enquanto o NetzDG enquadrava a desordem informacional em tipos penais já existentes, o DSA busca enfrentar os riscos decorrentes desse cenário, exigindo que as plataformas atuem também contra conteúdos que, embora não sejam necessariamente ilegais, contribuem para a intensificação de riscos sistêmicos, como a desinformação, sem, no entanto, definir com precisão o que entende por esse termo (Strowel; De Meyere, 2023).

As obrigações previstas no DSA são estruturadas de forma hierarquizada, de modo que quanto mais utilizado for o serviço enquadrado, maior será o conjunto de obrigações a que estará sujeito. A hierarquia, em ordem crescente de responsabilidades, é a seguinte: (i) provedores de serviços online intermediários; (ii) provedores de serviços de hospedagem de conteúdo; (iii) plataformas digitais, um subconjunto da categoria anterior; e, por fim, (iv) as

chamadas *very large online platforms* (“plataformas digitais muito grandes”; VLOPs), que são acompanhadas pelos *very large online search engines* (“buscadores digitais muito grandes”; VLOSEs) (European Parliament, 2022; Strowel; De Meyere, 2023).

Considera-se que as VLOPs e os VLOSEs[9] apresentam um alcance significativamente maior e exercem influência desproporcional sobre a forma como os cidadãos se informam e se comunicam online, em comparação com outros serviços digitais. Por isso, assumem mais obrigações e responsabilidades perante o público e as autoridades europeias, dado que os riscos associados à sua atuação tendem a ser amplificados (European Parliament, 2022), tanto que a regulação passou a ser aplicada a esses serviços seis meses antes de sua vigência plena, justamente para que servissem como referência para os demais posteriormente (Ghedin, 2023). Assim, são classificados com base no número de usuários, sendo considerados “muito grandes” quando atingem uma média mensal de 45 milhões de usuários na União Europeia, aproximadamente 10% da população do bloco.

O DSA surge com a promessa de garantir mais justiça, transparência e responsabilidade na segmentação de publicidade e na recomendação e moderação de conteúdo, aspectos centrais para este trabalho. A ideia é assegurar que as ações de moderação respeitem os direitos fundamentais, cabendo aos sistemas jurídicos nacionais definir se conteúdos difamatórios, degradantes, obscenos, invasivos ou prejudiciais podem permanecer no ar ou devem ser removidos (Eifert *et al.*, 2021; Leerssen, 2024). A concepção de moderação apresentada no DSA é relativamente ampla, incluindo não apenas a remoção de publicações e a suspensão de usuários, mas também ações como desmonetização, checagem de fatos e a redução da visibilidade e do alcance de conteúdos específicos, o que representa um avanço claro em relação à autogovernança das plataformas (Leerssen, 2024).

No Quadro 1, destacamos seis artigos do DSA que visam regular a moderação de conteúdo na União Europeia, para além das já mencionadas obrigações de garantir a possibilidade de agir de boa-fé na moderação de conteúdos problemáticos e de não monitorar integralmente o conteúdo publicado por seus usuários, apontadas, respectivamente, em seus artigos 7 (*Voluntary own-initiative investigations and legal compliance*) e 8 (*No general monitoring or active fact-finding obligations*). Para Frosio (2024), essas medidas, especialmente as delineadas nos artigos 16 e 20, ampliam a transparência e fortalecem os usuários ao garantir que suas demandas sejam tratadas de forma rápida e eficaz. Um detalhe

[9] No momento de finalização deste trabalho, havia vinte e três plataformas digitais designadas como *very large online platforms* e outros dois buscadores classificados como *very large online search engines* pela Comissão Europeia. A lista completa está disponível em: https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses. Acesso em: 21 jul. 2025.

importante é que, diferentemente de outras propostas regulatórias que impõem explicitamente a obrigação de moderação automatizada ou estruturam as exigências de forma que essa se torne a única solução viável, o DSA não obriga as plataformas a moderar conteúdo automaticamente, ainda que essa prática venha se consolidando como padrão (Kaushal *et al.*, 2024). Além disso, para evitar que se esquivem de obrigações de moderação, o DSA exige que as plataformas enquadradas designem representantes legais em toda a União Europeia (Dwivedi, 2022).

Quadro 1 – Principais obrigações de moderação de conteúdo previstas no *Digital Services Act*

Identificação	Nome do Artigo	A quem se aplica	O que prevê
Artigo 9	*Orders to act against illegal content*	● VLOPs e VLOSEs ● Plataformas digitais ● Provedores de serviços online intermediários	Autoridades públicas podem, com base nas leis aplicáveis, solicitar a remoção de conteúdos ilegais das plataformas, desde que suas solicitações atendam a uma série de requisitos, como a indicação clara da norma jurídica que fundamenta o pedido, a identificação da autoridade responsável e a inclusão de informações que permitam à plataforma localizar o conteúdo mencionado.
Artigo 16	*Notice and action mechanisms*	● VLOPs e VLOSEs ● Plataformas digitais	Regulamenta um novo sistema de *notice and takedown*, obrigando todas as plataformas digitais a implementarem sistemas específicos para que usuários possam submeter denúncias contra conteúdos potencialmente ilegais, que deverão ser analisadas em tempo hábil.
Artigo 17	*Statement of reasons*	● VLOPs e VLOSEs ● Plataformas digitais	Obriga as plataformas a notificarem todos os usuários que tiverem conteúdos moderados com base em seus termos de serviço e diretrizes da comunidade, exceto nos casos previstos no artigo 9, explicando a medida adotada e sua justificativa.
Artigo 20	*Internal complaint-handling system*	● VLOPs e VLOSEs ● Plataformas digitais	Garante que usuários possam contestar decisões de moderação de conteúdo.
Artigo 22	*Trusted flaggers*	● VLOPs e VLOSEs ● Plataformas digitais	Obriga as plataformas a darem prioridade às denúncias submetidas por *trusted flaggers*, designados por agências reguladoras e autoridades públicas dos Estados-membros da União Europeia, em razão de sua expertise na governança de plataformas.

Artigo 23	*Measures and protection against misuse*	● VLOPs e VLOSEs ● Plataformas digitais	Obriga as plataformas a suspenderem, por um período de tempo substancial e seguindo aviso prévio, o acesso de usuários que repetidamente disseminam conteúdo ilegal.

Fonte: European Parliament (2022).

As obrigações de transparência do DSA se dividem em cinco eixos, como aponta Dwivedi (2022): (i) Termos de serviço; (ii) Publicação de relatórios; (iii) Publicidade e conteúdo pago; (iv) Análise e gestão de riscos; e (v) Acesso a dados. Outros avanços do DSA, como as obrigações de que VLOPs e VLOSEs conduzam análises dos riscos sistêmicos decorrentes de suas funcionalidades e uso, publiquem os resultados em relatórios específicos e implementem medidas para mitigá-los, e promovam ações voltadas à redução das assimetrias informacionais (ver também Eifert *et al.*, 2021), merecem atenção. Contudo, devido ao escopo desta dissertação, apresentamos no Quadro 2 apenas os três artigos que preveem obrigações de publicação de relatórios de transparência de moderação de conteúdo. É possível ver que, quanto mais relevante for o status da plataforma, maiores são as responsabilidades de transparência que ela assume. Como anteriormente mencionado, as regras do DSA nesse aspecto não são inteiramente novas, mas ampliam o alcance, o escopo e a profundidade de práticas voluntárias já existentes (Leerssen, 2024; Llansó *et al.*, 2020).

Quadro 2 – Obrigações de transparência de moderação de conteúdo previstas no *Digital Services Act*

Identificação	Nome do Artigo	A quem se aplica	O que prevê
Artigo 15	*Transparency reporting obligations for providers of intermediary services*	● VLOPs e VLOSEs ● Plataformas digitais ● Provedores de serviços online intermediários	Publicação anual de relatórios de transparência de moderação de conteúdo, contendo, entre outros dados: - o número de solicitações de remoção recebidas de autoridades dos Estados-membros da União Europeia, classificadas por tipo de ilegalidade; - o volume de notificações submetidas por meio dos mecanismos de *notice and takedown*; - informações sobre contestações a decisões de moderação; - dados agregados sobre medidas que afetem a disponibilidade, visibilidade ou acessibilidade de conteúdos fornecidos por usuários, com destaque para o uso de sistemas automatizados nesses processos.

Artigo 24	*Transparency reporting obligations for providers of online platforms*	● VLOPs e VLOSEs ● Plataformas digitais	Publicação semestral de relatórios de transparência de moderação de conteúdo, incluindo: - dados sobre a resolução extrajudicial de disputas entre usuários e plataformas; - o número médio mensal de usuários ativos por Estado-membro da União Europeia; - informações sobre suspensões de contas por uso indevido da plataforma.
Artigo 42	*Transparency reporting obligations*	VLOPs e VLOSEs	Publicação semestral de relatórios de transparência de moderação de conteúdo, incluindo dados e estatísticas sobre: - as equipes de moderação dedicadas a cada idioma falado na União Europeia; - os índices de precisão e as taxas de erro dos sistemas algorítmicos utilizados nesses processos.

Fonte: European Parliament (2022).

A existência dessas obrigações, por mais rigorosas que pareçam ser, não representa, por si só, uma garantia de aprimoramento no nível de transparência da moderação de conteúdo das grandes plataformas de redes sociais, o que ficou particularmente evidente com a experiência do NetzDG na Alemanha, discutida anteriormente (Dwivedi, 2022; Heldt, 2019; Wagner *et al.*, 2020). Esse não é um problema restrito à regulação de plataformas de redes sociais: regimes regulatórios criados para garantir a transparência de corporações privadas comumente enfrentam dificuldades para assegurar que essa transparência seja de fato significativa (Wagner *et al.*, 2020). O DSA é uma experiência nova, da qual ainda não temos real dimensão dos efeitos práticos. Com a publicação dos primeiros relatórios exigidos pela regulação, análises preliminares apontaram que os avanços em transparência e "valor informacional" ficaram muito aquém do esperado (Miller, 2023). Kaushal *et al*. (2024, p. 1125, tradução do autor) observam, inclusive, que "pode-se argumentar que o DSA não acrescenta muito valor ao cenário de moderação já existente", uma vez que as diretrizes e políticas internas das plataformas continuam sendo seus pilares. No próximo capítulo, sistematizamos os avanços concretos – ou sua ausência – introduzidos pelo novo regime de transparência regulada da União Europeia no campo da moderação de conteúdo, em comparação com o modelo de transparência voluntária e autogovernada ao qual grande parte do mundo (ainda) se submete, e frente ao qual o DSA busca se consolidar como alternativa replicável (European Commission, 2024; Helberger; Samuelson, 2024).

4 A (FALTA DE) TRANSPARÊNCIA DA MODERAÇÃO DE CONTEÚDO: ENTRE O VOLUNTARISMO E A OBRIGAÇÃO REGULADA

Neste capítulo, apresentamos uma análise comparativa entre os relatórios de transparência de moderação de conteúdo publicados voluntariamente, em âmbito global, e aqueles divulgados sob exigência regulatória da União Europeia, no contexto do DSA, por quatro plataformas de redes sociais. A **seção 4.1** é dedicada à exposição e construção da abordagem metodológica adotada. Na **subseção 4.1.1**, explicamos os critérios de seleção das plataformas cujos relatórios foram analisados: Facebook, Instagram, YouTube e X/Twitter. Em seguida, a **subseção 4.1.2** descreve os documentos dessas plataformas em maior detalhe, sendo que as **subseções 4.1.2.1**, **4.1.2.2** e **4.1.2.3** têm um olhar individualizado para cada uma delas – com exceção da **4.1.2.1**, que trata conjuntamente dos relatórios do Facebook e do Instagram. Encerrando a explanação da abordagem metodológica, a **subseção 4.1.3** apresenta a construção do quadro analítico original desenvolvido para comparar os documentos selecionados. Esse quadro, disponível no **Apêndice**, é composto por 60 critérios de avaliação fundamentados em recomendações de especialistas em moderação de conteúdo, na literatura acadêmica especializada e nas exigências do próprio DSA. Os critérios buscam analisar a *granularidade* e a *qualidade* das informações dos relatórios de transparência selecionados.

A **seção 4.2** apresenta em detalhe os resultados da análise comparativa, com base na aplicação do quadro analítico desenvolvido. Os 60 critérios de avaliação estão organizados em cinco grandes eixos temáticos, que orientam a estrutura das subseções de resultados: a **subseção 4.2.1** aborda as *Disposições gerais* dos relatórios de transparência; a **4.2.2** trata da transparência da *Moderação de conteúdo por determinação da plataforma*; a **4.2.3**, da transparência das *Denúncias realizadas por usuários*; a **4.2.4**, da transparência da *Restauração de conteúdo e contestações à moderação*; e a **4.2.5**, da transparência de *Demandas de autoridades públicas*. A **subseção 4.2.6**, por fim, traça um panorama geral dos achados, respondendo às questões de pesquisa delineadas no início deste trabalho. Nela, discutimos em que medida o DSA representa um avanço concreto em relação à transparência da moderação de conteúdo das grandes plataformas na União Europeia, em comparação ao modelo voluntário e autogovernado, e apontamos aspectos que ainda permanecem à margem da nova legislação.

4.1 Materiais e métodos

4.1.1 Seleção das plataformas

Selecionamos para análise e comparação os relatórios de transparência de moderação de conteúdo publicados voluntariamente em âmbito global e, sob exigência regulatória, na União Europeia, por quatro grandes plataformas de redes sociais: Facebook, Instagram, YouTube e X/Twitter. Em primeiro lugar, vale ressaltar a ampla utilização dessas, com cerca de 112, 141, 144 e 16 milhões de usuários ativos no Brasil, respectivamente (Kemp, 2025); em todo o mundo, são aproximadamente 3, 2, 2,53 e 0,58 bilhões de usuários ativos (Dixon, 2025). Ao lado do WhatsApp e do TikTok, Facebook, Instagram e YouTube são as plataformas mais utilizadas globalmente. Na União Europeia, devido à sua ampla base de usuários, essas plataformas são classificadas como VLOPs pelo DSA, conforme definição abordada no capítulo anterior, e, por isso, estão sujeitas a todas as obrigações relativas à transparência da moderação de conteúdo previstas na regulação (European Parliament, 2022). Ao comparar documentos voltados à União Europeia com aqueles destinados ao "resto do mundo", a pesquisa evidencia como as plataformas moldam suas práticas de transparência de acordo com o grau de liberdade regulatória a que estão submetidas. Apesar de o número de plataformas analisadas ser limitado, consideramos que ele é suficiente para identificarmos tendências relevantes sobre o tema, oferecendo uma base sólida para futuras investigações.

É importante ressaltar, ainda, que Facebook e Instagram, as duas principais plataformas da Meta, e YouTube, uma das principais do Google, pertencem a duas das empresas mais valiosas do mundo e que são, seguramente, as mais centrais na consolidação do Capitalismo de Vigilância (Poell; Nieborg; van Dijck, 2019; Zuboff, 2020), além de exercerem um papel cada vez mais relevante como forças de lobby político anti-regulação (Popiel, 2018). Embora o X/Twitter não tenha o mesmo alcance global que as plataformas da Meta e do Google, seus relatórios de transparência de moderação de conteúdo foram selecionados neste trabalho por conta de sua relevância acadêmica e política (Leone de Castris, 2024). Como apontam Warnke, Maier e Gilbert (2024), Facebook, YouTube e X/Twitter são as plataformas mais utilizadas pelos usuários como fontes de informação. Por muito tempo, o X/Twitter foi a plataforma digital mais estudada pela comunidade científica, mesmo com uma base de usuários significativamente menor do que a de suas concorrentes, graças à sua relativa transparência e consequente facilidade de extração e manipulação de dados (Zuckerman, 2021). Esse cenário, contudo, tem mudado desde que Elon Musk assumiu

a empresa, transformando-a em um aparato progressivamente mais fechado, opaco e imprevisível (Blakey, 2024). Além das já mencionadas demissões em massa das equipes de *Trust & Safety* da plataforma (O'Brien; Ortutay, 2022), essa mudança se reflete no encerramento dos métodos gratuitos de coleta de dados, há anos utilizados por pesquisadores (Taylor, 2023).

O fato de todas essas plataformas serem sediadas nos EUA contribui para uma maior convergência político-ideológica em relação a temas como regulação, moderação de conteúdo e transparência, refletindo uma visão comum sobre o funcionamento da internet (Leone de Castris, 2024; Popiel, 2018; Stockmann, 2022). Como ressalta Leone de Castris (2024), Meta, Google e X/Twitter são as empresas de plataformas digitais que mais têm se mostrado engajadas em discussões institucionais sobre políticas de transparência. A escolha por essas plataformas não desconsidera a importância de demais concorrentes e prestadoras de serviços similares, mas reflete sua posição de destaque no setor, o que as torna casos-chave para compreender tendências e padrões mais amplos de governança.

As plataformas selecionadas compartilham arquiteturas semelhantes, estruturadas em torno da circulação, visibilidade e recomendação algorítmica de conteúdos gerados pelos usuários, das interações entre eles e da exibição de publicidade em *feeds* personalizados (Bossetta, 2018; Gorwa, 2024), o que favorece nossa análise comparativa. Ainda que serviços de mensageria como o WhatsApp e o Telegram, muito populares no Brasil e em outras partes do mundo, possam ser considerados plataformas de redes sociais (ver Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024), suas arquiteturas de uso substancialmente diferentes, aliadas ao fato de não estarem incluídos nas obrigações de transparência previstas pelo DSA na União Europeia, justificam sua não inclusão nesta análise. Além disso, os mecanismos de moderação de conteúdo adotados por essas plataformas diferem significativamente dos utilizados pelas plataformas selecionadas, o que comprometeria a realização de uma comparação adequada[10].

4.1.2 Levantamento dos relatórios de transparência

No final de fevereiro de 2025, reunimos os relatórios de transparência de moderação de conteúdo mais recentemente disponibilizados pelas plataformas de redes sociais selecionadas, referentes a diferentes períodos de 2024 e aos contextos regulatórios e

[10] Para discussões em torno da moderação de conteúdo em serviços e aplicativos de mensageria como WhatsApp e Telegram, ver Gillespie *et al*. (2020).

geográficos analisados. Os documentos foram obtidos na Central de Transparência da Meta[11] (para Facebook e Instagram), na seção de relatórios de transparência do Google[12] (para o YouTube) e na Central de Transparência do X/Twitter[13].

O Quadro 3 apresenta os relatórios de transparência de moderação de conteúdo publicados globalmente, de forma voluntária, pelas plataformas selecionadas e analisados neste estudo, acompanhados do período coberto por cada documento. Pode-se perceber, de início, que os relatórios de transparência sobre a aplicação das diretrizes da comunidade no Facebook, Instagram e YouTube são publicados separadamente daqueles que tratam da moderação decorrente de solicitações de autoridades públicas. Essa separação dificulta o acesso integrado às informações sobre moderação de conteúdo pelas partes interessadas. Nenhuma justificativa oficial é dada para tal fato. As análises dos relatórios referentes ao Facebook e ao Instagram serão realizadas em conjunto, tendo em vista que seus relatórios são publicados de forma unificada pela Meta, com as mesmas informações disponibilizadas para ambos. De todo modo, as informações são discriminadas por plataforma, de forma suficiente para permitir a análise individual de cada serviço, caso necessário. Como se pode observar, não há padronização entre os documentos, nem mesmo quando se trata de uma mesma plataforma: alguns relatórios são divulgados trimestralmente, enquanto outros são disponibilizados em intervalos semestrais.

Quadro 3 – Relatórios de transparência de moderação de conteúdo publicados de forma voluntária com escopo global selecionados para análise

Plataforma	Tipo de documento	Período do documento
Facebook & Instagram	Relatório de aplicação das diretrizes da comunidade	Out - Dez/2024
	Relatório de restrições de conteúdo com base na legislação local	Jan - Jun/2024
YouTube	Relatório de aplicação das diretrizes da comunidade	Out - Dez/2024
	Relatório de solicitações governamentais de remoção de conteúdo	Jul - Dez/2024
X/Twitter	Relatório de transparência global	Jul - Dez/2024

Fonte: elaboração própria.

O Quadro 4 apresenta os relatórios de transparência de moderação de conteúdo publicados de forma obrigatória pelas plataformas selecionadas na União Europeia, e indica o período coberto por cada documento analisado. Nota-se que, neste caso, todos os relatórios

[11] Disponível em: https://transparency.meta.com/pt-br/. Acesso em: 7 mar. 2025.
[12] Disponível em: https://transparencyreport.google.com/. Acesso em: 7 mar. 2025.
[13] Disponível em: https://transparency.x.com/en. Acesso em: 7 mar. 2025.

seguem uma periodicidade semestral, conforme previsto pelo DSA (European Commission, [*S. d.*]), e que apenas um relatório é publicado por plataforma.

Quadro 4 – Relatórios de transparência de moderação de conteúdo publicados sob exigência do *Digital Services Act* selecionados para análise

Plataforma	Tipo de documento	Período do documento
Facebook	Relatório de transparência publicado sob exigência do *Digital Services Act* (DSA)	Abr - Set/2024
Instagram	Relatório de transparência publicado sob exigência do *Digital Services Act* (DSA)	Abr - Set/2024
YouTube	Relatório de transparência publicado sob exigência do *Digital Services Act* (DSA)	Jul - Dez/2024
X/Twitter	Relatório de transparência publicado sob exigência do *Digital Services Act* (DSA)	Abr - Set/2024

Fonte: elaboração própria.

A seguir, oferecemos uma visão geral dos relatórios selecionados, destacando suas formas de acesso e disponibilização, bem como suas principais características, antes de detalharmos o método adotado para a análise comparativa dos mesmos.

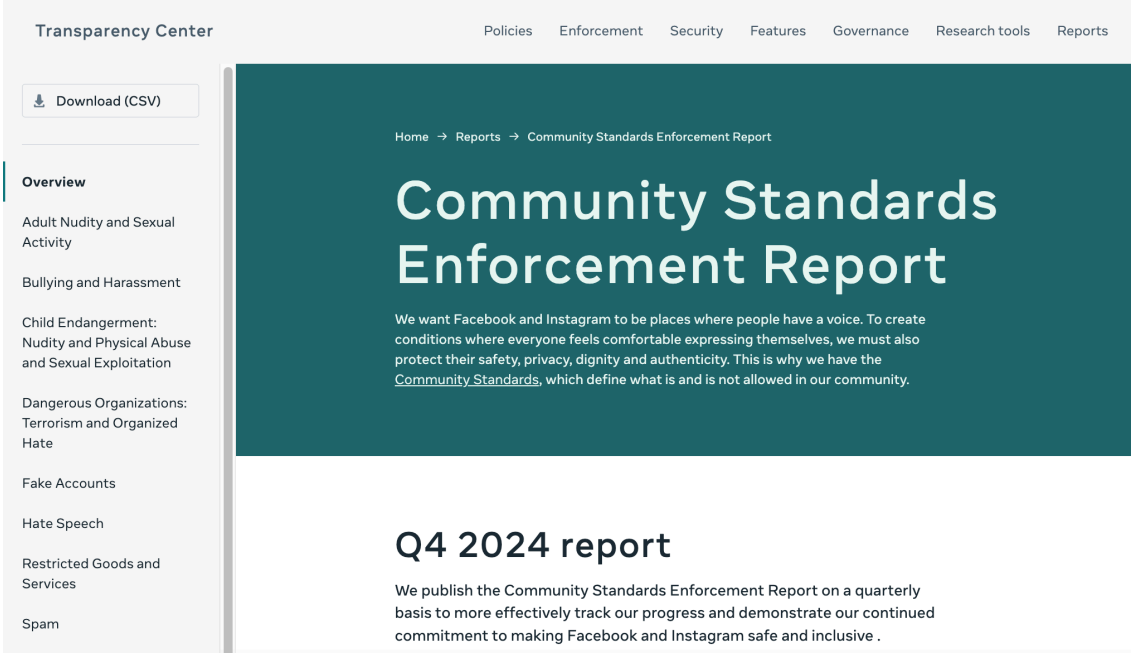
4.1.2.1 Facebook & Instagram

Ao acessar a Central de Transparência da Meta, o usuário se depara com uma lista de relatórios sobre diferentes temas, nem todos relacionados à moderação de conteúdo. Os dois relatórios de transparência voluntários aqui selecionados para análise redirecionam os usuários para outras páginas web, nas quais eles encontram seu conteúdo na íntegra, em uma estrutura muito similar. A Figura 3 apresenta a tela inicial do *relatório de aplicação das diretrizes da comunidade* (*Community Standards Enforcement Report*), com dados sobre as ações de moderação de conteúdo que ocorrem tanto no Facebook quanto no Instagram com base em violações julgadas pelas próprias plataformas.

O relatório tem estrutura relativamente simples: após uma breve introdução sobre seus objetivos e escopo, direciona o usuário à navegação pelas informações de moderação de conteúdo de ambos Facebook e Instagram, organizadas por plataforma e classificadas segundo os tipos de violação às diretrizes da comunidade da Meta. A lista completa de violações apresentadas inclui: (i) *Nudez adulta e atividades sexuais*; (ii) *Bullying e assédio*;

(iii) *Exploração sexual, abuso ou nudez infantil*; (iv) *Organizações e indivíduos perigosos: terrorismo e ódio organizado*; (v) *Contas falsas*; (vi) *Discurso de ódio*; (vii) *Produtos e serviços restritos*; (viii) *Spam*; (ix) *Suicídio, automutilação e distúrbios alimentares*; (x) *Violência e incitação*; e (xi) *Conteúdo violento e explícito*. Cada aba dedicada a um tipo de violação traz estatísticas sobre a aplicação das regras e detalhes qualitativos sobre como ela se dá. Como visto no canto superior esquerdo da Figura 3, também é possível exportar os dados apresentados no relatório em formato CSV (*comma-separated values*), o que permite seu tratamento e análise em ferramentas externas.

Figura 3 – Tela de apresentação do *relatório de aplicação das diretrizes da comunidade* das plataformas da Meta

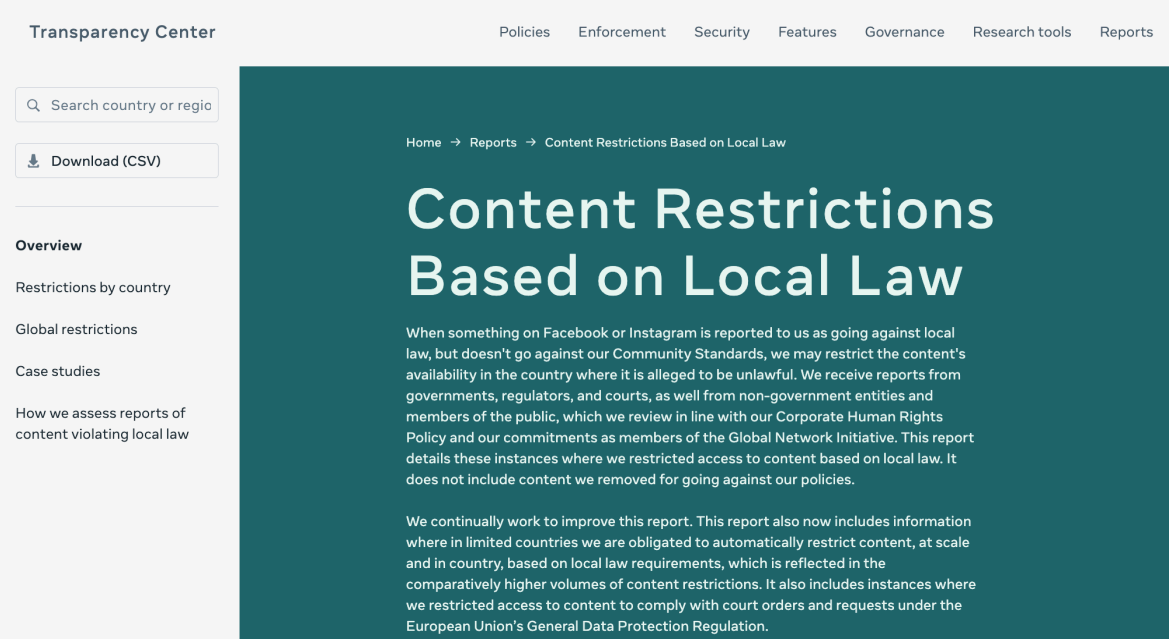


Fonte: captura de tela do autor.

O *relatório de restrições de conteúdo com base na legislação local* (*Content Restrictions Based on Local Law Report*), cuja página pode ser vista na Figura 4, apresenta dados agregados sobre o bloqueio geográfico (*geoblocking*) de conteúdos em países onde as plataformas da Meta operam, em decorrência de solicitações de autoridades públicas – como tribunais superiores e forças de segurança. Segundo a Meta ([*S. d.*]), restrições geográficas ocorrem quando determinado conteúdo é considerado ilegal conforme a legislação local, mas não viola as diretrizes da comunidade ou outras políticas internas da empresa. A página em que se localiza o relatório de transparência permite que os dados apresentados sejam filtrados

por países específicos, incluindo o Brasil. Como mostrado no canto superior esquerdo da Figura 4, os dados também podem ser exportados em formato CSV, permitindo sua análise em ferramentas externas.

Figura 4 – Tela de apresentação do *relatório de restrições de conteúdo com base na legislação local* das plataformas da Meta

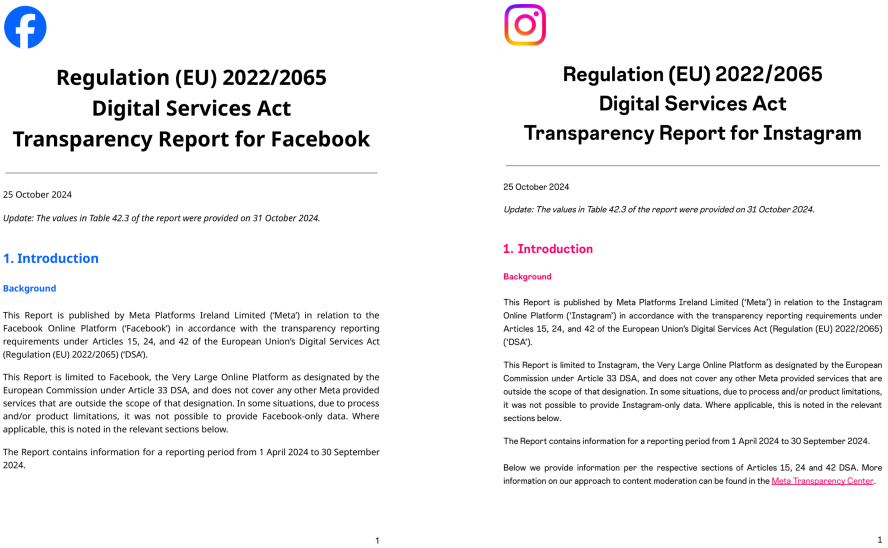


Fonte: captura de tela do autor.

Os relatórios de transparência exigidos pelo DSA são disponibilizados em uma página específica, voltada à publicação de documentos obrigatórios em atendimento a legislações e regulações vigentes em determinadas regiões. Essa página reúne não apenas relatórios sobre moderação de conteúdo, mas também documentos relacionados a técnicas de perfilamento de dados de usuários, deveres de diligência, auditorias diversas, entre outros. Diferentemente dos relatórios voluntários, esses relatórios são separados por plataforma, com documentos distintos para o Facebook e o Instagram. Os relatórios do DSA são disponibilizados exclusivamente em formato PDF (*Portable Document Format*) e seguem uma estrutura padronizada, organizada em seções e categorias que refletem diretamente as exigências de transparência previstas na regulação: *pedidos de moderação realizados por autoridades públicas*; *denúncias de usuários*; *moderação de conteúdo por determinação da própria plataforma*; *contestações à moderação de conteúdo*; *moderação de conteúdo por meios automatizados*; *recursos humanos dedicados à moderação de conteúdo*; *disputas extrajudiciais em torno da moderação de conteúdo*; *proteção contra a má utilização da*

plataforma por parte de usuários; e *número médio de usuários ativos mensalmente em cada país da União Europeia*. As capas dos relatórios de transparência do Facebook (28 páginas) e do Instagram (26 páginas) podem ser vistas na Figura 5.

Figura 5 – Capas dos relatórios de transparência de moderação de conteúdo publicados pela Meta para Facebook (esquerda) e Instagram (direita) na União Europeia em virtude das obrigações do *Digital Services Act*



Regulation (EU) 2022/2065
Digital Services Act
Transparency Report for Facebook

25 October 2024

Update: The values in Table 42.3 of the report were provided on 31 October 2024.

1. Introduction

Background

This Report is published by Meta Platforms Ireland Limited ('Meta') in relation to the Facebook Online Platform ('Facebook') in accordance with the transparency reporting requirements under Articles 15, 24, and 42 of the European Union's Digital Services Act (Regulation (EU) 2022/2065) ('DSA').

This Report is limited to Facebook, the Very Large Online Platform as designated by the European Commission under Article 33 DSA, and does not cover any other Meta provided services that are outside the scope of that designation. In some situations, due to process and/or product limitations, it was not possible to provide Facebook-only data. Where applicable, this is noted in the relevant sections below.

The Report contains information for a reporting period from 1 April 2024 to 30 September 2024.

1

Regulation (EU) 2022/2065
Digital Services Act
Transparency Report for Instagram

25 October 2024

Update: The values in Table 42.3 of the report were provided on 31 October 2024.

1. Introduction

Background

This Report is published by Meta Platforms Ireland Limited ('Meta') in relation to the Instagram Online Platform ('Instagram') in accordance with the transparency reporting requirements under Articles 15, 24, and 42 of the European Union's Digital Services Act (Regulation (EU) 2022/2065) ('DSA').

This Report is limited to Instagram, the Very Large Online Platform as designated by the European Commission under Article 33 DSA, and does not cover any other Meta provided services that are outside the scope of that designation. In some situations, due to process and/or product limitations, it was not possible to provide Instagram-only data. Where applicable, this is noted in the relevant sections below.

The Report contains information for a reporting period from 1 April 2024 to 30 September 2024.

Below we provide information per the respective sections of Articles 15, 24 and 42 DSA. More information on our approach to content moderation can be found in the Meta Transparency Center.

1

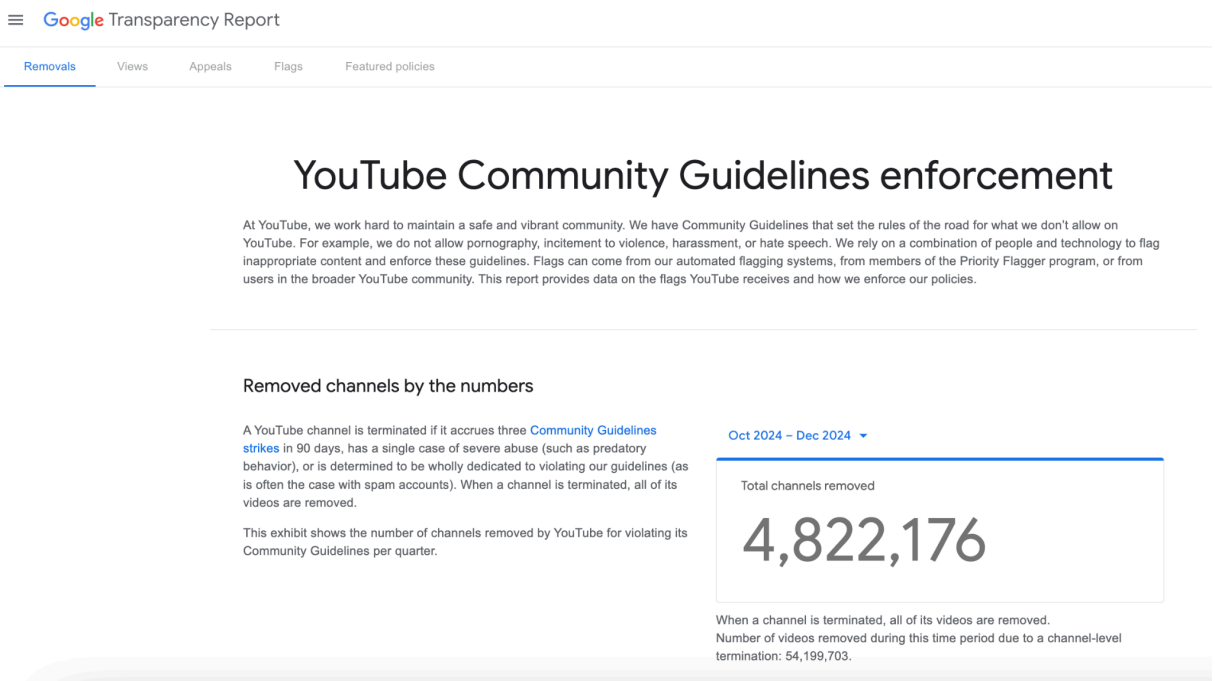
Fonte: Meta (2024a, b).

4.1.2.2 YouTube

O *relatório de aplicação das diretrizes da comunidade* do YouTube (*Community Guidelines Enforcement Report*) é disponibilizado em formato de página web, como ilustrado na Figura 6. Esse relatório de transparência é organizado em cinco seções principais: (i) *Remoções* (*Removals*), que apresenta estatísticas gerais e informações qualitativas sobre a exclusão de vídeos, canais e comentários, muitas vezes discriminadas por tipo de violação; (ii) *Visualizações* (*Views*), que traz o cálculo da taxa de visualizações com violações (*Violative View Rate*), um indicador percentual que estima o alcance global de vídeos posteriormente moderados pela plataforma; (iii) *Contestações* (*Appeals*), com dados estatísticos sobre os recursos interpostos por usuários que tiveram conteúdo moderado e informações sobre a apreciação destas contestações; (iv) *Denúncias* (*Flags*), que fornece informações sobre as sinalizações feitas por usuários a respeito de vídeos potencialmente problemáticos, bem como

explicações sobre o fluxo dessas denúncias, desde sua realização até a decisão final; e (v) *Políticas em destaque* (*Featured Policies*), com estatísticas e exemplos ilustrativos de ações de moderação decorrentes de violações específicas – extremismo violento, segurança de crianças e adolescentes e discurso de ódio.

Figura 6 – Tela de apresentação do *relatório de aplicação das diretrizes da comunidade* do YouTube



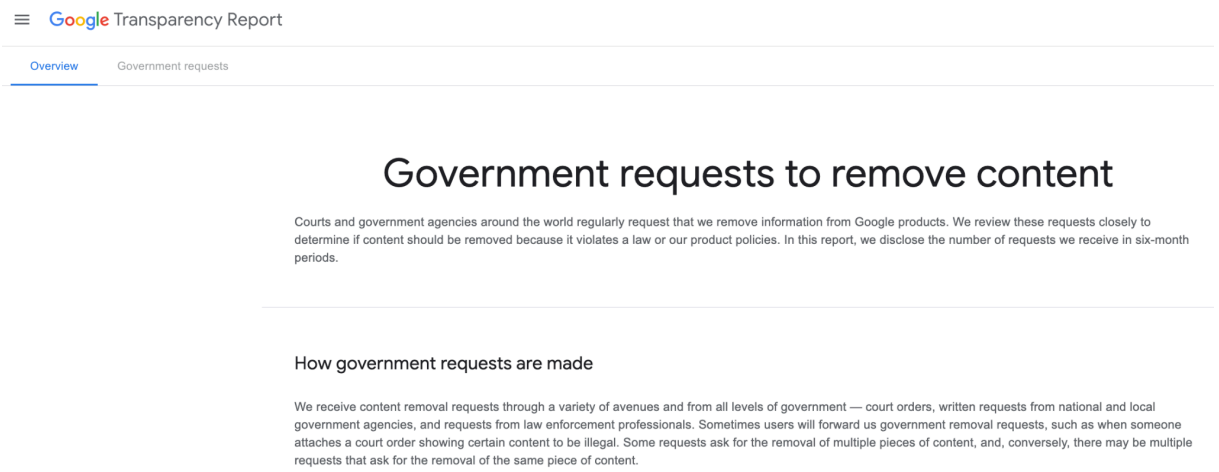
Fonte: captura de tela do autor.

O *relatório de solicitações governamentais de remoção de conteúdo* (*Government Requests to Remove Content Report*), ilustrado na Figura 7, apresenta informações sobre ações de moderação realizadas com base em solicitações de entes governamentais e estatais. Este relatório não é exclusivo ao YouTube, abrangendo também outras plataformas da Google, como Gmail, Google Maps e o próprio buscador da empresa. Como é possível filtrar, selecionar e visualizar especificamente as informações da plataforma ou serviço de interesse, essa organização da página não compromete a validade da análise restrita ao YouTube, preservando a coerência metodológica em relação às demais plataformas selecionadas.

O relatório começa apresentando informações gerais sobre o recebimento e a avaliação de pedidos de moderação de conteúdo por parte de autoridades públicas. Assim como a Meta, o Google afirma que, quando um conteúdo é considerado ilegal em uma determinada jurisdição, mas não viola suas políticas internas ou diretrizes da comunidade, a restrição é aplicada apenas naquela região. Neste caso, no entanto, o relatório especifica claramente

quando as ações de moderação resultam de cada uma dessas motivações. Ações de moderação motivadas por decisões judiciais também são incluídas no relatório. Em seguida, o relatório disponibiliza dados sobre esse tipo de moderação em diversas plataformas da empresa, incluindo o YouTube, com a possibilidade de visualização agregada globalmente ou segmentada por país.

Figura 7 – Tela de apresentação do *relatório de solicitações governamentais de remoção de conteúdo* do Google



Fonte: captura de tela do autor.

O relatório de transparência de moderação de conteúdo publicado na União Europeia, em conformidade com as obrigações regulatórias do DSA, cuja capa pode ser vista na Figura 8, está disponível para *download* em formato PDF (49 páginas) em uma página específica à parte, da mesma forma que ocorre com a Meta. No entanto, ao contrário do que faz a Meta, este relatório abrange informações sobre a moderação de conteúdo em todos os serviços do Google classificados pela legislação como VLOPs e, no caso de seu buscador, como VLOSEs. O documento também é organizado em seções temáticas, que refletem as exigências de transparência previstas pela regulação europeia: *pedidos de moderação realizados por autoridades públicas*; *denúncias de usuários*; *moderação de conteúdo por determinação da própria plataforma*; *contestações à moderação de conteúdo*; *disputas extrajudiciais em torno da moderação de conteúdo*; e *proteção contra a má utilização da plataforma por parte de usuários*. Essas seções são, por sua vez, subdivididas de acordo com as plataformas e ferramentas da empresa abrangidas pelo relatório. Também há subseções específicas destinadas a abordar pontos como a força de trabalho humana da moderação de conteúdo e a utilização de sistemas automatizados para moderação.

Figura 8 – Capa do relatório de transparência de moderação de conteúdo publicado pelo Google para suas *very large online platforms* e *very large online search engines* em virtude das obrigações do *Digital Services Act*

Fonte: Google (2025).

4.1.2.3 X/Twitter

O X/Twitter retomou a publicação de seus relatórios voluntários de transparência de moderação de conteúdo em 2024, após uma interrupção em 2021, durante o processo de aquisição da plataforma por Elon Musk (Joseph; Scanlon, 2024). Diferentemente das demais plataformas analisadas neste estudo, que publicam múltiplos relatórios de transparência para abordar diferentes aspectos de suas ações de governança, o X/Twitter disponibiliza apenas o seu *relatório de transparência global* (*Global Transparency Report*), visto na Figura 9. Enquanto os relatórios anteriores da plataforma ainda podem ser consultados em formato PDF, o relatório mais recente é disponibilizado exclusivamente em uma página web.

O relatório inicia com a apresentação de dados globais agregados sobre a remoção de publicações orgânicas e o banimento de usuários da plataforma, discriminados por tipo de violação. A lista completa de violações apresentadas inclui: (i) *Exploração sexual de*

menores; (ii) *Propagação de ódio*; (iii) *Comportamento inautêntico e spam*; (iv) *Entidades violentas*; (v) *Conteúdo violento*; (vi) *Automutilação e suicídio*; (vii) *Identidades enganosas e fraudulentas*; (viii) *Nudez não consensual*; (ix) *Informações privadas*; e (x) *Produtos e serviços restritos*. Também é apresentado um panorama geral das denúncias recebidas dos usuários, classificadas de acordo com o tipo de violação identificado, acompanhado de uma descrição do processo de análise dessas denúncias até a decisão final de moderação, semelhante ao que faz o YouTube. O relatório ainda cita uma relevante particularidade dos sistemas de moderação de conteúdo do X/Twitter: o princípio da "liberdade de expressão, mas não liberdade de alcance" ("*Freedom of Speech, not Freedom of Reach*"), segundo o qual conteúdos problemáticos, em sua maioria, têm seu alcance reduzido, ao invés de serem deletados (X Safety, 2023). O documento, entretanto, não apresenta informações sobre a aplicação prática desse princípio nem sobre a redução do alcance dessas publicações. Os dados relativos a pedidos de moderação de conteúdo por autoridades públicas são apresentados ao final do relatório, mas com a limitação de que poucas regiões são representadas, como será discutido em maior detalhe na próxima seção.

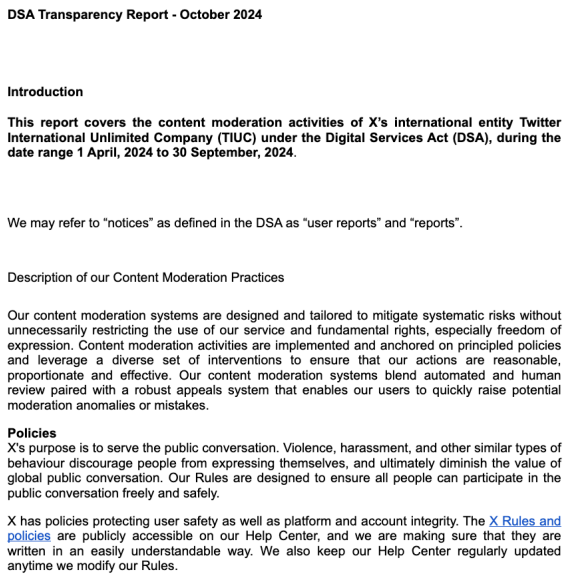
Figura 9 – Tela de apresentação do *relatório de transparência global* do X/Twitter

Fonte: captura de tela do autor.

Por fim, o X/Twitter é a única plataforma aqui analisada que não disponibiliza seu relatório de transparência de moderação de conteúdo publicado diante das obrigações do DSA em formato PDF, fazendo-o por meio de uma página web dedicada. Esta página apresenta uma diagramação mais simples, cuja introdução pode ser visualizada na Figura 10. Assim

como nas demais plataformas, o relatório é dividido em seções temáticas que abordam as diversas informações sobre as ações de moderação de conteúdo que devem ser tornadas transparentes conforme exigido pelo DSA: *recursos humanos dedicados à moderação de conteúdo*; *pedidos de moderação realizados por autoridades públicas*; *denúncias de usuários*; *moderação de conteúdo por determinação da própria plataforma*; *moderação de conteúdo por meios automatizados*; e *número médio de usuários ativos mensalmente em cada país da União Europeia*. Tal como ocorre com todos os outros relatórios de transparência de moderação de conteúdo publicados por exigência do DSA analisados, não é possível extrair os dados apresentados no relatório do X/Twitter em formato estruturado.

Figura 10 – Tela de apresentação do relatório de transparência de moderação de conteúdo publicado pelo X/Twitter em virtude das obrigações do *Digital Services Act*



DSA Transparency Report - October 2024

Introduction

This report covers the content moderation activities of X's international entity Twitter International Unlimited Company (TIUC) under the Digital Services Act (DSA), during the date range 1 April, 2024 to 30 September, 2024.

We may refer to "notices" as defined in the DSA as "user reports" and "reports".

Description of our Content Moderation Practices

Our content moderation systems are designed and tailored to mitigate systematic risks without unnecessarily restricting the use of our service and fundamental rights, especially freedom of expression. Content moderation activities are implemented and anchored on principled policies and leverage a diverse set of interventions to ensure that our actions are reasonable, proportionate and effective. Our content moderation systems blend automated and human review paired with a robust appeals system that enables our users to quickly raise potential moderation anomalies or mistakes.

Policies
X's purpose is to serve the public conversation. Violence, harassment, and other similar types of behaviour discourage people from expressing themselves, and ultimately diminish the value of global public conversation. Our Rules are designed to ensure all people can participate in the public conversation freely and safely.

X has policies protecting user safety as well as platform and account integrity. The X Rules and policies are publicly accessible on our Help Center, and we are making sure that they are written in an easily understandable way. We also keep our Help Center regularly updated anytime we modify our Rules.

Fonte: captura de tela do autor.

4.1.3 Construção do quadro analítico

É importante enfatizar que a literatura acadêmica mais recente sobre governança e regulação de plataformas digitais, especialmente a de caráter empírico, tem se beneficiado amplamente de análises comparativas. Ao longo desta seção, e à medida que apresentamos nossa abordagem, indicamos algumas influências metodológicas relevantes para este trabalho. Essas contribuições são especialmente valiosas porque, embora adotem perspectivas distintas, os estudos mencionados buscam evidenciar falhas na transparência voluntária das plataformas

e na aplicação desigual de marcos regulatórios, revelando disparidades entre regiões já submetidas à regulação e outras onde as plataformas ainda operam sob suas próprias regras.

Entre as referências metodológicas que embasam este trabalho, destacamos primeiramente o estudo de Urman e Makhortykh (2023), que, por sua vez, fundamenta-se na versão mais recente dos *Santa Clara Principles* (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021). Os *Santa Clara Principles* (SCP) são um conjunto de recomendações elaboradas por pesquisadores e organizações da sociedade civil[14] com o objetivo de assegurar o respeito ao devido processo e aos direitos dos usuários na moderação de conteúdo, além de fomentar a produção de relatórios de transparência mais detalhados, consistentes e acessíveis (Radsch, 2022). A primeira versão dos SCP foi publicada em 2018, como resultado das conferências *Content Moderation at Scale* (Moderação de Conteúdo em Larga Escala), que ocorreram na cidade de Santa Clara, na Califórnia. Após a publicação da primeira versão dos SCP, diversas empresas de tecnologia o endossaram publicamente, o que, segundo seus próprios responsáveis, fez com que essas empresas passassem a oferecer mais garantias processuais e transparência ao processo de moderação de conteúdo (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021). Entre 2020 e 2021, a segunda versão dos SCP foi formulada, com o objetivo de aprofundar as recomendações e abordar as desigualdades globais na publicação de relatórios de transparência de grandes plataformas digitais (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021).

Os autores, então, realizaram uma análise comparativa dos relatórios de transparência de moderação de conteúdo voluntários e globais de 13 empresas de plataformas digitais, incluindo as analisadas neste estudo, para avaliar o grau de conformidade desses documentos com as recomendações dos SCP (Urman; Makhortykh, 2023). O estudo revelou uma grande divergência nas formas de disponibilização das informações de transparência de moderação de conteúdo, o que, segundo os autores, contribui para o fenômeno do *transparency washing* – especialmente no que se refere às ações de moderação determinadas pelas próprias plataformas (Urman; Makhortykh, 2023; ver também Zalnieriute, 2021). Até então, o DSA ainda era uma legislação recém-aprovada pela União Europeia, e, embora a obrigação de publicar relatórios de transparência já estivesse prevista, levaria algum tempo até que esses

[14] Reunindo, além dos EUA, entidades de países como Costa Rica, Brasil e Equador, a lista de organizações e empresas que colaboraram para a construção dos SCP inclui: 7amleh, Association for Progressive Communications, Centre for Internet & Society, Meta, Fundación Acceso, GitHub, Institute for Research on Internet and Society, InternetLab, Laboratório de Políticas Públicas e Internet, Lawyers Hub, Montreal AI Ethics Institute, PEN America, Point of View, Public Knowledge, Taiwan Association for Human Rights, The Dialogue e Usuarios Digitales.

documentos se tornassem acessíveis e passíveis de comparação. Com isso, os próprios autores propuseram recomendações para a publicação de relatórios obrigatórios no contexto do DSA, incluindo sua padronização, tradução para diversos idiomas, disponibilização em formatos estruturados para análises externas e a inclusão de exemplos de casos politicamente sensíveis de moderação de conteúdo (Urman; Makhortykh, 2023).

Para este trabalho, desenvolvemos um quadro analítico original, composto por 60 critérios de avaliação, destinado à comparação crítica dos relatórios de transparência de moderação de conteúdo selecionados. O quadro não se propõe a analisar, necessariamente, as informações disponibilizadas em si, mas sim a identificar *quais* informações são oferecidas por cada plataforma em diferentes contextos, considerando dois aspectos principais: o nível de *granularidade* e a *qualidade* dessas informações. A granularidade refere-se ao grau de detalhamento dos dados – uma granularidade baixa indica informações mais gerais e agregadas. Já a qualidade diz respeito à relevância e utilidade dos dados, de modo que informações de qualidade são claras e efetivamente úteis para os objetivos de análise e transparência (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024).

A elaboração dos critérios de avaliação desta pesquisa deriva, em grande parte, da experiência acumulada com o desenvolvimento dos *Índice de Transparência de Dados das Plataformas de Redes Sociais* (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Medeiros; *et al.*, 2024) e *Índice de Transparência da Publicidade nas Plataformas de Redes Sociais* (Santini; Salles; Mattos; Canavarro; Barros; Moreira; Grael; *et al.*, 2024), publicados pelo NetLab UFRJ e dos quais fui coautor. Esses trabalhos resultaram na criação de quadros analíticos compostos, ao todo, por 100 critérios voltados à avaliação do acesso e da qualidade dos dados disponibilizados por grandes plataformas de redes sociais para fins de pesquisa acadêmica – o primeiro com foco em publicações e interações orgânicas; o segundo, em conteúdo publicitário. A participação ativa nesse processo de sistematização do escrutínio das práticas e políticas de transparência e acesso a dados de grandes plataformas de redes sociais foi fundamental para a construção do quadro analítico adotado nesta pesquisa.

Além de abranger, naturalmente, os SCP, que influenciaram diretamente 42 dos seus 60 critérios, nosso quadro analítico incorpora também contribuições de estudos como os de MacCarthy (2020), Radsch (2022) e Santini, Salles, Mattos, Canavarro, Barros, Moreira, Medeiros *et al.* (2024), bem como recomendações propostas por Urman e Makhortykh (2023). As determinações do DSA para a publicação de relatórios de transparência de moderação de conteúdo, já apresentadas no capítulo anterior desta dissertação, inspiram 33 dos 60 critérios desenvolvidos. Note-se que, não raramente, estas inspirações e referências se

sobrepõem no embasamento de um mesmo critério. A inclusão das exigências do DSA em nosso quadro analítico, em particular, permite identificar lacunas de transparência não abordadas pela própria legislação. Com isso, torna-se possível avaliar em que medida o DSA representa um avanço concreto na transparência da moderação de conteúdo na União Europeia, em comparação com outras regiões, além de examinar as limitações dessa transparência agora regulada, em contraste com os padrões voluntários vigentes na maior parte do mundo.

Mesmo nos critérios de avaliação que não se baseiam diretamente nos SCP, adotamos uma de suas diretrizes centrais: a de que as informações sobre ações de moderação de conteúdo, nos relatórios de transparência, devem ser apresentadas de forma discriminada por país ou região, assim como por tipo de violação percebida (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021). Isso porque a divulgação apenas de estatísticas globais e agregadas limita a compreensão das ações de moderação de conteúdo, dificultando que pesquisadores, usuários e demais interessados compreendam o funcionamento e os impactos reais desses sistemas (Wagner *et al.*, 2020).

Além disso, o quadro analítico resultou de um processo iterativo, no qual os relatórios de transparência foram lidos e examinados paralelamente à elaboração e ao refinamento dos critérios de avaliação, possibilitando que ele fosse continuamente aprimorado conforme os padrões e temas emergentes identificados nos documentos selecionados (Forman *et al*., 2008). Com base em Morgan (2022) e Bowen (2009), os documentos foram submetidos a uma leitura atenta e comparada, focada em identificar padrões de consistência e variação na qualidade e nas formas de exposição das informações adotadas nos relatórios de transparência. Esse processo possibilitou incorporar ao nosso quadro analítico práticas identificadas em alguns dos relatórios de transparência analisados, partindo do entendimento de que, idealmente, essas práticas poderiam ser aprofundadas e também adotadas por outras plataformas. As possibilidades de escrutínio dos relatórios de transparência selecionados derivam, em grande parte, do trabalho de Trans *et al.* (2024), que realizaram uma análise documental comparativa das políticas internas de quatro grandes plataformas de redes sociais, com o objetivo de entender como, no contexto regulatório europeu, essas empresas possibilitam (ou não) o acesso a dados para pesquisas de interesse público. Além de conduzirem uma das primeiras investigações sistemáticas sobre as práticas de transparência relacionadas ao DSA, os autores também contribuíram diretamente para os procedimentos desta pesquisa, especialmente no que se refere à forma de análise e organização dos dados para viabilizar sua comparação.

A íntegra do quadro analítico, com o detalhamento do embasamento de cada critério de avaliação desenvolvido para esta dissertação, pode ser consultada no **Apêndice**. Há de se notar uma opção metodológica por repetir, em diversos casos, os mesmos enunciados dos critérios de avaliação para dar conta, separadamente, da transparência da moderação de publicações orgânicas, anúncios e usuários. Essa escolha foi diretamente inspirada nos SCP, permitindo uma avaliação mais precisa das diferentes práticas de moderação de conteúdo adotadas pelas plataformas e, assim, a identificação de lacunas específicas de transparência em cada uma dessas frentes. De acordo com o site oficial dos SCP,

> Esta segunda versão dos SCP amplia o escopo da transparência exigida no que diz respeito ao que é considerado "conteúdo" e às "ações" tomadas por uma plataforma. O termo "conteúdo" refere-se a todo conteúdo gerado por usuários, pago ou não, incluindo publicidade. Já os termos "ação" e "acionado" referem-se a qualquer tipo de medida de aplicação adotada por uma empresa em relação ao conteúdo ou a uma conta, devido ao descumprimento de suas regras e políticas, incluindo, mas não se limitando a remoção de conteúdo, rebaixamento algorítmico e suspensão, temporária ou permanente, de contas (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021, n.p., tradução do autor).

Nós organizamos os critérios de avaliação em cinco grandes eixos: (i) *Disposições gerais*, cujos dez critérios avaliam aspectos contextuais e estruturais dos relatórios, tais como sua frequência de publicação e acessibilidade dos dados; (ii) *Moderação de conteúdo por determinação da plataforma*, cujos quatorze critérios tratam da divulgação de informações sobre ações de moderação de conteúdo determinadas exclusivamente pelas plataformas, sem interferências externas; (iii) *Denúncias realizadas por usuários*, cujos quinze critérios se referem à divulgação de informações sobre denúncias feitas por usuários, que usualmente fundamentam ações de moderação; (iv) *Restauração de conteúdo e contestações à moderação*, cujos seis critérios abrangem a divulgação de informações sobre retificações proativas ou reativas de decisões equivocadas de moderação; e (v) *Demandas de autoridades públicas*, cujos quinze critérios contemplam a divulgação de informações sobre ações de moderação motivadas por solicitações ou ordens de órgãos estatais ou governamentais. Além de poderem ser consultados no **Apêndice**, os enunciados de todos os critérios que compõem cada um destes eixos serão apresentados na próxima seção deste capítulo.

Os critérios de avaliação podiam ser respondidos de três maneiras distintas, seguindo uma escala gradativa: (i) *Atende positivamente ao enunciado*, indicando cumprimento total de todos os requisitos do enunciado do critério, em todos os níveis de detalhamento e com granularidade esperada; (ii) *Atende parcialmente ao enunciado*, indicando cumprimento limitado dos requisitos do enunciado do critério, sem o detalhamento e a granularidade esperados; e (iii) *Não atende ao enunciado*, que indica o não cumprimento de nenhum dos

requisitos do enunciado, em qualquer nível de detalhamento e granularidade. Essa escala foi inspirada no trabalho de Hawker *et al*. (2022), que desenvolveram um quadro analítico próprio para comparar as políticas de transparência publicitária de cinco grandes plataformas, com critérios avaliados também por meio de uma escala tripartida.

Na próxima seção, apresentamos os resultados de nossa análise. Iniciamos com uma visão comparativa dos relatórios publicados pelas plataformas selecionadas, com base nos eixos que estruturam os critérios de avaliação, a fim de evidenciar o desempenho relativo de cada uma. Em seguida, oferecemos uma visão geral e uma discussão final dos achados, com o objetivo de responder às perguntas de pesquisa delineadas na introdução deste trabalho e, principalmente, analisar os avanços e retrocessos promovidos pelo DSA em relação à transparência da moderação de conteúdo.

4.2 Resultados e discussão

4.2.1 Disposições gerais

De modo geral, os relatórios publicados em conformidade com as exigências do DSA apresentam desempenho superior nos critérios de nosso primeiro eixo de avaliação, quando comparados aos relatórios de transparência voluntários, conforme ilustrado na Figura 11. Os relatórios de transparência de moderação de conteúdo publicados voluntariamente por Facebook e Instagram contabilizam apenas duas avaliações parciais e duas positivas. Já os documentos publicados pelas mesmas plataformas no contexto europeu apresentam desempenho significativamente melhor, com seis avaliações positivas e duas parciais. O YouTube também demonstra avanços, passando de três avaliações positivas e uma parcial para seis positivas e uma parcial. O destaque, no entanto, é o X/Twitter. Enquanto seu relatório voluntário de transparência teve apenas uma avaliação positiva, o documento publicado na União Europeia obteve sete avaliações positivas e uma parcial. O detalhamento destas avaliações pode ser conferido na Figura 12.

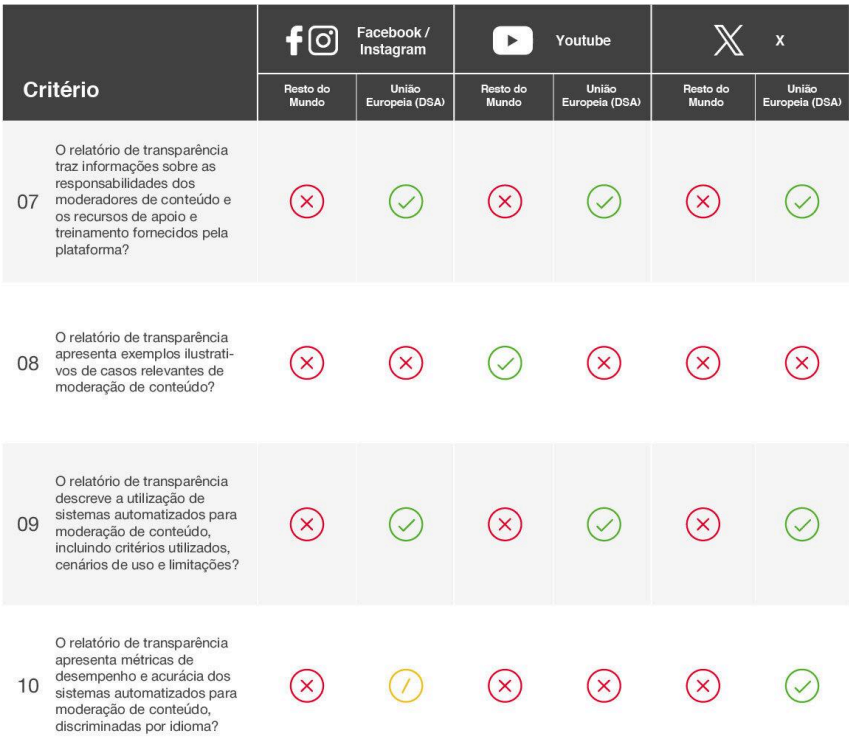
Figura 11 – Avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Disposições gerais*, segundo o tipo de relatório

Fonte: elaboração própria.

Figura 12 – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Disposições gerais*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
Disposições Gerais							
01	O relatório de transparência é publicado com uma periodicidade fixa, no mínimo semestral?	✓	✓	✓	✓	✗	✓
02	O relatório de transparência informa a data em que foi publicado, além do período coberto por seus dados?	/	✓	✗	✓	✗	/
03	É possível extrair os dados do relatório de transparência em formato estruturado e legível por máquinas?	✓	✗	/	✗	✗	✗
04	O relatório de transparência apresenta um glossário ou seções explicativas que descrevem de forma clara as dimensões dos dados apresentados?	/	✓	✓	✓	✓	✓
05	O relatório de transparência apresenta o número médio mensal de usuários da plataforma em cada país de atuação?	✗	/	✗	/	✗	✓
06	O relatório de transparência apresenta o número de moderadores especializados em cada idioma e/ou dedicados a cada país de atuação?	✗	✓	✗	✓	✗	✓

Figura 12 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Disposições gerais*, segundo o tipo de relatório



Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
07	O relatório de transparência traz informações sobre as responsabilidades dos moderadores de conteúdo e os recursos de apoio e treinamento fornecidos pela plataforma?	✗	✓	✗	✓	✗	✓
08	O relatório de transparência apresenta exemplos ilustrativos de casos relevantes de moderação de conteúdo?	✗	✗	✓	✗	✗	✗
09	O relatório de transparência descreve a utilização de sistemas automatizados para moderação de conteúdo, incluindo critérios utilizados, cenários de uso e limitações?	✗	✓	✗	✓	✗	✓
10	O relatório de transparência apresenta métricas de desempenho e acurácia dos sistemas automatizados para moderação de conteúdo, discriminadas por idioma?	✗	/	✗	✗	✗	✓

Fonte: elaboração própria.

Um primeiro aspecto negativo é que nenhuma das plataformas analisadas disponibiliza os dados de seus relatórios de transparência publicados em cumprimento ao DSA em formato estruturado para análise (Critério 3). O texto normativo apenas determina que os relatórios sejam disponibilizados publicamente, com fácil acesso, em um formato legível por máquina (European Parliament, 2022). Em contraste, os relatórios voluntários do Facebook, Instagram e YouTube são acompanhados por bases de dados em formato CSV, o que facilita sua manipulação e análise por meio de ferramentas de planilhas ou linguagens de programação. Inclusive, os dados dessas bases estruturadas foram utilizados na avaliação dos critérios dos eixos de avaliação subsequentes. Essa ausência representa uma falha importante da regulação: o DSA não estabelece um formato específico para a publicação dos relatórios, tampouco exige que os dados sejam oferecidos de maneira estruturada. Outro aspecto relevante é que, entre os documentos analisados, o relatório de transparência publicado voluntariamente pelo YouTube é o único a apresentar exemplos concretos de conteúdos moderados (Critério 8), uma prática recomendada na literatura por conferir maior materialidade às ações de

moderação das plataformas (Radsch, 2022; Urman; Makhortykh, 2023). Trata-se, logo, de mais um ponto negligenciado pelas determinações do DSA.

Apesar dessas limitações, é interessante que o DSA exija que as plataformas forneçam informações mais contextualizadas e aprofundadas sobre diferentes etapas de seus processos de moderação de conteúdo. Um exemplo concreto desse avanço é a obrigação, seguida por todas as plataformas analisadas (Critério 6), de detalhar o número de moderadores especializados para cada idioma do bloco europeu, o que possibilita identificar possíveis lacunas na supervisão em certas regiões – um problema histórico e amplamente documentado da falta de transparência das grandes plataformas (Nicholas; Bhatia, 2023; Shahid, 2024; Waldron, 2023). Além disso, os relatórios apresentam informações relevantes sobre as atribuições dos moderadores de conteúdo, grupo que, como discutido no capítulo anterior, representa uma força de trabalho silenciosa e invisibilizada (Critério 7), contribuindo para uma compreensão mais aprofundada de seus papéis na estrutura de governança de cada plataforma.

Os relatórios exigidos pelo DSA também oferecem informações mais detalhadas sobre o uso de sistemas automatizados nos processos de moderação de conteúdo. Enquanto os relatórios de transparência voluntários geralmente se limitam a declarações genéricas, como "*n% do conteúdo moderado foi avaliado automaticamente*", é obrigatório que os documentos publicados sob exigência do DSA descrevam de forma mais precisa as etapas de detecção de conteúdos violativos e a aplicação das diretrizes da comunidade por esses modelos, conferindo maior concretude e inteligibilidade a seu funcionamento (Critério 9). Nesse mesmo sentido, as plataformas são obrigadas a divulgar indicadores de acurácia e taxas de erro desses sistemas automatizados, ecoando recomendações expressas nos SCP (*Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2021). Curiosamente, essa exigência não é plenamente atendida pelo Google no caso do YouTube (Critério 10). Outra exigência do DSA que não é devidamente cumprida pela empresa diz respeito à divulgação do número de usuários por país, dado essencial para contextualizar as ações de moderação em escala local e regional (Critério 5). Em vez de apresentar essas informações diretamente no relatório, a empresa apenas redireciona para uma página externa que as reúne, o que compromete tanto a acessibilidade quanto a completude do documento.

4.2.2 Moderação de conteúdo por determinação da plataforma

Como discutido anteriormente, todas as grandes plataformas de redes sociais precisam, por diferentes razões, moderar o conteúdo e as atividades de seus usuários. Muitas, porém, preferem não tratar deste fato voluntária e abertamente (Gillespie, 2018a), algo que os resultados da nossa análise parecem confirmar, ao menos em certa medida. Como mostra a Figura 13, Facebook, Instagram e X/Twitter apresentaram avanços nos níveis de transparência das ações de moderação de conteúdo realizadas com base em seus próprios termos, quando comparamos os relatórios de transparência exigidos pelo DSA com os relatórios voluntários. Facebook e Instagram saltam de duas para oito avaliações parciais na comparação entre os relatórios voluntários e os exigidos pelo DSA, enquanto o X/Twitter apresenta o melhor desempenho: seu relatório sob o DSA alcança sete avaliações positivas, frente a apenas quatro parciais conquistas por seu relatório de transparência voluntário. Na contagem dos critérios de avaliação, o desempenho dos relatórios de transparência de moderação de conteúdo publicados pelo YouTube permanece o mesmo, com seis avaliações parciais. O detalhamento destas avaliações pode ser conferido na Figura 14.

Figura 13 – Avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Moderação de conteúdo por determinação da plataforma*, segundo o tipo de relatório

Fonte: elaboração própria.

Figura 14 – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Moderação de conteúdo por determinação da plataforma*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
Moderação de conteúdo por determinação da plataforma							
11	O relatório de transparência apresenta o número de publicações orgânicas removidas por determinação da plataforma, discriminadas por país e tipo de violação?	/	/	/	/	/	✓
11.1	O relatório de transparência especifica o número de publicações orgânicas removidas proativa e reativamente por determinação da plataforma, discriminadas por país e tipo de violação?	/	✗	/	✗	✗	✓
11.2	O relatório de transparência especifica o número de publicações orgânicas removidas por determinação dos sistemas automatizados da plataforma, discriminadas por país e tipo de violação?	✗	/	/	/	/	✓
11.3	O relatório de transparência apresenta informações agregadas ou médias do engajamento e/ou alcance das publicações orgânicas removidas no momento da moderação, discriminados por país e tipo de violação?	✗	✗	/	✗	✗	✗
12	O relatório de transparência apresenta o número de publicações orgânicas com alcance reduzido por determinação da plataforma, discriminadas por país?	✗	/	✗	/	✗	✓
13	O relatório de transparência apresenta o número de anúncios removidos por determinação da plataforma, discriminados por país e tipo de violação?	✗	/	✗	/	✗	✗

Figura 14 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Moderação de conteúdo por determinação da plataforma*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
13.1	O relatório de transparência especifica o número de anúncios removidos proativa e reativamente por determinação da plataforma, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✗
13.2	O relatório de transparência especifica o número de anúncios removidos por determinação dos sistemas automatizados da plataforma, discriminados por país e tipo de violação?	✗	/	✗	/	✗	✗
13.3	O relatório de transparência apresenta informações agregadas ou médias do alcance dos anúncios removidos no momento da moderação, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✗
14	O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos por determinação da plataforma, discriminados por país e tipo de violação?	✗	/	/	/	/	✓
14.1	O relatório de transparência especifica o número de usuários restritos, suspensos e/ou banidos proativa e reativamente por determinação da plataforma, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✓
14.2	O relatório de transparência especifica o número de usuários suspensos e/ou banidos por determinação dos sistemas automatizados da plataforma, discriminados por país e tipo de violação?	✗	/	✗	✗	/	✓

Figura 14 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Moderação de conteúdo por determinação da plataforma*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
14.3	O relatório de transparência apresenta informações agregadas ou médias do engajamento e/ou alcance dos usuários restritos, suspensos e/ou banidos no momento da moderação, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✗
15	O relatório de transparência apresenta informações sobre outros tipos de conteúdos e interações orgânicas removidos e/ou restritos por determinação da plataforma (por exemplo, comentários, classificados de *marketplace*)?	✗	/	/	✗	✗	✗

Fonte: elaboração própria.

Os avanços são particularmente evidentes no caso do X/Twitter, cujo relatório de transparência publicado em cumprimento ao DSA e voltado ao público da União Europeia é o único a apresentar conformidade plena para com quaisquer critérios deste eixo. Isso se deve ao fato de que as informações sobre moderação realizada por iniciativa da própria plataforma são apresentadas de forma detalhada e com a granularidade esperada. É possível consultar, sem dificuldade, dados sobre a remoção de publicações orgânicas (Critérios 11, 11.1, 11.2) e o banimento de usuários (Critérios 14, 14.1, 14.2) segmentados por país, por método de detecção (por denúncias de usuários ou detecção automática proativa), por tipo de agente responsável pela moderação (moderadores humanos ou sistemas automatizados) e por tipo de violação, sendo ainda viável cruzar essas variáveis entre si. Vale destacar que este é o único relatório de transparência analisado que discrimina as ações de redução de visibilidade de publicações orgânicas por país (Critério 12), prática reconhecida pelo DSA como moderação de conteúdo (European Parliament, 2022). Além disso, o relatório informa os métodos usados para detectar essas publicações e indica o tipo de agente responsável pela moderação.

Por outro lado, apesar de o relatório voluntário do X/Twitter apresentar o princípio da "liberdade de expressão, mas não liberdade de alcance", já aludido aqui, ele não inclui informações sobre a diminuição do alcance das publicações na plataforma (Critério 12). O relatório permite apenas a navegação por dados agregados globalmente sobre remoção de publicações orgânicas (Critérios 11 e 11.2) e banimento de usuários (Critérios 14 e 14.2), segmentados por tipo de violação e por agente de moderação, sem oferecer a possibilidade de

filtragem por região geográfica. Ambos os relatórios publicados pelo X/Twitter e analisados neste trabalho apresentam uma limitação significativa ao não disponibilizarem informações sobre a moderação de anúncios (Critério 13), como se tais conteúdos estivessem isentos de moderação, nem sobre o engajamento alcançado por publicações (Critério 11.3) e usuários moderados (Critério 14.3).

Assim sendo, o X/Twitter é a única plataforma analisada que não divulga qualquer informação sobre a moderação de conteúdo publicitário em seus relatórios de transparência, sejam eles voluntários ou exigidos pelo DSA. Isso reforça alegações de que as plataformas tendem a ser menos propensas tanto a moderar quanto a divulgar informações sobre suas operações comerciais, justamente aquelas que têm impacto mais direto sobre seus modelos de negócios e sua saúde financeira (Santini *et al.*, 2025). Portanto, mesmo que o DSA estabeleça que as plataformas devem divulgar informações sobre qualquer ação de moderação de conteúdo realizada no período em questão (European Parliament, 2022), é nítido que certas ações ainda são ofuscadas e ocultadas do público geral em benefício próprio.

Os documentos publicados pela Meta para o Facebook e o Instagram também revelam uma melhora limitada nos níveis de transparência, quando comparados os relatórios voluntários com aqueles divulgados em cumprimento às exigências do DSA. Os relatórios de transparência voluntários revelam pouquíssimas informações de fato relevantes, apenas a quantidade de publicações orgânicas removidas em todo o mundo pelas plataformas, discriminando-as por tipo de violação (Critério 11) e por tipo de detecção proativa ou reativa (Critério 11.1). A escolha das informações divulgadas, neste caso, é estratégica, uma vez que as plataformas relatam altos índices de ações de moderação de conteúdo *proativas*, ao invés de *reativas* – cerca de 95% para *exploração sexual, abuso ou nudez infantil*, 99% para *terrorismo e ódio organizado* e 99,5% para *discurso de ódio* no Facebook ao longo do período do relatório analisado. Assim, transmitem às partes interessadas a imagem de plataformas socialmente responsáveis e constantemente vigilantes, adotando uma postura de segurança perante seus anunciantes e afastando possíveis regulações e intervenções tidas como indesejadas por autoridades públicas e legisladores.

Enquanto isso, os relatórios publicados pela Meta por exigência do DSA divulgam informações não só sobre as ações de remoção de conteúdo orgânico por determinação de ambas as plataformas (Critério 11), mas também de banimento e restrição de usuários (Critério 14) e deleção de anúncios (Critério 13). Mesmo assim, apenas as estatísticas de publicações orgânicas removidas são destrinchadas por tipo de violação; as demais informações, por outro lado, são apresentadas de maneira agregada para toda a União

Europeia, acompanhadas apenas de dados auxiliares sobre o volume de ações de moderação realizadas por sistemas automatizados (Critérios 11.2, 13.2 e 14.2).

Já a comparação entre os relatórios publicados pelo Google para o YouTube indica que, embora as exigências previstas no DSA possam ser úteis, elas também têm o potencial de abrir novas lacunas de transparência. Assim como no caso do Facebook e do Instagram, o DSA impulsiona avanços ao levar o Google a ser mais transparente quanto à interrupção da circulação de anúncios veiculados em vídeos do YouTube – mesmo que, mais uma vez, os números dessas ações de moderação não sejam discriminados por tipo de violação (Critério 13). Em contraste, os relatórios voluntários de moderação de conteúdo publicados pelo YouTube se destacam por serem os únicos a apresentar dados de engajamento das publicações moderadas pela plataforma (Critério 11.3). Mais especificamente, esses relatórios apresentam a métrica da taxa de visualizações com violações, que estima a proporção de visualizações recebidas por vídeos moderados ao longo de um determinado intervalo de tempo (Google, [*S. d.*]b). Da mesma forma, informações sobre a moderação de comentários em vídeos do YouTube estão ausentes dos relatórios de transparência disponibilizados para o público e autoridades europeias (Critério 15). Esses elementos indicam que a regulação poderia ter sido mais ambiciosa, já que as plataformas demonstram possuir os recursos técnicos e operacionais necessários para atender a padrões de transparência mais rigorosos, reforçando a urgência de se estabelecer critérios técnicos claros e mínimos diretamente na legislação.

4.2.3 Denúncias realizadas por usuários

Dada a centralidade das denúncias feitas por usuários nas ações de moderação de conteúdo realizadas pelas grandes plataformas de redes sociais, reforçada pela introdução de novos marcos normativos, seria razoável esperar que esse aspecto recebesse maior destaque nos relatórios de transparência. Na prática, porém, isso não se confirma na maioria dos documentos avaliados: como mostra a Figura 15, com detalhamento na Figura 16, alguns relatórios ignoram completamente o tema, com destaque negativo para os relatórios voluntários publicados pelas plataformas em nível global. Os relatórios de transparência de moderação de conteúdo voluntários publicados por Facebook e Instagram recebem somente avaliações negativas neste eixo, mas os relatórios de transparência publicados por exigência do DSA alcançam seis avaliações parciais e duas positivas. Também se observam avanços nos relatórios de transparência do YouTube e do X/Twitter, que passam, respectivamente, de duas

avaliações parciais e uma positiva, e de uma avaliação parcial, para dez avaliações parciais e seis positivas.

Figura 15 – Avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Denúncias realizadas por usuários*, segundo o tipo de relatório

Fonte: elaboração própria.

Figura 16 – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Denúncias realizadas por usuários*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
Denúncias realizadas por usuários							
16	O relatório de transparência apresenta o número de denúncias de publicações orgânicas feitas por usuários, discriminadas por país e tipo de violação?	✗	/	/	/	/	✓
16.1	O relatório de transparência apresenta o número de publicações orgânicas removidas após denúncias de usuários, discriminadas por país e tipo de violação?	✗	✗	/	/	✗	✓
16.2	O relatório de transparência apresenta o número de denúncias de publicações orgânicas feitas por usuários, discriminadas por tipo de processamento (automático ou manual)?	✗	✓	✗	/	✗	✓
16.3	O relatório de transparência apresenta o número de denúncias de publicações orgânicas, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?	✗	/	✓	/	✗	✓
16.4	O relatório de transparência apresenta o tempo médio para lidar com denúncias de publicações orgânicas feitas por usuários, discriminado por país e tipo de violação?	✗	/	✗	/	✗	✓
17	O relatório de transparência apresenta o número de denúncias de anúncios feitas por usuários, discriminadas por país e tipo de violação?	✗	/	✗	/	✗	✗

Figura 16 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Denúncias realizadas por usuários*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
17.1	O relatório de transparência apresenta o número de anúncios removidos após denúncias de usuários, discriminados por país e tipo de violação?	✗	✗	✗	/	✗	✗
17.2	O relatório de transparência apresenta o número de denúncias de anúncios feitas por usuários, discriminadas por tipo de processamento (automático ou manual)?	✗	✓	✗	/	✗	✗
17.3	O relatório de transparência apresenta o número de denúncias de anúncios, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?	✗	/	✗	/	✗	✗
17.4	O relatório de transparência apresenta o tempo médio para lidar com denúncias de anúncios feitas por usuários, discriminado por país e tipo de violação?	✗	/	✗	/	✗	✗
18	O relatório de transparência apresenta o número de denúncias de usuários feitas por outros usuários, discriminadas por país e tipo de violação?	✗	✗	✗	✗	✗	✗
18.1	O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos após denúncias de outros usuários, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✓

Figura 16 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Denúncias realizadas por usuários*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
18.2	O relatório de transparência apresenta o número de denúncias de usuários feitas por outros usuários, discriminadas por tipo de processamento (automático ou manual)?	⊗	⊗	⊗	⊗	⊗	⊗
18.3	O relatório de transparência apresenta o número de denúncias de usuários, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?	⊗	⊗	⊗	⊗	⊗	⊗
18.4	O relatório de transparência apresenta o tempo médio para lidar com denúncias de usuários feitas por outros usuários, discriminado por país e tipo de violação?	⊗	⊗	⊗	⊗	⊗	⊗

Fonte: elaboração própria.

O maior nível de detalhamento sobre denúncias feitas por usuários nos relatórios de transparência exigidos pelo DSA se deve à exigência, prevista no artigo 16 (*Notice and Action Mechanisms*), de que as plataformas implementem mecanismos específicos e acessíveis para que usuários possam reportar conteúdos potencialmente ilegais em território europeu (European Parliament, 2022). Assim como outros mecanismos previstos pelo DSA, não necessariamente relacionados à moderação de conteúdo, esses dispositivos são acompanhados por obrigações rigorosas de transparência, que envolvem a divulgação clara e acessível de dados e informações sobre sua utilização, abrangendo desde o número de denúncias recebidas até os resultados e desdobramentos de ações concretas de moderação.

Por exemplo, embora os relatórios de transparência voluntários da Meta para Facebook e Instagram não mencionem nem divulguem informações sobre denúncias feitas por usuários para remoção de conteúdo, os documentos exigidos pelo DSA trazem um nível de detalhamento consideravelmente maior nesse aspecto. Neles, a empresa orienta os usuários a utilizarem os canais específicos de denúncia e afirma que, caso o conteúdo viole suas diretrizes da comunidade ou políticas de publicidade, ele é removido. A Meta também declara que todas as notificações feitas com base no artigo 16 do DSA são revisadas manualmente por seus moderadores, sem o uso de sistemas automatizados (Critérios 16.2 e 17.2).

Em termos de dados estatísticos, é apresentada uma tabela por plataforma com o número total de denúncias feitas por usuários, organizadas por tipo de violação e pelo desfecho da análise realizada – ou seja, se as denúncias foram acatadas ou não (Critérios 16 e 17). Um dos principais problemas dessas tabelas, no entanto, é a ausência de distinção entre conteúdos orgânicos e anúncios, tanto nas denúncias quanto nas ações de remoção, apesar de a própria Meta reconhecer que ambos os tipos podem ser alvo de moderação. Além disso, os dados não são segmentados por país, sendo apresentados de forma agregada para toda a União Europeia, o que resulta em um nível de granularidade inferior ao considerado adequado pelos critérios do nosso quadro analítico.

Como apresentado anteriormente, YouTube e X/Twitter destinam seções e/ou subseções específicas de seus relatórios de transparência voluntários ao detalhamento deficiente de informações sobre denúncias de usuários. Primeiramente, o YouTube divulga a quantidade de vídeos denunciados por seus usuários no período coberto pelo relatório, com dados globais. As denúncias são discriminadas de acordo com os tipos de violações apontadas, com a observação de que um mesmo vídeo pode ser denunciado múltiplas vezes e por diferentes motivos (Critério 16). A plataforma também exibe um breve ranking dos países de onde mais se originaram essas denúncias, embora não apresente dados ou estatísticas que sustentem essa lista. Já o relatório de transparência voluntário do X/Twitter recebeu apenas uma avaliação parcial no primeiro critério deste eixo, tendo em vista que, apesar de apresentar o número global de denúncias feitas por usuários contra publicações orgânicas, classificadas por tipo de violação, não há discriminação desses dados por país (Critério 16).

Os relatórios publicados por ambas as plataformas em cumprimento às exigências do DSA apresentam avanços parciais na granularidade dos dados. Por exemplo, o relatório do Google traz informações sobre denúncias feitas por usuários contra vídeos e anúncios no YouTube, classificadas por tipo de violação – embora sem separar os dados conforme o tipo de conteúdo (Critérios 16 e 17). Além disso, as denúncias são discriminadas segundo o perfil do usuário, com destaque para o volume realizado por *trusted flaggers* (Critérios 16.3 e 17.3), e o método de processamento, automático ou manual (Critérios 16.2 e 17.2). Já o relatório do X/Twitter oferece tabelas detalhadas tanto sobre as denúncias contra publicações orgânicas quanto sobre as medidas adotadas, com dados discriminados por tipo de violação, perfil do usuário, país e método de processamento (Critérios 16, 16.1, 16.2, 16.3 e 16.4). Contudo, devido a lacunas significativas na granularidade, como a ausência de informações específicas sobre denúncias contra anúncios no X/Twitter (Critério 17) e a falta de detalhamento por país no caso do YouTube (Critério 16), as informações de transparências disponibilizadas por

exigência da regulação ainda não refletem plenamente a diversidade das ações de moderação baseadas nas denúncias dos usuários.

4.2.4 Restauração de conteúdo e contestações à moderação

Reconhecer que as ações de moderação de conteúdo não são permanentes constitui uma importante salvaguarda dos direitos dos usuários em ambientes online. Embora as plataformas de redes sociais frequentemente não sejam transparentes com os usuários afetados pela moderação quanto aos motivos que levaram à aplicação dessas medidas (Suzor *et al.*, 2019), muitas vezes oferecem oportunidades para que esses usuários contestem tais decisões. Além disso, em certas ocasiões, as publicações e/ou perfis são restaurados proativamente pelas próprias plataformas, a partir do reconhecimento de eventuais erros. Contudo, ainda é possível questionar a efetividade real das contestações permitidas pelas plataformas (Are, 2024), assim como o alcance efetivo das ações proativas de restauração de conteúdo moderado por elas realizadas (Clune; McDaid, 2023).

Com exceção do relatório voluntário publicado globalmente pelo X/Twitter, todos os demais documentos analisados abordam esses pontos de alguma forma, como demonstrado na Figura 17 e na Figura 18. Na comparação entre os relatórios de transparência voluntários e os produzidos em conformidade com o DSA, o número de avaliações parciais sobe de três para quatro no caso do Facebook e do Instagram, enquanto o YouTube passa de duas para quatro. Se o relatório voluntário do X/Twitter recebeu apenas avaliações negativas, o documento publicado na União Europeia obteve quatro avaliações parciais.

Figura 17 – Avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Restauração de conteúdo e contestações à moderação*, segundo o tipo de relatório

Fonte: elaboração própria.

Figura 18 – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Restauração de conteúdo e contestações à moderação*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
Restauração de conteúdo e contestações à moderação							
19	O relatório de transparência apresenta o número de contestações à remoção de publicações orgânicas, discriminadas por país e tipo de violação?	/	/	/	/	✗	/
19.1	O relatório de transparência apresenta o número de publicações orgânicas restauradas, discriminadas por país e tipo de violação, após contestações de usuários?	/	/	/	/	✗	/
20	O relatório de transparência apresenta o número de publicações orgânicas restauradas, discriminadas por país e tipo de violação, após a identificação proativa de erros na moderação?	/	✗	✗	✗	✗	✗
21	O relatório de transparência apresenta o número de contestações à restrição, suspensão e/ou ao banimento de usuários, discriminadas por país?	✗	/	✗	/	✗	/
21.1	O relatório de transparência apresenta o número de usuários restaurados, discriminados por país, após contestações dos próprios usuários moderados?	✗	/	✗	/	✗	/
22	O relatório de transparência apresenta o número de usuários restaurados discriminados por país, após a identificação proativa de erros na moderação?	✗	✗	✗	✗	✗	✗

Fonte: elaboração própria.

Assim como no caso das denúncias realizadas por usuários, o DSA também avança ao exigir, em seu artigo 20 (*Internal complaint-handling system*), que as plataformas digitais implementem mecanismos destinados a receber contestações de usuários que se sintam prejudicados por ações de moderação (European Parliament, 2022). Como já discutido anteriormente, essa exigência vem acompanhada de obrigações específicas de transparência quanto ao funcionamento e aos resultados desses mecanismos. Porém, mesmo diante de obrigações legais específicas, chama a atenção, de forma negativa, a discrepância entre os

níveis de detalhamento e qualidade das informações apresentadas nos diferentes relatórios exigidos pelo DSA.

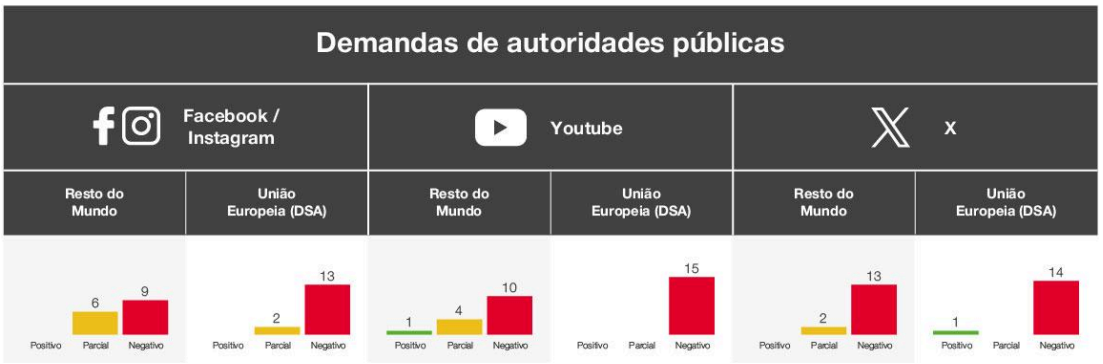
O relatório de transparência do X/Twitter voltado à União Europeia, por exemplo, apresenta estatísticas sobre contestações à remoção de publicações orgânicas (Critério 19) e ao banimento de usuários (Critério 21), discriminadas por país, além do número de decisões de moderação revertidas em decorrência dessas contestações (Critérios 19.1 e 21.1). Já os documentos publicados pela Meta para o Facebook e o Instagram incluem dados sobre contestações à remoção de conteúdos orgânicos, mas sem segmentação geográfica, de modo que as informações são somente organizadas por tipo de violação (Critério 19). No caso de suspensões de contas, os números são apresentados de forma agregada, sem qualquer detalhamento quanto à natureza da violação ou à localização dos usuários afetados (Critério 21). O relatório do YouTube, por sua vez, exibe um nível ainda mais limitado de detalhamento: apresenta apenas estatísticas gerais sobre contestações a decisões de moderação, sem qualquer desagregação, não havendo distinção por país nem pelo tipo de violação alegada (Critérios 19 e 21). Logo, observamos um descompasso evidente entre o nível de transparência adotado por cada plataforma, o que impede que os dados divulgados alcancem um grau satisfatório de granularidade, mesmo sob um regime regulatório que, em tese, deveria restringir a arbitrariedade das empresas quanto ao que revelar ou ocultar.

Outro aspecto que merece destaque nessa comparação diz respeito às ações *proativas* de restauração de conteúdo. Entre todos os relatórios analisados, apenas os documentos voluntários publicados globalmente pelo Facebook e pelo Instagram disponibilizam informações sobre esse tipo de medida. Neles, a Meta diferencia os conteúdos restaurados com ou sem a necessidade de contestação, organizando os dados por tipo de violação, embora sem qualquer segmentação por país (Critério 20). Isso não significa, porém, que as demais plataformas não adotem práticas semelhantes; o problema está na ausência de transparência sobre essas ações. Como o DSA não impõe exigências específicas para a divulgação dessas informações, as empresas seguem sem torná-las públicas, mesmo em contextos em que há uma expectativa maior de abertura. Isso contribui para ofuscar o fato de que decisões de moderação também são tomadas de forma autônoma pelas próprias plataformas, independentemente de pressões externas, inclusive de seus próprios usuários.

4.2.5 Demandas de autoridades públicas

Contrariando o que aponta parte da literatura sobre transparência em plataformas de redes sociais (ver Leone de Castris, 2024; Urman; Makhortykh, 2023; Vergara; Jain; Mehta, 2024), a transparência em relação a solicitações e ordens de moderação de conteúdo por autoridades públicas foi a pior avaliada em nossa análise, tanto no contexto da transparência voluntária global quanto no marco regulatório da União Europeia. Dos quinze critérios avaliados, os relatórios de transparência voluntários do X/Twitter e Facebook e Instagram obtiveram apenas duas e seis avaliações parciais, respectivamente, enquanto o YouTube apresentou desempenho um pouco superior, com quatro avaliações parciais e uma positiva; já nos relatórios previstos pelo DSA, Facebook e Instagram caíram para duas avaliações parciais, ao passo que o X/Twitter teve uma avaliação positiva e o YouTube foi avaliado de maneira negativa em todos os critérios, como mostram a Figura 19 e a Figura 20.

Figura 19 – Avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Demandas de autoridades públicas*, segundo o tipo de relatório



Fonte: elaboração própria.

Figura 20 – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Demandas de autoridades públicas*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
Demandas de autoridades públicas							
23	O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição e/ou remoção de publicações orgânicas, discriminados por país e tipo de violação?	✗	/	✓	✗	/	✓
23.1	O relatório de transparência apresenta o número de publicações orgânicas restritas e/ou removidas por determinação de autoridades públicas, discriminadas por país e tipo de violação?	/	✗	/	✗	/	✗
23.2	O relatório de transparência indica se as publicações orgânicas restritas e/ou removidas após determinação de autoridades públicas foram moderadas por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?	✗	✗	/	✗	✗	✗
23.3	O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição e/ou remoção de publicações orgânicas, discriminadas por país?	/	✗	/	✗	✗	✗
23.4	O relatório de transparência informa a quantidade de pedidos de remoção de publicações orgânicas que tiveram origem em ordens judiciais, discriminados por país?	/	✗	/	✗	✗	✗
24	O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição e/ou remoção de anúncios, discriminados por país e tipo de violação?	✗	/	✗	✗	✗	✗

Figura 20 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Demandas de autoridades públicas*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
24.1	O relatório de transparência apresenta o número de anúncios restritos e/ou removidos por determinação de autoridades públicas, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✗
24.2	O relatório de transparência indica se os anúncios restritos e/ou removidos após determinação de autoridades públicas foram moderados por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?	✗	✗	✗	✗	✗	✗
24.3	O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição e/ou remoção de anúncios, discriminadas por país?	✗	✗	✗	✗	✗	✗
24.4	O relatório de transparência informa a quantidade de pedidos de remoção de anúncios que tiveram origem em ordens judiciais, discriminados por país?	✗	✗	✗	✗	✗	✗
25	O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição, suspensão e/ou banimento de usuários, discriminados por país e tipo de violação?	✗	✗	✗	✗	✗	✗
25.1	O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos por determinação de autoridades públicas, discriminados por país e tipo de violação?	/	✗	✗	✗	✗	✗

Figura 20 (continuação) – Detalhamento das avaliações dos relatórios de transparência de moderação de conteúdo selecionados nos critérios que compõem o eixo de *Demandas de autoridades públicas*, segundo o tipo de relatório

Critério		Facebook / Instagram		Youtube		X	
		Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)	Resto do Mundo	União Europeia (DSA)
25.2	O relatório de transparência indica se os usuários restritos, suspensos e/ou banidos após determinação de autoridades públicas foram moderados por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?	✗	✗	✗	✗	✗	✗
25.3	O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição, suspensão e/ou banimento de usuários, discriminadas por país?	/	✗	✗	✗	✗	✗
25.4	O relatório de transparência informa a quantidade de pedidos de restrição, suspensão e/ou banimento de usuários que tiveram origem em ordens judiciais, discriminados por país?	/	✗	✗	✗	✗	✗

Fonte: elaboração própria.

Começando pelo X/Twitter, plataforma que recebeu a pior avaliação entre os relatórios voluntários de moderação de conteúdo nesse eixo, as informações disponibilizadas são pouco reveladoras. A empresa apresenta uma tabela com o total de solicitações de moderação feitas por autoridades públicas globalmente (Critério 23), indicando quantas dessas foram efetivamente atendidas (Critério 23.1). A plataforma alega não atender "solicitações que sejam juridicamente defeituosas, excessivamente amplas e/ou que pareçam impor restrições indevidas à liberdade de expressão" (X/Twitter, 2025, n.p., tradução do autor). Em seguida, esses dados são detalhados apenas para quatro regiões específicas: Turquia, Japão, Coreia do Sul e União Europeia. Embora o relatório explique os diferentes tipos de pedidos de moderação realizados por autoridades, não há identificação ou detalhamento dessas entidades, mesmo nas regiões destacadas (Critério 23.3).

No relatório de transparência de moderação de conteúdo publicado pela plataforma em cumprimento ao DSA, são apresentadas apenas as quantidades de pedidos de remoção de conteúdo por país e tipo de violação (Critério 23). Esse nível de detalhamento é, em algum nível, significativo, em especial porque as demais plataformas analisadas não oferecem esse tipo de cruzamento de informações, mesmo nos relatórios exigidos pela regulação. Por outro lado, representa um certo retrocesso em relação ao relatório voluntário publicado pela mesma

plataforma, uma vez que não é possível verificar quantas dessas solicitações resultaram em ações concretas de moderação (Critério 23.1).

No *relatório de solicitações governamentais de remoção de conteúdo*, publicado globalmente, o Google apresenta dados sobre o YouTube e outras plataformas da empresa. O documento inclui: (i) solicitações de entes governamentais que alegam violações a leis locais; (ii) ordens judiciais determinando a remoção de conteúdo; e (iii) pedidos de revisão para verificar a conformidade de conteúdos com as diretrizes internas das plataformas (Google, [*S. d.*]a). O relatório informa, inicialmente, o número total de solicitações recebidas desde 2011, bem como o volume de itens de conteúdo abrangidos no período. Em seguida, são disponibilizados gráficos interativos que permitem filtrar os dados por país, motivo da solicitação, plataforma envolvida e entidade requerente da remoção.

O principal problema, contudo, é a impossibilidade de cruzar esses filtros na interface de sua página web, o que impede a visualização de dados específicos por plataforma em cada país, limitando o usuário a informações agregadas sobre todas as plataformas do Google. Apesar disso, o relatório recebeu avaliação positiva no primeiro critério deste eixo, por disponibilizar o número de pedidos de remoção de vídeos no YouTube por país e tipo de violação, ainda que essas informações só estejam acessíveis via arquivo CSV para *download* (Critério 23). Importante notar que o relatório diferencia entre conteúdos removidos por violação às políticas internas da empresa e aqueles restritos e/ou removidos em razão do descumprimento de legislações locais, o que permite, em certo grau, compreender o impacto real das solicitações governamentais (Critério 23.2), destacando-se como a única, entre todas as plataformas avaliadas, a fazê-lo.

Essa é, no entanto, toda a transparência fornecida pelo Google sobre as solicitações de moderação de conteúdo no YouTube feitas por autoridades públicas. No relatório exigido pelo DSA, a empresa apenas informa que esses dados já estão disponíveis em seu relatório global de transparência, sugerindo que, diante das exigências pouco rigorosas da regulação, seu padrão voluntário de transparência é suficiente para atender à nova legislação europeia. Por essa razão, o relatório do YouTube submetido ao DSA recebeu avaliações negativas nos critérios correspondentes.

O *relatório de restrições de conteúdo com base na legislação local* publicado pela Meta para Facebook e Instagram também é consideravelmente incompleto. O documento afirma que a Meta recebe "notificações de governos, órgãos reguladores e tribunais, bem como de entidades não governamentais, que são analisados de acordo com nossa Política Corporativa de Direitos Humanos e nossos compromissos como membros da *Global Network*

Initiative" (Meta, [*S. d.*], n.p., tradução do autor). O principal problema é o fato de o relatório detalhar apenas os casos em que o acesso a determinados conteúdos – como publicações orgânicas, eventos, comentários e perfis – foi *restringido* com base em legislações locais (Critério 23.2). Ou seja, o documento exclui situações em que conteúdos foram *removidos* após notificações de autoridades públicas e confirmação posterior de violação às políticas internas da plataforma, uma distinção fundamental apontada pelos SCP (2021). Afinal, sem esse nível de detalhamento, não é possível identificar quem está, de fato, influenciando as decisões de moderação de conteúdo das plataformas, tampouco compreender os efeitos concretos dessas decisões.

Ao contrário do X/Twitter, a Meta, ao menos, divulga alguns detalhes sobre as autoridades responsáveis pelas solicitações de moderação de conteúdo em suas plataformas (Critério 23.3). No período analisado, é revelado que, das cerca de 14,7 mil restrições regionais impostas em suas plataformas no Brasil, aproximadamente 2.370 no Facebook e 12.340 no Instagram, entre publicações orgânicas, perfis e comentários, 9,7 mil se deram por pedidos da Agência Nacional de Vigilância Sanitária (Anvisa) enquanto outros 3,8 mil se deram por ordens de tribunais locais relacionadas a processos cíveis, criminais e eleitorais (Meta, [*S. d.*]). No entanto, a granularidade dos dados é, novamente, insatisfatória. As descrições dos pedidos feitos por autoridades são vagas, sem detalhar as solicitações específicas de restrição de conteúdo. Por exemplo: quais foram os motivos que levaram a Anvisa a solicitar a remoção de tantos conteúdos diferentes? Que tipos de material foram alvo desses pedidos? Nada disso pode ser inferido das informações fornecidas pela Meta.

Os relatórios de transparência da Meta exigidos pelo DSA também trazem poucas informações relevantes. Eles apresentam uma breve explicação sobre os poderes de moderação de conteúdo atribuídos a autoridades públicas, conforme previsto no artigo 9 da regulação, e afirmam que, diante de uma solicitação para agir contra conteúdos supostamente ilegais, o material é primeiro avaliado com base nas diretrizes da comunidade e em outras políticas da empresa, como as de publicidade. Se o conteúdo violar essas políticas internas, é removido; se não as violar, mas for considerado ilegal, seu acesso é restringido na jurisdição da autoridade que fez o pedido (Meta, 2024a, b).

A partir disso, são apresentadas duas tabelas: a primeira indica o número de solicitações de autoridades públicas para moderação de conteúdo por país da União Europeia; a segunda, o número de solicitações por tipo de violação (Critérios 23 e 24). Como não é possível cruzar os dados entre ambas, qualquer análise mais aprofundada das informações é limitada. Assim, faltam dados sobre quais autoridades públicas fizeram os pedidos (Critérios

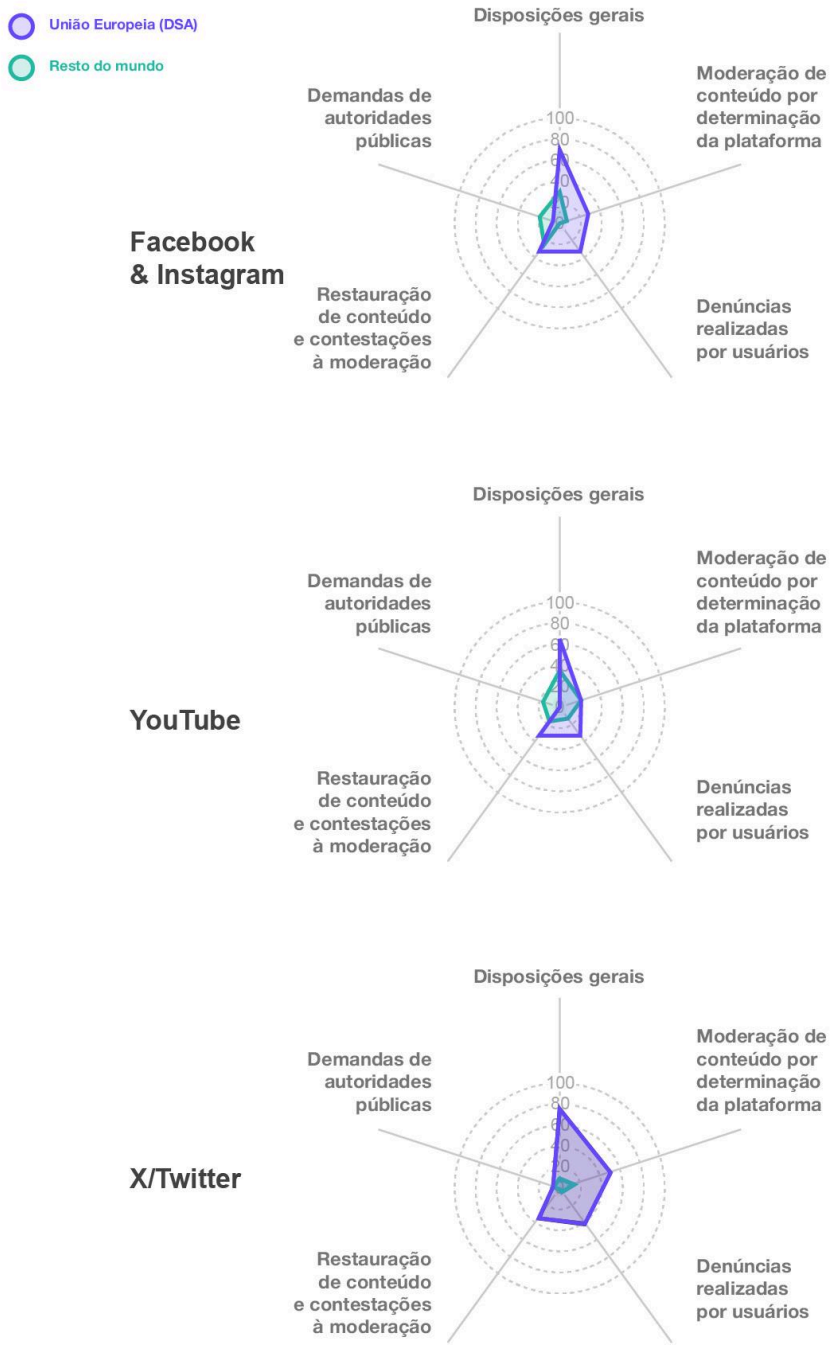
23.3 e 24.3), bem como informações sobre os efeitos dessas solicitações – incluindo se foram atendidas com base nas políticas internas das plataformas ou apenas por violarem legislações locais (Critérios 23.2 e 24.2). Apesar da menção da Meta a violações de suas políticas de publicidade, não há distinção entre solicitações de moderação de anúncios e publicações orgânicas (Critérios 23 e 24).

Vale destacar, por fim, que, embora a Meta e o YouTube indiquem explicitamente, em seus relatórios voluntários de moderação de conteúdo, quando as remoções resultam de ordens judiciais (Critério 23.4), essa informação não é apresentada nos relatórios publicados em cumprimento ao DSA. Essa omissão compromete a transparência e dificulta análises mais precisas sobre a origem das decisões de moderação, reduzindo a possibilidade de examinar criticamente o papel de diferentes atores na governança do discurso online e enfraquecendo, consequentemente, a prestação de contas e o escrutínio público sobre esses processos.

4.2.6 Visão geral dos relatórios de transparência

Para ilustrar de forma mais clara o desempenho geral dos relatórios de transparência analisados e, assim, responder às nossas questões de pesquisa, elaboramos os gráficos apresentados na Figura 21. Nela, os cinco eixos do nosso quadro analítico estão dispostos nas extremidades dos gráficos de radar, e quanto maior a área sombreada, melhor o desempenho do relatório em cada critério. As áreas sombreadas correspondentes aos relatórios publicados na União Europeia por exigência do DSA, em roxo, se sobrepõem àquelas dos relatórios voluntários divulgados em escala global, em verde. A pontuação em cada eixo foi obtida pela soma do percentual de avaliações positivas com metade do percentual das avaliações parciais.

Figura 21 – Desempenho comparado dos relatórios de transparência voluntários e exigidos pelo *Digital Services Act* analisados, por eixo de avaliação e por plataforma selecionada



Fonte: elaboração própria.

Em relação à **QP1** (*Qual é o nível de transparência voluntária da moderação de conteúdo das plataformas selecionadas em escala global?*), é possível observar que nenhum dos relatórios de transparência de moderação de conteúdo voluntários apresenta desempenho elevado em quaisquer eixos analisados. Pelo contrário, os resultados demonstram um nível

consistentemente baixo, mesmo diante dos compromissos públicos assumidos pelas plataformas em favor da transparência, como o apoio declarado aos SCP (Urman; Makhortykh, 2023). Nesse contexto de baixa transparência, o relatório mais equilibrado em termos positivos é o do YouTube, que apresenta pontuações relativamente consistentes em todos os eixos analisados; no extremo oposto, encontra-se o relatório voluntário do X/Twitter, cujo desempenho geral é praticamente nulo. O relatório de transparência da Meta, referente ao Facebook e Instagram, apresenta lacunas evidentes, destacando a divulgação de ações de moderação motivadas por demandas de autoridades públicas, em detrimento das baseadas nas políticas internas das plataformas.

Más práticas de transparência latentes, no que diz respeito à **QP2** (*Quais práticas de moderação de conteúdo são mais opacas nos relatórios de transparência voluntários publicados globalmente pelas plataformas selecionadas?*), incluem, notadamente, a ausência de qualquer menção, por parte das plataformas selecionadas, à moderação de conteúdo publicitário, o que contribui para a manutenção da opacidade que rege suas operações comerciais. Afinal, fornecer mais detalhes publicamente pode fazer com que elas acabem desmascarando um ecossistema publicitário e de monetização de conteúdo extremamente nocivo e permissivo, contrariando seus próprios termos de serviço e diretrizes da comunidade (Santini *et al*., 2025).

De modo geral, a granularidade dos dados e das informações divulgadas nos relatórios é insatisfatória. As informações são apresentadas de forma agregada e genérica, raramente sendo detalhadas por tipo de violação que motivou a ação de moderação ou pelo país em que ela ocorreu, ou, ainda, pelo país de residência dos usuários diretamente afetados. Essa limitação dificulta a realização de análises contextuais com o devido aprofundamento e a comparação das informações disponibilizadas entre plataformas. Mais especificamente, chama a atenção o fato de o relatório publicado voluntariamente pela Meta para Facebook e Instagram ter um desempenho completamente nulo no eixo de *Denúncias realizadas por usuários*, apesar de estas representarem um dos principais alicerces das ações de moderação de conteúdo das grandes plataformas de redes sociais (Crawford; Gillespie, 2016). Por outro lado, este é o único relatório a revelar informações sobre as ações de restauração de conteúdo conduzidas proativamente após a moderação, informações indisponíveis em qualquer outro relatório analisado. Informações contextuais que facilitariam a compreensão dos relatórios de transparência pelas partes interessadas, conforme previsto no eixo de *Disposições gerais*, também estão reiteradamente ausentes nos documentos publicados de forma voluntária. Por fim, o nível de transparência voluntária do X/Twitter em relação às ações de moderação

motivadas por solicitações de autoridades públicas é mínimo. Uma possível hipótese para tal é que a omissão dessas informações contribui para sustentar a narrativa promovida por Elon Musk de que a plataforma seria alvo desproporcional de "censura estatal" (Lima-Strong, 2024), afastando as partes interessadas de dados que poderiam contestar essa alegação.

Já quanto à **QP3** (*Em que medida o Digital Services Act contribui para o aprimoramento da transparência da moderação de conteúdo das plataformas selecionadas na União Europeia?*), a Figura 21 evidencia, à primeira vista, os avanços no nível de transparência dos relatórios de transparência de moderação de conteúdo selecionados. Mais do que isso, os gráficos revelam um padrão consistente nas áreas sombreadas, cujos contornos se repetem de forma notavelmente similar entre os relatórios analisados. O DSA leva as plataformas a adotarem maior transparência em relação às ações de moderação realizadas por iniciativa própria – muito mais, aliás, do que em relação às ações executadas por determinação de autoridades públicas. Isso fica nítido quando todas as plataformas analisadas passam a fornecer informações sobre publicações que tiveram seu alcance reduzido, algo que não é feito, voluntariamente, por nenhuma delas, contribuindo para o entendimento dessa prática como uma ação de moderação de conteúdo. Além disso, contribui para equilibrar os níveis de transparência das ações de restauração de conteúdo, especialmente quando realizadas de forma reativa, em resposta a contestações de usuários. No eixo de *Restauração de conteúdo e contestações à moderação*, todos os relatórios de transparência publicados obrigatoriamente tiveram a mesma pontuação. Resultados esperados, considerando que a legislação institui mecanismos específicos para que usuários possam contestar ações de moderação e denunciar conteúdos potencialmente ilegais.

Porém, apesar dos avanços proporcionados pelo DSA, os resultados ainda deixam a desejar: o volume expressivo de avaliações negativas indica que os problemas de transparência da moderação de conteúdo persistem de forma relevante. Em resposta à **QP4** (*Quais práticas de moderação de conteúdo permanecem mais opacas nos relatórios de transparência publicados pelas plataformas em cumprimento ao Digital Services Act na União Europeia?*), nossa análise indica que as regras e critérios propostos por especialistas para promover a transparência das ações de moderação de conteúdo continuam amplamente desatendidos, mesmo sob a vigência da regulação da União Europeia. Este fato preocupa, especialmente, quando levamos em consideração que o DSA se apresenta como o principal norte para a regulação de plataformas e serviços digitais em regiões diversas do mundo (European Commission, 2024), a exemplo da própria América Latina (Bizberge; Mastrini; Gómez, 2023; Helberger; Samuelson, 2024).

Mais especificamente, o DSA não é capaz de solucionar, por exemplo, o problema da granularidade dos dados e informações disponibilizados nos relatórios de transparência. A única plataforma analisada a repetidamente apresentar informações com a granularidade esperada – isto é, discriminando as ações de moderação de conteúdo aplicadas por país e tipo de violação percebida – é o X/Twitter. Não à toa, a plataforma tem quatorze avaliações positivas somando os eixos de *Moderação de conteúdo por determinação da plataforma*, *Denúncias realizadas por usuários*, *Restauração de conteúdo e contestações à moderação* e *Demandas de autoridades públicas*, contra apenas duas do Facebook e Instagram e nenhuma do YouTube. Com exceção do X/Twitter, as plataformas também perdem muito na transparência do eixo de *Demandas de autoridades públicas*, como já evidenciado pela Figura 21 – o YouTube, como notado anteriormente, nem sequer pontua nele.

Preocupa, ainda, o fato de que algumas recomendações de transparência incorporadas ao nosso quadro analítico sejam observadas pelas plataformas em seus relatórios voluntários, mas não nos relatórios publicados em conformidade com o DSA. Esperava-se que o DSA consolidasse e aprofundasse práticas voluntárias já recorrentemente adotadas pelas grandes plataformas de redes sociais (Leerssen, 2024). Na prática, muitos desses esforços mínimos são negligenciados, e as plataformas deixam de cumprir até mesmo o piso de transparência que elas próprias haviam estabelecido, desestimulando o avanço coletivo em práticas oficiais de prestação de contas. Um indicativo disso é a divulgação, por parte da Meta, de informações sobre a restauração proativa de conteúdo, sem necessidade de contestação por parte dos usuários, tanto no Facebook quanto no Instagram. Outro ponto é que os dados dos relatórios de transparência de moderação de conteúdo publicados em cumprimento à legislação não são disponibilizados em formato estruturado e legível por máquinas, o que dificulta sua análise por ferramentas externas.

Respondendo, enfim, à **QP5** (*Quais lacunas são identificadas nas exigências de transparência de moderação de conteúdo estabelecidas pelo Digital Services Act?*), argumentamos que o cenário apresentado se deve principalmente à permissão do DSA para que as plataformas mantenham controle significativo tanto sobre o que optam por tornar transparente quanto sobre a forma como essas informações são apresentadas. A regulação define, de forma geral, o que deve ser divulgado, mas não especifica como isso deve ser feito. Como consequência, muitas das informações apresentadas nos relatórios de transparência carecem de real valor informacional, um problema já observado nos relatórios exigidos pelo NetzDG na Alemanha (Heldt, 2019). Há divergências significativas na granularidade dos dados apresentados nestes relatórios, já que cada plataforma define, por conta própria, como

atender aos requisitos mínimos da nova legislação. Logo, as plataformas continuam a controlar o escopo e os termos de suas medidas de transparência, limitando significativamente o escrutínio externo e dificultando sua responsabilização por abusos e falhas na moderação.

Estes problemas poderiam ser facilmente solucionados com a instituição de padrões a serem seguidos obrigatoriamente por todas as plataformas. A ausência de diretrizes técnicas na legislação para a publicação dos relatórios de transparência representa uma limitação significativa do DSA. Essa lacuna tem sido explorada pelas empresas para construir um novo cenário de visibilidade controlada de suas práticas de moderação de conteúdo, que serve mais aos seus interesses privados do que ao interesse público. Com isso, abre-se espaço para níveis excessivos de discricionariedade e arbitrariedade. Afinal, o simples fato de haver regras não garante que elas serão cumpridas conforme o esperado, nem que contribuirão verdadeiramente para aprimorar a governança e a transparência das plataformas. Em uma análise preliminar, Shattock (2021) já havia diagnosticado que, embora representasse um avanço importante, o DSA poderia permanecer atrelado ao modelo ineficaz da autogovernança, justamente o que ele propunha superar, também reforçando preocupações sobre uma mera regulamentação do *transparency washing* (Zalnieriute, 2021). Em última instância, nossa avaliação aponta para a frustração de um dos principais objetivos do DSA: garantir, com base em evidências, a proteção dos usuários frente aos riscos e danos significativos associados ao modelo de negócios das plataformas digitais (European Commission, 2024).

5 CONSIDERAÇÕES FINAIS

A pressão regulatória direcionada às plataformas digitais, especialmente às de redes sociais, tem se concentrado cada vez mais em sua transparência. Apesar de uma retórica pública voltada, nas palavras de seus próprios executivos, à "transparência radical" (Gorwa; Garton Ash, 2020), as grandes plataformas demonstram, dia após dia, que sabemos menos sobre elas do que seria razoável esperar. À revelia do interesse público, sabemos pouco sobre operações movidas por interesses econômicos que afetam diretamente a expressão pública e o comportamento dos usuários – e o que sabemos decorre, em grande parte, de investigações independentes conduzidas em meio à resistência dessas plataformas. A opacidade não é fruto do acaso, mas uma estratégia deliberada e coordenada para evitar pressões externas sobre a forma como essas empresas conduzem seus negócios e para dificultar sua responsabilização por fenômenos extremos que facilitam e hospedam.

O cenário de baixa transparência é particularmente evidente quando nos voltamos à moderação de conteúdo. A moderação de conteúdo constitui um dos principais instrumentos por meio dos quais essas plataformas exercem sua governança, impondo limites ao que os usuários podem dizer ou fazer em seus domínios, em consonância com interesses e disputas predominantemente comerciais. Mais do que os simples atos de remover publicações e suspender usuários, entendemos que a moderação engloba ações de curadoria mais amplas, indissociáveis do que é recomendado – ou não – a seus usuários (Alizadeh *et al.*, 2022; Roberts, 2016; Santini; Salles; Mattos, 2023).

No entanto, mais difícil do que entender o conceito de moderação de conteúdo que as próprias plataformas buscam estabelecer perante o público, em consonância com seu modelo de negócios, é compreender como a moderação se concretiza na prática. As diretrizes e políticas internas que guiam as ações de moderação são guardadas como segredos industriais e os documentos que são liberados ao público, excessivamente genéricos e com poucas explicações. Grande parte dos usuários afetados pela moderação sequer sabe por que foi alvo dessas medidas e dispõe de poucos meios para contestá-las. Diante das pressões para tornar a moderação mais escalável, sistemas algorítmicos, cujos parâmetros de treinamento e diretrizes de funcionamento são envoltos em sigilo, passam a assumir o papel que antes ficava majoritariamente a cargo de pessoas reais, mas sem a mesma possibilidade de responsabilização que elas.

À medida que críticas e escândalos públicos se acumulavam, as plataformas de redes sociais passaram a adotar iniciativas de transparência voluntárias e autogovernadas, prometendo, enfim, conceder ao público, pesquisadores, formuladores de políticas públicas e outros especialistas o acesso a informações há muito demandadas – ao menos, era o que se esperava. No campo da moderação de conteúdo, essas iniciativas de transparência se materializaram sobretudo na publicação dos chamados relatórios de transparência. Todavia, esses documentos são amplamente criticados por carecerem de transparência real e por desviarem a atenção dos aspectos mais problemáticos da moderação (Zalnieriute, 2021).

Esses esforços voluntários de transparência parecem, em certa medida, ter gerado o efeito oposto ao pretendido pelas plataformas: se foram adotados como forma de evitar regulações vinculantes e a introdução de novos marcos normativos, acabaram por se tornar referência para a formulação de estruturas regulatórias orientadas à transparência. Em 2017, o NetzDG, na Alemanha, foi o primeiro projeto de regulação de plataformas digitais que tornou mandatória a publicação de relatórios de transparência de moderação de conteúdo. Cinco anos depois, em linha com a estratégia da União Europeia de coordenar, em nível regional, as crescentes pressões regulatórias que vinham surgindo de forma dispersa entre seus Estados-membros, foi aprovado o DSA. A nova legislação europeia tem um objetivo claro: estabelecer um regime de *diligência* e *transparência* para plataformas digitais, incluindo as de redes sociais, para que elas atuem – e comprovem essa atuação – em prol do bem-estar dos usuários e do interesse público. Entre suas obrigações, também está a publicação de relatórios de transparência de moderação de conteúdo, conforme uma série de exigências específicas.

Neste trabalho, demonstramos que os relatórios de transparência de moderação de conteúdo publicados por quatro grandes plataformas de redes sociais – Facebook, Instagram, YouTube e X/Twitter – em cumprimento ao DSA na União Europeia representam um avanço notável na qualidade das informações divulgadas. Por meio de um quadro analítico original, composto por 60 critérios distribuídos em cinco eixos, verificamos que esses relatórios obrigatórios apresentam desempenho superior em quase todos os aspectos quando comparados aos documentos voluntários que as mesmas plataformas ainda publicam para o restante do mundo.

Inegavelmente, a transparência regulada apresenta diversos benefícios e parece se colocar, neste momento, como a principal e mais tangível alternativa ao regime de opacidade processual que as plataformas de redes sociais mantêm há tantos anos. Contudo, nossa análise evidencia falhas expressivas na forma como essa nova transparência vem sendo implementada. Na prática, o DSA ainda concede ampla margem para a autogovernança das

plataformas, permitindo que elas definam os principais rumos e condições sob os quais essa nova transparência deve ser implementada, confirmando alertas já apontados anteriormente (ver Shattock, 2021). Como a legislação não determina critérios e padrões claros para a publicação destes relatórios, as informações presentes em cada um deles variam significativamente. Os dados são apresentados com granularidades variadas, raramente alcançando um nível satisfatório, o que, em última análise, dificulta a comparação das informações divulgadas, comprometendo um dos principais objetivos dos relatórios de transparência de moderação de conteúdo (Wagner *et al.*, 2020).

Os relatórios exigidos pelo DSA indicam, especialmente, um retrocesso na transparência das ações de moderação feitas em resposta a solicitações de autoridades públicas. Por muito tempo, esses dados foram destacados para reforçar a narrativa de que existiria um aparato estatal de censura contra plataformas e serviços digitais (Urman; Makhortykh, 2023). Atualmente, observa-se um desequilíbrio significativo no valor informacional desses relatórios, porém em sentido oposto ao anteriormente constatado.

Portanto, qualquer projeto de regulação de plataformas digitais e de redes sociais baseado na transparência deve assegurar a definição de padrões e estruturas informacionais claras, que orientem de forma consistente o que deve ser reportado. Isso impediria que as plataformas se limitassem a cumprir apenas o mínimo necessário, aproveitando brechas legais, e garantiria que os resultados regulatórios fossem minimamente comparáveis, viabilizando a fiscalização e a auditabilidade de seus processos e operações. No caso específico analisado neste estudo, as informações de transparência sobre os processos de moderação de conteúdo deveriam ser suficientemente relevantes para permitir a responsabilização das plataformas diante da identificação de vieses ou de riscos a que seus usuários estejam expostos. No entanto, esse objetivo ainda está longe de ser alcançado: observa-se uma transparência formalizada, mas cujas informações carecem de utilidade real.

As limitações deste trabalho derivam do recorte escolhido e de seu escopo. A primeira delas diz respeito ao número reduzido de plataformas analisadas. Estudos anteriores que serviram de base para esta pesquisa investigaram um conjunto maior de plataformas (ver Urman; Makhortykh, 2023), o que lhes permitiu identificar tendências mais abrangentes em relação à (falta de) transparência no ecossistema informacional contemporâneo. Além disso, nossa análise concentrou-se exclusivamente na estrutura e nos critérios formais de apresentação dos relatórios de transparência, sem examinar o conteúdo específico das informações divulgadas. Isso nos impediu de interpretar os sentidos atribuídos pelas próprias plataformas às suas práticas de moderação de conteúdo, restringindo nossa avaliação à

presença ou ausência de determinados dados, e não à forma como esses dados são narrados, enquadrados ou operacionalizados.

Ressaltamos que futuras pesquisas sobre a transparência regulada da moderação de conteúdo devem se concentrar na sua efetiva operacionalização – isto é, no que deve ser tornado transparente e como isso deve ocorrer – e não apenas em proposições teóricas sobre a quem se destina, com que propósito e sob quais limitações. Inevitavelmente, esse processo exige que a própria legislação estabeleça padrões mínimos e bem definidos de transparência, os quais possam ser adotados de forma comum por toda a indústria. Além disso, é fundamental explicitar de maneira clara quais seriam os ganhos concretos e os efeitos esperados da adoção desses padrões, tanto para a sociedade quanto para a integridade dos ecossistemas digitais.

Adicionalmente, é necessário ampliar o debate sobre formas complementares – ou mesmo alternativas – aos relatórios de transparência para reduzir a opacidade dos processos de moderação de conteúdo. O próprio DSA consagra um avanço nesse sentido ao prever, de forma inédita, a criação de uma base de dados online alimentada pelas notificações individuais enviadas aos usuários afetados por ações de moderação (ver Kaushal *et al.*, 2024). A adoção de sistemas semelhantes, centrados na disponibilização de dados mais granulares em vez da simples divulgação de estatísticas gerais e agregadas, pode representar um caminho promissor rumo a uma transparência mais robusta, mas mesmo essas iniciativas dependem da discussão para definição de padrões mínimos e critérios claros que orientem como essa transparência deve ser implementada na prática – e com que finalidade.

O fortalecimento da transparência das plataformas é condição indispensável para a construção de uma "nova esfera pública", como tantos projetos regulatórios ao redor do mundo ambicionam alcançar (Schlesinger, 2020). É fundamental, porém, que a regulação não se limite a burocratizar as frágeis infraestruturas de transparência existentes, mas que promova uma ruptura efetiva rumo a formas mais sólidas de transparência, baseadas em processos e critérios claramente definidos. Com este trabalho, esperamos ter contribuído com bases empíricas para esse debate complexo, indicando caminhos possíveis para superar as deficiências da transparência tanto voluntária quanto regulada.

REFERÊNCIAS

ABBOTT, Kenneth W.; SNIDAL, Duncan. The governance triangle: regulatory standards institutions and the shadow of the state. *In*: MATTLI, Walter; WOODS, Ngaire (Org.). **The politics of global regulation**. Princeton: Princeton University Press, 2009. p. 44–88. Disponível em: https://www.researchgate.net/publication/228677087_The_Governance_Triangle_Regulatory_Standards_Institutions_and_the_Shadow_of_the_State. Acesso em: 15 jan. 2025.

ÅKERLUND, Mathilda. Politics of deliberate inaction: the disconnect between platform justifications and user imaginaries on content moderation in a 'free speech' online forum. **New Media & Society**, [*S. l.*], v. 27, n. 3, p. 1235–1255, ago. 2023. Disponível em: https://doi.org/10.1177/14614448231190905. Acesso em: 5 jun. 2025.

ALIZADEH, Meysam *et al*. Content moderation as a political issue: the Twitter discourse around Trump's ban. **Journal of Quantitative Description: Digital Media**, [*S. l.*], v. 2, out. 2022. Disponível em: https://journalqd.org/article/view/3424. Acesso em: 23 jan. 2025.

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, dez. 2016. Disponível em: https://doi.org/10.1177/1461444816676645. Acesso em: 5 abr. 2025.

ANANNY, Mike; GILLESPIE, Tarleton. Public platforms: beyond the cycle of shocks and exceptions. *In*: INTERNATIONAL COMMUNICATION ASSOCIATION. Annual Conference of the International Communications Association, 67., 2017, San Diego. **Anais** [...]. Washington, D.C.: International Communication Association, 2017. Disponível em: https://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf. Acesso em: 7 jan. 2025.

ANGWIN, Julia; GRASSEGGER, Hannes. Facebook's secret censorship rules protect white men from hate speech but not black children. **ProPublica**, [*S. l.*], 28 jun. 2017. Disponível em: https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms. Acesso em: 30 jan. 2025.

AOS FATOS. Tudo sobre as novas políticas de checagem e moderação da Meta. **Aos Fatos**, [*S. l.*], 14 jan. 2025. Disponível em: https://www.aosfatos.org/noticias/tudo-sobre-novas-politicas-checagem-moderacao-meta/. Acesso em: 28 mar. 2025.

ARE, Carolina. 'Dysfunctional' appeals and failures of algorithmic justice in Instagram and TikTok content moderation. **Information, Communication & Society**, [*S. l.*], 2024. Disponível em: https://doi.org/10.1080/1369118X.2024.2396621. Acesso em: 20 maio 2025.

ARRESE, Ángel. Cultural dimensions of fake news exposure: a cross-national analysis among European Union countries. **Mass Communication and Society**, [*S. l.*], v. 27, n. 5, p. 827–850, out. 2022. Disponível em: https://doi.org/10.1080/15205436.2022.2123278. Acesso em: 4 jul. 2025.

BADOUARD, Romain; BELLON, Anne. Introduction to the special issue on content

moderation on digital platforms. **Internet Policy Review**, [*S. l.*], v. 14, n. 1, 31 mar. 2025. Disponível em: https://policyreview.info/articles/analysis/introduction-content-moderation-digital-platforms. Acesso em: 4 abr. 2025.

BASSAN, Sharon. Transparency ≠ accountability? Rethinking voluntary vs. mandatory content moderation reports. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2025. Disponível em: https://papers.ssrn.com/abstract=5143075. Acesso em: 27 fev. 2025.

BASTOS, Marco *et al.* Reverse influence: the social production of disinformation in the 2022 Brazilian general election. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2025. Disponível em: https://ssrn.com/abstract=5039468. Acesso em: 30 abr. 2025.

BBC. Trump sues Twitter, Google and Facebook alleging "censorship". **BBC**, [*S. l.*], 7 jul. 2021. Disponível em: https://www.bbc.com/news/world-us-canada-57754435. Acesso em: 25 jan. 2025.

BBC NEWS BRASIL. Marco civil da internet: como Estados Unidos e Europa tratam as "big techs". **BBC News Brasil**, [*S. l.*], 13 jun. 2025. Disponível em: https://www.bbc.com/portuguese/articles/c17rvkd2qxgo. Acesso em: 18 jun. 2025.

BECHMANN, Anja. Tackling disinformation and infodemics demands media policy changes. **Digital Journalism**, [*S. l.*], v. 8, n. 6, p. 855–863, jun. 2020. Disponível em: https://doi.org/10.1080/21670811.2020.1773887. Acesso em: 3 mar. 2025.

BEDI, Suneal. The myth of the chilling effect. **Harvard Journal of Law & Technology**, [*S. l.*], v. 35, n. 1, p. 267–307, 2021. Disponível em: https://heinonline.org/HOL/LandingPage?handle=hein.journals/hjlt35&div=8&id=&page=. Acesso em: 26 jul. 2025.

BENKLER, Yochai; FARIS, Robert; ROBERTS, Hal. **Network propaganda**: manipulation, disinformation, and radicalization in American politics. Nova York: Oxford University Press, 2018.

BEZERRA, Arthur Coelho. Vigilância e cultura algorítmica no novo regime global de mediação da informação. **Perspectivas em Ciência da Informação**, [*S. l.*], v. 22, n. 4, p. 68–81, dez. 2017. Disponível em: https://doi.org/10.1590/1981-5344/2936. Acesso em: 5 abr. 2025.

BIZBERGE, Ana; MASTRINI, Guillermo; GÓMEZ, Rodrigo. Discussing internet platform policy and regulation in Latin America. **Journal of Digital Media & Policy**, [*S. l.*], v. 14, n. Emerging debates on internet platform policy and regulation in Latin America, p. 135–148, jun. 2023. Disponível em: https://doi.org/10.1386/jdmp_00118_2. Acesso em: 5 abr. 2025.

BLAKEY, Elizabeth. The day data transparency died: how Twitter/X cut off access for social research. **Contexts**, [*S. l.*], v. 23, n. 2, p. 30–35, mai. 2024. Disponível em: https://doi.org/10.1177/15365042241252125. Acesso em: 7 jul 2025.

BOSSETTA, Michael. The digital architectures of social media: comparing political campaigning on Facebook, Twitter, Instagram, and Snapchat in the 2016 U.S. election. **Journalism & Mass Communication Quarterly**, [*S. l.*], v. 95, n. 2, p. 471–496, mar. 2018.

Disponível em: https://doi.org/10.1177/1077699018763307. Acesso em: 5 abr. 2025.

BOSSETTA, Michael. Scandalous design: how social media platforms' responses to scandal impacts campaigns and elections. **Social Media + Society**, [*S. l.*], v. 6, n. 2, p. 1–4, jun. 2020. Disponível em: https://doi.org/10.1177/2056305120924777. Acesso em: 5 abr. 2025.

BOWEN, Glenn. Document analysis as a qualitative research method. **Qualitative Research Journal**, [*S. l.*], v. 9, n. 2, p. 27–40, ago. 2009. Disponível em: https://doi.org/10.3316/QRJ0902027. Acesso em: 5 abr. 2025.

BRADSHAW, Tim; STACEY, Kiran. Facebook chose to maximise engagement at users' expense, whistleblower says. **Financial Times**, [*S. l.*], 5 out. 2021. Disponível em: https://www.ft.com/content/41b657c8-d716-436b-a06d-19859f0f6ce4. Acesso em: 3 jul. 2025.

BRASIL. **Lei nº 12.965, de 23 de abril de 2014**. Marco Civil da Internet. *Diário Oficial da União*: seção 1, Brasília, DF, 24 abr. 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l12965.htm. Acesso em: 8 abr. 2025.

BREGA, Gabriel R. A regulação de conteúdo nas redes sociais: uma breve análise comparativa entre o NetzDG e a solução brasileira. **Revista Direito GV**, [*S. l.*], v. 19, p. 1–27, 2023. Disponível em: https://doi.org/10.1590/2317-6172202305. Acesso em: 1 jul. 2025.

BROMELL, David. **Regulating free speech in a digital age**: hate, harm and the limits of censorship. Cham: Springer International Publishing, 2022.

BRUNS, Axel. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. **Information, Communication & Society**, [*S. l.*], v. 22, n. 11, p. 1544–1566, jul. 2019. Disponível em: https://doi.org/10.1080/1369118X.2019.1637447. Acesso em: 3 mar. 2025.

BRUNS, Axel. Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. *In*: PÉREZ-ESCOLAR, Marta; NOGUERA-VIVO, José M. (Org.). **Hate speech and polarization in participatory society**. Londres: Routledge, 2021. p. 33–48. Disponível em: https://www.taylorfrancis.com/chapters/oa-edit/10.4324/9781003109891-4/echo-chambers-filter-bubbles-misleading-metaphors-obscure-real-problem-axel-bruns. Acesso em: 15 jan. 2025.

BUENO, Thales M.; CANAAN, Renan G. The Brussels effect in Brazil: analysing the impact of the EU digital services act on the discussion surrounding the fake news bill. **Telecommunications Policy**, [*S. l.*], v. 48, n. 5, p. 1–15, jun. 2024. Disponível em: https://doi.org/10.1016/j.telpol.2024.102757. Acesso em: 5 abr. 2025.

BUNI, Catherine; CHEMALY, Soraya. The secret rules of the internet. **The Verge**, [*S. l.*], 13 abr. 2016. Disponível em: https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech. Acesso em: 20 jan. 2025.

BYERS, Dylan. "This is not normal": behind the decisions at Facebook and Twitter to deplatform Trump. **NBC News**, [*S. l.*], 14 jan. 2021. Disponível em: https://www.nbcnews.com/tech/tech-news/how-facebook-twitter-decided-take-down-trump-s-

accounts-n1254317. Acesso em: 25 jan. 2025.

CALIFANO, Bernadette. Privacy and data protection in Latin America: regulatory initiatives and collisions with the right to freedom of expression on the internet. **Journal of Digital Media & Policy**, [*S. l.*], v. 14, n. Emerging debates on internet platform policy and regulation in Latin America, p. 207–224, jun. 2023. Disponível em: https://doi.org/10.1386/jdmp_00122_1. Acesso em: 5 abr. 2025.

CALVO-GUTIÉRREZ, Elvira; MARÍN-LLADÓ, Carles. Combatting fake news: a global priority post Covid-19. **Societies**, [*S. l.*], v. 13, n. 7, jul. 2023. Disponível em: https://doi.org/10.3390/soc13070160. Acesso em: 22 ago. 2024.

CÂMARA DOS DEPUTADOS. **Projeto de Lei n.º 2630, de 2020**. Brasília, DF: Câmara dos Deputados, 2020. Disponível em: https://www.camara.leg.br/proposicoesWeb/fichadetramitacao?idProposicao=2256735. Acesso em: 28 jan. 2025.

CARAH, Nicholas *et al.* Observing "tuned" advertising on digital platforms. **Internet Policy Review**, [*S. l.*], v. 13, n. 2, jun. 2024. Disponível em: https://doi.org/10.14763/2024.2.1779. Acesso em: 25 set. 2024.

CHANDER, Anupam; KRISHNAMURTHY, Vivek. The myth of platform neutrality. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2018. Disponível em: https://papers.ssrn.com/abstract=4849156. Acesso em: 26 jun. 2025.

CLUNE, Conor; MCDAID, Emma. Content moderation on social media: constructing accountability in the digital space. **Accounting, Auditing & Accountability Journal**, [*S. l.*], v. 37, n. 1, p. 257–279, maio 2023. Disponível em: https://doi.org/10.1108/AAAJ-11-2022-6119. Acesso em: 5 jun. 2025.

COBBE, Jennifer. Algorithmic censorship by social platforms: power and resistance. **Philosophy & Technology**, [*S. l.*], v. 34, n. 4, p. 739–766, dez. 2021. Disponível em: https://doi.org/10.1007/s13347-020-00429-0. Acesso em: 5 abr. 2025.

COLLIER, Kevin; KOLODNY, Lora. Should I delete my DMs? What Twitter has on you, and what you can and can't do about it. **NBC News**, [*S. l.*], 16 nov. 2022. Disponível em: https://www.nbcnews.com/tech/security/twitter-dms-data-delete-rcna57031. Acesso em: 13 mar. 2025.

COMMON, MacKenzie F. Fear the reaper: how content moderation rules are enforced on social media. **International Review of Law, Computers & Technology**, [*S. l.*], v. 34, n. 2, p. 126–152, mar. 2020. Disponível em: https://doi.org/10.1080/13600869.2020.1733762. Acesso em: 5 abr. 2025.

CONSELHO ADMINISTRATIVO DE DEFESA ECONÔMICA. **Mercados de Plataformas Digitais**. Brasília: Conselho Administrativo de Defesa Econômica, ago. 2023. Disponível em: https://cdn.cade.gov.br/Portal/centrais-de-conteudo/publicacoes/estudos-economicos/cadernos-do-cade/Caderno_Plataformas-Digitais_Atualizado_29.08.pdf. Acesso em: 29 jul. 2025.

COSTA, Bruno F. Return to censorship: Portuguese perceptions of digital disinformation regulation. *In*: FILIBELI, Tirşe E.; ÖZBEK, Melis Ö. (Org.). **Mapping lies in the global media sphere**. Londres: Routledge, 2023. p. 148–164. Disponível em:

https://www.taylorfrancis.com/books/9781003403203/chapters/10.4324/9781003403203-13. Acesso em: 12 jan. 2025.

COUNTS, Aisha; NAKANO, Eari. Twitter's surge in harmful content keeps advertiser away. **Time**, [*S. l.*], 19 jul. 2023. Disponível em: https://time.com/6295711/twitters-hate-content-advertisers/. Acesso em: 6 jan. 2025.

CRAMER, Benjamin W. From liability to accountability: the ethics of citing Section 230 to avoid the obligations of running a social media platform. **Journal of Information Policy**, [*S. l.*], v. 10, p. 123–150, maio 2020. Disponível em: https://doi.org/10.5325/jinfopoli.10.2020.0123. Acesso em: 4 abr. 2025.

CRAWFORD, Kate; GILLESPIE, Tarleton. What is a flag for? Social media reporting tools and the vocabulary of complaint. **New Media & Society**, [*S. l.*], v. 18, n. 3, p. 410–428, mar. 2016. Disponível em: https://doi.org/10.1177/1461444814543163. Acesso em: 5 abr. 2025.

CULPEPPER, Pepper D.; THELEN, Kathleen. Are we all Amazon Primed? Consumers and the politics of platform power. **Comparative Political Studies**, [*S. l.*], v. 53, n. 2, p. 288–318, jun. 2019. Disponível em: https://doi.org/10.1177/0010414019852687. Acesso em: 5 abr. 2025.

DANTAS, Marcos. Mais-valia 2.0: produção e apropriação de valor nas redes do capital. **Revista Eletrônica Internacional de Economia Política da Informação da Comunicação e da Cultura**, [*S. l.*], v. 16, n. 2, p. 85–108, maio 2014. Disponível em: https://periodicos.ufs.br/eptic/article/view/2167. Acesso em: 1 jul. 2025.

DANTAS, Marcos; CANAVARRO, Marcela; BARROS, Marina. Trabalho gratuito nas redes: de como o ativismo de 99% pode gerar ainda mais lucros para 1%. **Liinc em Revista**, [*S. l.*], v. 10, n. 1, p. 22–43, maio 2014. Disponível em: https://doi.org/10.18617/liinc.v10i1.696. Acesso em: 5 abr. 2025.

DANTAS, Marcos; RAULINO, Gabriela. Trabalho da audiência e renda informacional no Facebook e YouTube. **Revista Eletrônica Internacional de Economia Política da Informação da Comunicação e da Cultura**, [*S. l.*], v. 22, n. 1, p. 123–141, fev. 2020. Disponível em: https://periodicos.ufs.br/eptic/article/view/12215. Acesso em: 1 jul. 2025.

DE GREGORIO, Giovanni; STREMLAU, Nicole. Platform governance at the periphery: moderation, shutdowns and intervention. *In*: JUDIT, Bayer *et al*. (Org.). **Perspectives on platform regulation**: concepts and models of social media governance across the globe. Baden-Baden: Nomos, 2022. Disponível em: https://dx.doi.org/10.2139/ssrn.4045036. Acesso em: 22 jan. 2025.

DÍAZ, Ángel; HECHT-FELELLA, Laura. Double standards in social media content moderation. **Brennan Center for Justice**, Nova York, 4 ago. 2021. Disponível em: https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation. Acesso em: 25 jan. 2025.

DIAZ RUIZ, Carlos. Disinformation on digital media platforms: a market-shaping approach. **New Media & Society**, [*S. l.*], v. 27, n. 4, p. 2188–2211, out. 2023. Disponível em: https://doi.org/10.1177/14614448231207644. Acesso em: 5 abr. 2025.

DIXON, Stacy Jo. Biggest social media platforms by users 2025. **Statista**, [*S. l.*], 26 mar.

2025. Disponível em: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Acesso em: 7 maio 2025.

DOBBER, Tom *et al.* Shielding citizens? Understanding the impact of political advertisement transparency information. **New Media & Society**, [*S. l.*], v. 26, n. 11, p. 6715–6735, mar. 2023. Disponível em: https://doi.org/10.1177/14614448231157640. Acesso em: 5 abr. 2025.

DOMMETT, Katharine. Regulating digital campaigning: the need for precision in calls for transparency. **Policy & Internet**, [*S. l.*], v. 12, n. 4, p. 432–449, 2020. Disponível em: https://doi.org/10.1002/poi3.234. Acesso em: 1 mar. 2025.

DOUEK, Evelyn. Governing online speech: from "posts-as-Trumps" to proportionality and probability. **Columbia Law Review**, [*S. l.*], v. 121, n. 3, p. 759–834, 2021. Disponível em: https://heinonline.org/HOL/LandingPage?handle=hein.journals/clr121&div=20&id=&page=. Acesso em: 5 abr. 2025.

DWIVEDI, Rekha. **Will the DSA fix it?** A critical analysis of transparency obligations under the Digital Services Act. 2022. 111 f. Dissertação (Mestrado em Direito) – Universidade de Oslo, Oslo, 2022. Disponível em: https://www.duo.uio.no/handle/10852/97974. Acesso em: 10 jun. 2025.

EDELSON, Laura; LAUINGER, Tobias; MCCOY, Damon. A security analysis of the Facebook Ad Library. *In*: IEEE. 2020 IEEE Symposium on Security and Privacy, 2020, San Francisco, CA. **Anais** [...]. New York: IEEE, 2020. p. 661–678. Disponível em: https://ieeexplore.ieee.org/document/9152626/. Acesso em: 29 set. 2023.

EDELSON, Laura *et al.* Universal digital ad transparency. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2021. Disponível em: https://www.ssrn.com/abstract=3898214. Acesso em: 30 abr. 2024.

EIFERT, Martin *et al.* Taming the giants: the DMA/DSA package. **Common Market Law Review**, [*S. l.*], v. 58, n. 4, p. 987–1028, ago. 2021. Disponível em: https://doi.org/10.54648/cola2021065. Acesso em: 26 fev. 2025.

EISENSTAT, Yaël; GILMAN, Nils. The myth of tech exceptionalism. **Noema**, [*S. l.*], 10 fev. 2022. Disponível em: https://www.noemamag.com/the-myth-of-tech-exceptionalism. Acesso em: 24 dez. 2024.

ELSWAH, Mona. **Investigating content moderation systems in the Global South**. Washington, D.C.: Center for Democracy and Technology, 30 jan. 2024. Disponível em: https://cdt.org/insights/investigating-content-moderation-systems-in-the-global-south/. Acesso em: 26 fev. 2025.

EUROPEAN COMMISSION. Questions and answers on the Digital Services Act. **European Commission**, Bruxelas, 22 fev. 2024. Disponível em: https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_2348. Acesso em: 5 maio 2025.

EUROPEAN COMMISSION. How the Digital Services Act enhances transparency online. **European Commission**, [*S. l.*], [*S. d.*]. Disponível em: https://digital-strategy.ec.europa.eu/en/policies/dsa-brings-transparency. Acesso em: 7 mar.

2025.

EUROPEAN PARLIAMENT. **Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)**. Bruxelas: Parlamento Europeu, 2022. Disponível em: http://data.europa.eu/eli/reg/2022/2065/oj/eng. Acesso em: 4 jan. 2025.

FITZGERALD, James; LOKMANOGLU, Ayse D. Special issue: the (international) politics of content takedowns: theory, practice, ethics. **Policy & Internet**, [*S. l.*], v. 15, n. 4, p. 456–465, nov. 2023. Disponível em: https://doi.org/10.1002/poi3.375. Acesso em: 5 abr. 2025.

FLEW, Terry. Platforms on trial. **Intermedia**, [*S. l.*], v. 46, n. 2, p. 24–29, jul. 2018. Disponível em: https://www.iicom.org/wp-content/uploads/im-july2018-platformsontrial-min.pdf. Acesso em: 1 maio 2025.

FLEW, Terry; MARTIN, Fiona; SUZOR, Nicolas. Internet regulation as media policy: rethinking the question of digital communication platform governance. **Journal of Digital Media & Policy**, [*S. l.*], v. 10, n. 1, p. 33–50, mar. 2019. Disponível em: https://doi.org/10.1386/jdmp.10.1.33_1. Acesso em: 5 abr. 2025.

FLEW, Terry; GILLETT, Rosalie. Platform policy: evaluating different responses to the challenges of platform power. **Journal of Digital Media & Policy**, [*S. l.*], v. 12, n. Platform governance: power, diversity and accountability, p. 231–246, jun. 2021. Disponível em: https://doi.org/10.1386/jdmp_00061_1. Acesso em: 5 abr. 2025.

FLEW, Terry. The return of the regulatory state: nation-states as policy actors in digital platform governance. *In*: PADOVANI, Claudia *et al.* (Org.). **Global communication governance at the crossroads**. Cham: Springer International Publishing, 2024. p. 161–178. Disponível em: https://doi.org/10.1007/978-3-031-29616-1_10. Acesso em: 15 jan. 2025.

FORMAN, Jane *et al.* Qualitative research methods: key features and insights gained from use in infection prevention research. **American Journal of Infection Control**, [*S. l.*], v. 36, n. 10, p. 764–771, dez. 2008. Disponível em: https://doi.org/10.1016/j.ajic.2008.03.010. Acesso em: 8 jul. 2025.

FREDERICK, Kara. The infodemic: regulating the new public square. **Observer Research Foundation**, [*S. l.*], 22 abr. 2021. Disponível em: https://www.orfonline.org/expert-speak/infodemic-regulating-new-public-square. Acesso em: 15 jan. 2025.

FREITAS AQUINO, Nick R. Antinomia jurídica entre o Marco Civil da Internet e o Código de Defesa do Consumidor em matéria de responsabilidade civil dos provedores de aplicações de internet. **Jusbrasil**, [*S. l.*], 14 set. 2015. Disponível em: https://www.jusbrasil.com.br/artigos/antinomia-juridica-entre-o-marco-civil-da-internet-e-o-codigo-de-defesa-do-consumidor-em-materia-de-responsabilidade-civil-dos-provedores-de-aplicacoes-de-internet/232516149. Acesso em: 11 jun. 2025.

FREITAS NETTO, Sebastião V. de. *et al.* Concepts and forms of greenwashing: a systematic review. **Environmental Sciences Europe**, [*S. l.*], v. 32, n. 1, p. 1–19, fev. 2020. Disponível em: https://doi.org/10.1186/s12302-020-0300-3. Acesso em: 27 jul. 2025.

FROSIO, Giancarlo. From the e-Commerce Directive to the Digital Services Act. Rochester, **SSRN Scholarly Paper**, NY: Social Science Research Network, ago. 2024. Disponível em: https://papers.ssrn.com/abstract=4914816. Acesso em: 30 abr. 2024.

FUCHS, Christian. **The online advertising tax as the foundation of a public service internet**. Londres: University of Westminster Press, 2018.

G1. Além do X: veja 10 mudanças no Twitter sob o comando de Elon Musk. **G1**, [*S. l.*], 25 jul. 2023. Disponível em: https://g1.globo.com/tecnologia/noticia/2023/07/25/alem-do-x-veja-10-mudancas-no-twitter-sob-o-comando-de-elon-musk.ghtml. Acesso em: 7 dez. 2023.

GARCÍA SILVA, Mateo; CHANDUVI, Maria F. A review of content moderation policies in Latin America. **Tech Policy Press**, [*S. l.*], 8 jul. 2024. Disponível em: https://techpolicy.press/a-review-of-content-moderation-policies-in-latin-america. Acesso em: 24 fev. 2025.

GENG, Yinuo. Transparency for what purpose? Designing outcomes-focused transparency tactics for digital platforms. **Policy & Internet**, [*S. l.*], v. 16, n. 1, p. 83–103, ago. 2023. Disponível em: https://doi.org/10.1002/poi3.362. Acesso em: 29 jun. 2025.

GHEDIN, Rodrigo. DSA: lei contra desinformação e conteúdos danosos começa a valer na Europa. **Núcleo Jornalismo**, [*S. l.*], 25 ago. 2023. Disponível em: https://nucleo.jor.br/raiox/2023-08-25-digital-services-act-dsa-europa-inicio/. Acesso em: 21 jul. 2025.

GHOSH, Dipayan. It's all in the business model: the internet's economic logic and the instigation of disinformation, hate, and discrimination. **Georgetown Journal of International Affairs**, [*S. l.*], v. 21, p. 129–135, 2020. Disponível em: https://doi.org/10.1353/gia.2020.0012. Acesso em: 5 abr. 2025.

GILLESPIE, Tarleton. The politics of 'platforms'. **New Media & Society**, [*S. l.*], v. 12, n. 3, p. 347–364, fev. 2010. Disponível em: https://doi.org/10.1177/1461444809342738. Acesso em: 4 maio 2025.

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

GILLESPIE, Tarleton. Governance of and by platforms. *In*: BURGESS, Jean; MARWICK, Alice; POELL, Thomas (Org.). **The SAGE handbook of social media**. Londres: SAGE, 2018b. p. 254–278. Disponível em: https://sk.sagepub.com/reference/the-sage-handbook-of-social-media/i2081.xml. Acesso em: 3 jan. 2025.

GILLESPIE, Tarleton. Platforms are not intermediaries. **Georgetown Law Technology Review**, [*S. l.*], v. 2, n. 2, p. 198–216, jul. 2018c. Disponível em: https://georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Gilespie-pp-198-216.pdf. Acesso em: 5 abr. 2025.

GILLESPIE, Tarleton. Content moderation, AI, and the question of scale. **Big Data & Society**, [*S. l.*], v. 7, n. 2, p. 1–5, ago. 2020. Disponível em: https://doi.org/10.1177/2053951720943234. Acesso em: 5 abr. 2025.

GILLESPIE, Tarleton *et al.* Expanding the debate about content moderation: scholarly research agendas for the coming policy debates. **Internet Policy Review**, [*S. l.*], v. 9, n. 4, p. 1–29, 2020. Disponível em: https://doi.org/10.14763/2020.4.1512. Acesso em: 5 abr. 2025.

GILLESPIE, Tarleton. Do not recommend? Reduction as a form of content moderation. **Social Media + Society**, [*S. l.*], v. 8, n. 3, ago. 2022. Disponível em: https://doi.org/10.1177/20563051221117552. Acesso em: 22 set. 2022.

GLOBAL WITNESS. How Big Tech platforms are neglecting their non-English language users. **Global Witness**, [*S. l.*], 30 nov. 2023. Disponível em: https://globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/. Acesso em: 24 fev. 2025.

GOMEZ, Juan F. *et al.* Algorithmic arbitrariness in content moderation. *In*: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2024, Rio de Janeiro. **Anais** [...]. Nova York: Association for Computing Machinery, 2024. p. 2234–2253. Disponível em: https://dl.acm.org/doi/10.1145/3630106.3659036. Acesso em: 23 jan. 2025.

GOOGLE. EU DSA transparency report – July 1, 2024 to December 31, 2024. **Google**, [*S. l.*], 28 fev. 2025. Disponível em: https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2024-7-1_2024-12-31_en_v1.pdf. Acesso em: 28 abr. 2025.

GOOGLE. Government removal requests – transparency report. **Google**, [*S. l.*], [*S. d.*]a. Disponível em: https://transparencyreport.google.com/government-removals/overview. Acesso em: 28 abr. 2025.

GOOGLE. YouTube policy removals – transparency report. **Google**, [*S. l.*], [*S. d.*]b. Disponível em: https://transparencyreport.google.com/youtube-policy/removals. Acesso em: 28 abr. 2025.

GOOGLE CLOUD. Definição de Big Data: exemplos e benefícios. **Google Cloud**, [*S. l.*], [*S. d.*]. Disponível em: https://cloud.google.com/learn/what-is-big-data. Acesso em: 13 mar. 2025.

GORWA, Robert. Platform moderation and its discontents. **Los Angeles Review of Books**, [*S. l.*], 10 ago. 2018. Disponível em: https://lareviewofbooks.org/article/platform-moderation-and-its-discontents. Acesso em: 4 fev. 2025.

GORWA, Robert. The platform governance triangle: conceptualising the informal regulation of online content. **Internet Policy Review**, [*S. l.*], v. 8, n. 2, p. 1–22, 2019a. Disponível em: https://doi.org/10.14763/2019.2.1407. Acesso em: 5 abr. 2025.

GORWA, Robert. What is platform governance? **Information, Communication & Society**, [*S. l.*], v. 22, n. 6, p. 854–871, 2019b. Disponível em: https://doi.org/10.1080/1369118X.2019.1573914. Acesso em: 5 abr. 2025.

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, p. 1–15, fev. 2020. Disponível em:

https://doi.org/10.1177/2053951719897945. Acesso em: 5 abr. 2025.

GORWA, Robert; GARTON ASH, Timothy. Democratic transparency in the platform society. *In*: PERSILY, Nathaniel; TUCKER, Joshua A. (Org.). **Social media and democracy**: the state of the field, prospects for reform. Cambridge: Cambridge University Press, 2020. p. 286–312. Disponível em: https://www.cambridge.org/core/books/social-media-and-democracy/democratic-transparency-in-the-platform-society/F4BC23D2109293FB4A8A6196F66D3E41. Acesso em: 3 jan. 2025.

GORWA, Robert. Elections, institutions, and the regulatory politics of platform governance: the case of the German NetzDG. **Telecommunications Policy**, [*S. l.*], v. 45, n. 6, p. 1–14, jul. 2021. Disponível em: https://doi.org/10.1016/j.telpol.2021.102145. Acesso em: 5 abr. 2025.

GORWA, Robert. **The politics of platform regulation**: how governments shape online content moderation. Nova York: Oxford University Press, 2024.

GORWA, Robert; LECHOWSKI, Grzegorz; SCHNEISS, Daniel. Platform lobbying: policy influence strategies and the EU's Digital Services Act. **Internet Policy Review**, [*S. l.*], v. 13, n. 2, p. 1–26, jun. 2024. Disponível em: https://doi.org/10.14763/2024.2.1782. Acesso em: 24 dez. 2024.

GRIMMELMANN, James. The virtues of moderation. **Yale Journal of Law & Technology**, [*S. l.*], v. 17, p. 42–109, abr. 2015. Disponível em: https://scholarship.law.cornell.edu/facpub/1486/. Acesso em: 5 abr. 2025.

GUESS, Andrew M.; LYONS, Benjamin A. Misinformation, disinformation, and online propaganda. *In*: PERSILY, Nathaniel; TUCKER, Joshua A. (Org.). **Social media and democracy**: the state of the field, prospects for reform. Cambridge: Cambridge University Press, 2020. p. 10–33. Disponível em: https://www.cambridge.org/core/books/social-media-and-democracy/misinformation-disinformation-and-online-propaganda/D14406A631AA181839ED896916598500. Acesso em: 3 jan. 2025.

HART, Robert. Elon Musk is restoring banned Twitter accounts—here's why the most controversial users were removed and who's already back. **Forbes**, Londres, 25 nov. 2022. Disponível em: https://www.forbes.com/sites/roberthart/2022/11/25/elon-musk-is-restoring-banned-twitter-accounts-heres-why-the-most-controversial-users-were-suspended-and-whos-already-back/. Acesso em: 22 mar. 2025.

HATEAID. Mixed feelings: Digital Services Act replaces NetzDG. **HateAid**, [*S. l.*], 16 fev. 2024. Disponível em: https://hateaid.org/en/mixed-feelings-digital-services-act-replaces-netzdg/. Acesso em: 24 mar. 2025.

HAWKER, Kiah *et al.* Advertisements on digital platforms: how transparent and observable are they? **Foundation for Alcohol Research and Education**, Queensland, 1 set. 2022. Disponível em: https://fare.org.au/transparency-report/. Acesso em: 29 maio 2025.

HELBERGER, Natali; PIERSON, Jo; POELL, Thomas. Governing online platforms: from contested to cooperative responsibility. **The Information Society**, [*S. l.*], v. 34, n. 1, p. 1–14, dez. 2017. Disponível em: https://doi.org/10.1080/01972243.2017.1391913. Acesso em: 5

jan. 2025.

HELBERGER, Natali. The political power of platforms: how current attempts to regulate misinformation amplify opinion power. **Digital Journalism**, [*S. l.*], v. 8, n. 6, p. 842–854, jul. 2020. Disponível em: https://doi.org/10.1080/21670811.2020.1773888. Acesso em: 3 abr. 2025.

HELBERGER, Natali; SAMUELSON, Pamela. The Digital Services Act as a global transparency regime. **Verfassungsblog**, [*S. l.*], 7 mar. 2024. Disponível em: https://verfassungsblog.de/the-digital-services-act-as-a-global-transparency-regime/. Acesso em: 7 jan. 2025.

HELDT, Amélie. Reading between the lines and the numbers: an analysis of the first NetzDG reports. **Internet Policy Review**, [*S. l.*], v. 8, n. 2, p. 1–18, jun. 2019. Disponível em: https://doi.org/10.14763/2019.2.1398. Acesso em: 10 abr. 2025.

HELMOND, Anne. The platformization of the web: making web data platform ready. **Social Media + Society**, [*S. l.*], v. 1, n. 2, p. 1–11, jul. 2015. Disponível em: https://doi.org/10.1177/2056305115603080. Acesso em: 5 abr. 2025.

HENDRIX, Justin. Transcript: Mark Zuckerberg announces major changes to Meta's content moderation policies and operations. **Tech Policy Press**, [*S. l.*], 7 jan. 2025. Disponível em: https://techpolicy.press/transcript-mark-zuckerberg-announces-major-changes-to-metas-content-moderation-policies-and-operations. Acesso em: 22 jan. 2025.

HERRMAN, John. Why nothing on your phone is safe from ads. **New York Magazine**, [*S. l.*], 21 ago. 2023. Disponível em: https://nymag.com/intelligencer/2023/08/why-every-tech-company-turns-into-an-ad-company.html. Acesso em: 27 dez. 2023.

HERN, Alex. Trump's vote fraud claims go viral on social media despite curbs. **The Guardian**, [*S. l.*], 10 nov. 2020. Disponível em: https://www.theguardian.com/us-news/2020/nov/10/trumps-vote-claims-go-viral-on-social-media-despite-curbs. Acesso em: 25 jan. 2025.

HILL, Stephanie; SHTERN, Jeremy. Techlash, platformization and the struggle to govern online content. *In*: PADOVANI, Claudia *et al.* (Org.). **Global communication governance at the crossroads**. Cham: Springer International Publishing, 2024. p. 315–332. Disponível em: https://doi.org/10.1007/978-3-031-29616-1_18. Acesso em: 21 fev. 2025.

HOVYADINOV, Serhiy. Toward a more meaningful transparency: examining Twitter, Google, and Facebook's transparency reporting and removal practices in Russia. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2019. Disponível em: https://papers.ssrn.com/abstract=3535671. Acesso em: 15 abr. 2025.

HUFFPOST. Google CEO on privacy: "If you have something you don't want anyone to know, maybe you shouldn't be doing it". **HuffPost**, [*S. l.*], 18 mar. 2010. Disponível em: https://www.huffpost.com/entry/google-ceo-on-privacy-if_n_383105. Acesso em: 14 mar. 2025.

JALLI, Nuurrianti. Holding social media companies accountable for enabling hate and disinformation. **Open Research Oklahoma**, [*S. l.*], 11 jul. 2024. Disponível em:

https://openresearch.okstate.edu/entities/publication/d2d28779-eed0-4120-abbb-c7315ce20710. Acesso em: 20 jan. 2025.

JAMISON, Amelia M. *et al.* Vaccine-related advertising in the Facebook Ad Archive. **Vaccine**, [*S. l.*], v. 38, n. 3, p. 512–520, jan. 2020. Disponível em: https://doi.org/10.1016/j.vaccine.2019.10.066. Acesso em: 5 abr. 2025.

JIANG, Sujia; FANG, Wei. Misinformation and disinformation in science: examining the social diffusion of rumours about GMOs. **Cultures of Science**, [*S. l.*], v. 2, n. 4, p. 327–340, dez. 2019. Disponível em: https://doi.org/10.1177/209660831900200407. Acesso em: 5 jan. 2025.

JOSEPH, Seb; SCANLON, Krystal. X brings back its transparency report for the first time since 2021. **Digiday**, [*S. l.*], 25 set. 2024. Disponível em: https://digiday.com/marketing/x-brings-back-its-transparency-report-for-the-first-time-since-2021/. Acesso em: 12 maio 2025.

KAUSHAL, Rishabh *et al.* Automated transparency: a legal and empirical analysis of the Digital Services Act Transparency Database. In: ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2024, Rio de Janeiro. **Anais** [...]. Nova York: Association for Computing Machinery, 2024. p. 1121–1132. Disponível em: https://doi.org/10.1145/3630106.3658960. Acesso em: 3 jan. 2025.

KAYYALI, Dia. An advocate's guide to automated content moderation. **Tech Policy Press**, [*S. l.*], 12 fev. 2025. Disponível em: https://techpolicy.press/an-advocates-guide-to-automated-content-moderation. Acesso em: 24 fev. 2025.

KEMP, Simon. Digital 2025: Brazil. **DataReportal**, [*S. l.*], 3 mar. 2025. Disponível em: https://datareportal.com/reports/digital-2025-brazil. Acesso em: 7 mar. 2025.

KIKERPILL, Kristjan; SIIBAK, Andra. Abusing the Covid-19 pan(dem)ic: a perfect storm for online scam. *In*: POLLOCK, John C.; VAKOCH, Douglas A. (Org.). **Covid-19 in international media**: global pandemic perspectives. Abingdon: Routledge, 2021. p. 249–258. Disponível em: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003181705-25/abusing-covid-19-pan-dem-ic-kristjan-kikerpill-andra-siibak. Acesso em: 4 maio 2025.

KIRKPATRICK, David. **The Facebook effect**: the real inside story of Mark Zuckerberg and the world's fastest-growing company. Londres: Virgin Books, 2011.

KLONICK, Kate. The new governors: the people, rules, and processes governing online speech. **Harvard Law Review**, [*S. l.*], v. 131, p. 1598–1670, mar. 2017. Disponível em: https://ssrn.com/abstract=2937985. Acesso em: 5 abr. 2025.

KNOLL, Johannes; PROKSCH, Ramona. Why we watch others' responses to online advertising – investigating users' motivations for viewing user-generated content in the context of online advertising. **Journal of Marketing Communications**, [*S. l.*], v. 23, n. 4, p. 400–412, jun. 2015. Disponível em: https://doi.org/10.1080/13527266.2015.1051092. Acesso em: 5 abr. 2025.

KOMAITIS, Konstantinos; CARTER, Jordan. Internet governance and digital governance:

are they the same thing? **Tech Policy Press**, [*S. l.*], 13 set. 2023. Disponível em: https://www.techpolicy.press/internet-governance-and-digital-governance-are-they-the-same-thing/. Acesso em: 24 fev. 2025.

KOSTA, Eleni; BREWCZYŃSKA, Magdalena. Government access to user data: towards more meaningful transparency reports. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2020. Disponível em: https://ssrn.com/abstract=3601661. Acesso em: 15 abr. 2025.

KULKARNI, Arti. Introducing the Ad Archive Report: a closer look at political and issue ads. **Meta Newsroom**, [*S. l.*], 23 out. 2018. Disponível em: https://about.fb.com/news/2018/10/ad-archive-report/. Acesso em: 26 dez. 2023.

LANGVARDT, Kyle. Regulating online content moderation. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, ago. 2017. Disponível em: https://papers.ssrn.com/abstract=3024739. Acesso em: 30 abr. 2024.

LAUER, David. Facebook's ethical failures are not accidental; they are part of the business model. **AI and Ethics**, [*S. l.*], v. 1, n. 4, p. 395–403, jun. 2021. Disponível em: https://doi.org/10.1007/s43681-021-00068-x. Acesso em: 5 abr. 2025.

LEAL, Hugo; FELSBERGER, Stefanie; NEFF, Gina. Harmful by design: current approaches to misinformation and how to improve harm mitigation. **Minderoo Centre for Technology & Democracy**, Cambridge, 18 dez. 2024. Disponível em: https://www.mctd.ac.uk/wp-content/uploads/2025/01/Written-Evidence-Harmful-by-Design-current-approaches-to-misinformation-and-how-to-improve-harm-mitigation.pdf. Acesso em: 26 jun. 2025.

LEE, Ricki. Here today, Elon tomorrow: are advertisers abandoning X? **TechInformed**, [*S. l.*], 23 ago. 2024. Disponível em: https://techinformed.com/why-advertisers-are-boycotting-x-elon-musk-impact-2024/. Acesso em: 6 jan. 2025.

LEERSSEN, Paddy *et al.* Platform ad archives: promises and pitfalls. **Internet Policy Review**, [*S. l.*], v. 8, n. 4, 2019. Disponível em: https://doi.org/10.14763/2019.4.1421. Acesso em: 5 abr. 2023.

LEERSSEN, Paddy. Outside the black box: from algorithmic transparency to platform observability in the Digital Services Act. **Weizenbaum Journal of the Digital Society**, [*S. l.*], v. 4, n. 2, 2024. Disponível em: https://doi.org/10.34669/wi.wjds/4.2.3. Acesso em: 6 maio 2025.

LEONE DE CASTRIS, Arcangelo. Types of platform transparency: an analysis of discourse around transparency and global digital platforms. **Public Integrity**, [*S. l.*], v. 27, n. 3, p. 340–354, fev. 2024. Disponível em: https://doi.org/10.1080/10999922.2024.2304741. Acesso em: 3 jan. 2025.

LEWIS, Rebecca; MARWICK, Alice. Media manipulation and disinformation online. **Data & Society**, [*S. l.*], 15 maio 2017. Disponível em: https://datasociety.net/library/media-manipulation-and-disinfo-online/. Acesso em: 17 jun. 2025.

LIMA-STRONG, Cristiano. Musk decries government ‘censorship.’ His X has been more compliant. **Washington Post**, [*S. l.*], 25 set. 2024. Disponível em: https://www.washingtonpost.com/technology/2024/09/25/elon-musk-x-twitter-free-speech-government-requests/. Acesso em: 7 maio 2025.

LLANSÓ, Emma *et al.* Artificial intelligence, content moderation, and freedom of expression. **University of Amsterdam**, Amsterdã, 26 fev. 2020. Disponível em: https://dare.uva.nl/search?identifier=34ef500a-67d4-4e27-afd1-26e8f4c6b80e. Acesso em: 7 jan. 2025.

LU, Sylvia. Tech law in 2025: a look ahead at AI, privacy and social media regulation under the new Trump administration. **The Conversation**, [*S. l.*], 3 jan. 2025. Disponível em: http://theconversation.com/tech-law-in-2025-a-look-ahead-at-ai-privacy-and-social-media-regulation-under-the-new-trump-administration-245425. Acesso em: 1 jul. 2025.

MAASS, Sabrina; WORTELKER, Jil; ROTT, Armin. Evaluating the regulation of social media: an empirical study of the German NetzDG and Facebook. **Telecommunications Policy**, [*S. l.*], v. 48, n. 5, jun. 2024. Disponível em: https://doi.org/10.1016/j.telpol.2024.102719. Acesso em: 27 jul. 2025.

MACCARTHY, Mark. Transparency requirements for digital social media platforms: recommendations for policy makers and industry. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2020. Disponível em: https://www.ssrn.com/abstract=3615726. Acesso em: 30 abr. 2024.

MASSANARI, Adrienne. #Gamergate and the fappening: how Reddit’s algorithm, governance, and culture support toxic technocultures. **New Media & Society**, [*S. l.*], v. 19, n. 3, p. 329–346, out. 2015. Disponível em: https://doi.org/10.1177/1461444815608807. Acesso em: 4 mar. 2025.

MCKEOWN, Caitlin. Facebook, defamation, and terrorism: who is responsible for dangerous posts on social media. **Tulane Journal of International and Comparative Law**, [*S. l.*], v. 26, n. 1, p. 163–187, 2017. Disponível em: https://journals.tulane.edu/jicl/article/view/3110. Acesso em: 5 maio 2025.

MEISNER, Colten. The weaponization of platform governance: mass reporting and algorithmic punishments in the creator economy. **Policy & Internet**, [*S. l.*], v. 15, n. 4, p. 466–477, 2023. Disponível em: https://doi.org/10.1002/poi3.359. Acesso em: 5 abr. 2025.

META. DSA transparency report – April - September 2024: Facebook. **Meta Transparency Center**, [*S. l.*], 25 out. 2024a. Disponível em: https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebook. Acesso em: 28 abr. 2025.

META. DSA transparency report – April - September 2024: Instagram. **Meta Transparency Center**, [*S. l.*], 25 out. 2024b. Disponível em: https://scontent.fsdu8-2.fna.fbcdn.net/v/t39.8562-6/466943155_1291701138400105_7867447844898917200_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=ApoOoCEllfAQ7kNvwFRQnAn&_nc_oc=Adm81Piz8hRSBE9lKzNsjbzyj7uOa-QNz_DKGX1JmlYSsZfuq_D_YJ-B69QCMNvSQ_KK5i1t8NUDv2N-ILhawLah&_nc_zt=14&_nc_ht=scontent.fsdu8-2.fna&_nc_gid=cVUCVFcM3VMSnrWeQ7cX3Q&oh=00_AfQF03TjrSKEtruB1B2WQcsOfs3S

LtlbN5HBMdu6nIMAaQ&oe=688EDADD. Acesso em: 28 abr. 2025.

META. Content restrictions based on local law. **Meta Transparency Center**, [*S. l.*], [*S. d.*]. Disponível em: https://transparency.meta.com/reports/content-restrictions/. Acesso em: 28 abr. 2025.

MILLER, Gabby. First transparency reports under Digital Services Act are difficult to compare. **Tech Policy Press**, [*S. l.*], 22 nov. 2023. Disponível em: https://techpolicy.press/first-transparency-reports-under-digital-services-act-are-difficult-to-compare. Acesso em: 12 mar. 2025.

MORAN, Rachel Elizabeth *et al.* The end of trust and safety? Examining the future of content moderation and upheavals in professional online safety efforts. *In:* CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2025, Yokohama. **Anais** [...]. Nova York: Association for Computing Machinery, 2025. p. 1–14. Disponível em: https://dl.acm.org/doi/10.1145/3706598.3713662. Acesso em: 4 maio 2025.

MORGAN, Hani. Conducting a qualitative document analysis. **The Qualitative Report**, [*S. l.*], v. 27, n. 1, p. 64–77, 2022. Disponível em: https://doi.org/10.46743/2160-3715/2022.5044. Acesso em: 3 jun. 2025.

MOROZOV, Evgeny. **The net delusion**: the dark side of internet freedom. Nova York: PublicAffairs, 2012.

MORRISON, Sara. Section 230, the internet law that's under threat, explained. **Vox**, [*S. l.*], 23 fev. 2023. Disponível em: https://www.vox.com/recode/2020/5/28/21273241/section-230-explained-supreme-court-social-media. Acesso em: 9 jan. 2025.

MOZUR, Paul. A genocide incited on Facebook, with posts from Myanmar's military. **The New York Times**, Nepiedó, 15 out. 2018. Disponível em: https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html. Acesso em: 12 jan. 2025.

MYERS WEST, Sarah. Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. **New Media & Society**, [*S. l.*], v. 20, n. 11, p. 4366–4383, mai. 2018. Disponível em: https://doi.org/10.1177/1461444818773059. Acesso em: 5 abr. 2025.

NAHON, Karine. Where there is social media there is politics. *In*: BRUNS, Axel *et al.* (Org.). **The Routledge companion to social media and politics**. Nova York: Routledge, 2015. p. 39–55. Disponível em: https://www.taylorfrancis.com/chapters/edit/10.4324/9781315716299-5/social-media-politics-karine-nahon. Acesso em: 22 jan. 2025.

NAPOLI, Philip; CAPLAN, Robyn. Why media companies insist they're not media companies, why they're wrong, and why it matters. **First Monday**, [*S. l.*], v. 22, n. 5, maio 2017. Disponível em: http://journals.uic.edu/ojs/index.php/fm/article/view/7051. Acesso em: 27 fev. 2023.

NAPOLI, Philip. **Social media and the public interest**: media regulation in the disinformation age. Nova York: Columbia University Press, 2019.

NEITSCH, Joana. Barroso: Vamos ter que julgar aqui no STF, se Congresso não regular plataformas digitais. **Jota**, [*S. l.*], 13 jun. 2022. Disponível em: https://www.jota.info/stf/do-supremo/barroso-vamos-ter-que-julgar-aqui-no-stf-se-congresso-nao-regular-plataforma-digitais. Acesso em: 1 ago. 2025.

NEMER, David. Responsabilização das plataformas digitais sobre conteúdo que publicam no Brasil está nas mãos do STF. **The Conversation**, [*S. l.*], 18 jun. 2025. Disponível em: http://theconversation.com/responsabilizacao-das-plataformas-digitais-sobre-conteudo-que-publicam-no-brasil-esta-nas-maos-do-stf-259247. Acesso em: 18 jun. 2025.

NEWTON, Casey. The secret lives of Facebook moderators in America. **The Verge**, [*S. l.*], 25 fev. 2019. Disponível em: https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona. Acesso em: 21 mar. 2025.

NICHOLAS, Gabriel; BHATIA, Aliya. Toward better automated content moderation in low-resource languages. **Journal of Online Trust and Safety**, [*S. l.*], v. 2, n. 1, p. 1–11, set. 2023. Disponível em: https://doi.org/10.54501/jots.v2i1.150. Acesso em: 23 fev. 2025.

O'BRIEN, Matt; ORTUTAY, Barbara. Musk's Twitter disbands its trust and safety advisory group. **AP News**, [*S. l.*], 13 dez. 2022. Disponível em: https://apnews.com/article/elon-musk-twitter-inc-technology-business-a9b795e8050de12319b82b5dd7118cd7. Acesso em: 6 jan. 2025.

O'REILLY, Lara. Advertisers say Meta's content-moderation changes make them uneasy. They won't stop spending. **Business Insider**, [*S. l.*], 8 jan. 2025. Disponível em: https://www.businessinsider.com/meta-content-moderation-update-impact-ad-business-brand-safety-2025-1. Acesso em: 24 mar. 2025.

OBAR, Jonathan A. Sunlight alone is not a disinfectant: consent and the futility of opening Big Data black boxes (without assistance). **Big Data & Society**, [*S. l.*], v. 7, n. 1, p. 1–5, jun. 2020. Disponível em: https://doi.org/10.1177/2053951720935615. Acesso em: 5 abr. 2025.

PAPAEVANGELOU, Charis; VOTTA, Fabio. Content moderation and platform observability in the Digital Services Act. **Tech Policy Press**, [*S. l.*], 29 maio 2024. Disponível em: https://techpolicy.press/content-moderation-and-platform-observability-in-the-digital-services-act. Acesso em: 4 jan. 2025.

PAPAKYRIAKOPOULOS, Orestis *et al.* Social media and microtargeting: political data processing and the consequences for Germany. **Big Data & Society**, [*S. l.*], v. 5, n. 2, p. 1–15, nov. 2018. Disponível em: https://doi.org/10.1177/2053951718811844. Acesso em: 5 abr. 2025.

PASQUALE, Frank. **The black box society**: the secret algorithms that control money and information. Cambridge; Londres: Harvard University Press, 2016.

PEDERSEN, Morten Axel; ALBRIS, Kristoffer; SEAVER, Nick. The political economy of attention. **Annual Review of Anthropology**, [*S. l.*], v. 50, n. 1, p. 309–325, out. 2021. Disponível em: https://doi.org/10.1146/annurev-anthro-101819-110356. Acesso em: 5 abr. 2025.

PERSILY, Nathaniel; TUCKER, Joshua A. (Org.). **Social media and democracy**: the state of

the field, prospects for reform. Cambridge: Cambridge University Press, 2020.

POELL, Thomas; NIEBORG, David; VAN DIJCK, José. Platform power & public value. **AoIR Selected Papers of Internet Research**, [*S. l.*], out. 2018. Disponível em: https://spir.aoir.org/ojs/index.php/spir/article/view/10501. Acesso em: 12 jan. 2025.

POELL, Thomas; NIEBORG, David; VAN DIJCK, José. Platformisation. **Internet Policy Review**, [*S. l.*], v. 8, n. 4, p. 1–4, nov. 2019. Disponível em: https://policyreview.info/node/1425. Acesso em: 25 maio 2022.

POPIEL, Pawel. The tech lobby: tracing the contours of new media elite lobbying power. **Communication, Culture and Critique**, [*S. l.*], v. 11, n. 4, p. 566–585, out. 2018. Disponível em: https://doi.org/10.1093/ccc/tcy027. Acesso em: 5 abr. 2025.

POPIEL, Pawel. Digital platforms as policy actors. *In*: FLEW, Terry; MARTIN, Fiona (Org.). **Digital platform regulation**: global perspectives on internet governance. Cham: Springer International Publishing, 2022. p. 131–150. Disponível em: https://link.springer.com/10.1007/978-3-030-95220-4_7. Acesso em: 9 jan. 2025.

POPIEL, Pawel; VASUDEVAN, Krishnan. Platform frictions, platform power, and the politics of platformization. **Information, Communication & Society**, [*S. l.*], v. 27, n. 10, p. 1867–1883, jun. 2024. Disponível em: https://doi.org/10.1080/1369118X.2024.2361095. Acesso em: 5 abr. 2025.

RADSCH, Courtney. Transparency reporting: good practices and lessons from global assessment frameworks. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2022. Disponível em: https://dx.doi.org/10.2139/ssrn.4416400. Acesso em: 28 abr. 2025.

RECUERO, Raquel. Mídia social, plataforma digital, site de rede social ou rede social? Não é tudo a mesma coisa? **Medium**, [*S. l.*], 9 jul. 2019. Disponível em: https://medium.com/@raquelrecuero/m%C3%ADdia-social-plataforma-digital-site-de-rede-social-ou-rede-social-n%C3%A3o-%C3%A9-tudo-a-mesma-coisa-d7b54591a9ec. Acesso em: 2 jul. 2025.

RECUERO, Raquel; SOARES, Felipe B.; GRUZD, Anatoliy. Hyperpartisanship, disinformation and political conversations on Twitter: the Brazilian presidential election of 2018. **Proceedings of the International AAAI Conference on Web and Social Media**, [*S. l.*], v. 14, n. 1, p. 569–578, maio 2020. Disponível em: https://doi.org/10.1609/icwsm.v14i1.7324. Acesso em: 5 maio 2025.

REUTERS. Musk says “possible” that Twitter gave preference to leftists during Brazil election. **Reuters**, São Paulo, 3 dez. 2022. Disponível em: https://www.reuters.com/technology/musk-says-possible-that-twitter-gave-preference-leftists-during-brazil-election-2022-12-03/. Acesso em: 22 mar. 2025.

REVIGLIO, Urbano; AGOSTI, Claudio. Thinking outside the black-box: the case for “algorithmic sovereignty” in social media. **Social Media + Society**, [*S. l.*], v. 6, n. 2, p. 1–12, abr. 2020. Disponível em: https://doi.org/10.1177/2056305120915613. Acesso em: 5 maio 2025.

RIBEIRO, Filipe N. *et al.* On microtargeting socially divisive ads: a case study of

Russia-linked ad campaigns on Facebook. *In*: FAT '19: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2019, Atlanta. **Anais** [...]. Nova York: Association for Computing Machinery, jan. 2019. p. 140–149. Disponível em: https://dl.acm.org/doi/10.1145/3287560.3287580. Acesso em: 2 jan. 2024.

RIEDER, Bernhard; HOFMANN, Jeanette. Towards platform observability. **Internet Policy Review**, [*S. l.*], v. 9, n. 4, dez. 2020. Disponível em: https://policyreview.info/articles/analysis/towards-platform-observability. Acesso em: 29 set. 2023.

ROBERTS, Sarah T. Commercial content moderation: digital laborers' dirty work. *In*: NOBLE, Safiya Umoja; TYNES, Brendesha S. (Org.). **The intersectional internet**: race, sex, class and culture online. 1. ed. Nova York: Peter Lang Publishing, Inc, 2016. p. 147–160. Disponível em: https://www.researchgate.net/publication/283421772_Commercial_Content_Moderation_Digital_Laborers'_Dirty_Work. Acesso em: 29 jul. 2025.

ROBERTS, Sarah T. Digital detritus: "error" and the logic of opacity in social media content moderation. **First Monday**, [*S. l.*], v. 23, n. 3, mar. 2018. Disponível em: https://firstmonday.org/ojs/index.php/fm/article/view/8283. Acesso em: 20 jan. 2025.

ROBERTS, Sarah T. **Behind the screen**: content moderation in the shadows of social media. New Haven: Yale University Press, 2019.

ROBERTSON, Adi. What Elon Musk's Twitter 'free speech' promises miss. **The Verge**, [*S. l.*], 15 abr. 2022. Disponível em: https://www.theverge.com/2022/4/15/23025120/elon-musk-twitter-free-speech-government-censorship. Acesso em: 22 mar. 2025.

ROBISON, Kylie. Elon Musk's open source code is 'completely dishonest'. **Fortune**, [*S. l.*], 2 abr. 2023. Disponível em: https://fortune.com/2023/04/02/elon-musk-open-source-twitter-algorithm-gets-cheers-jeers-confusion/. Acesso em: 7 abr. 2025.

SALLES, Débora *et al.* Unfair play: digital platforms' abuse of power to influence Brazilian policy agenda. **AoIR Selected Papers of Internet Research**, [*S. l.*], p. 1–5, 2024. Disponível em: https://spir.aoir.org/ojs/index.php/spir/article/view/14053. Acesso em: 24 fev. 2025.

SAN MARTIN, Pamela. Meta's Oversight Board: challenges of content moderation on the internet. **Erasmus Law Review**, [*S. l.*], v. 16, p. 124–137, 2023. Disponível em: http://dx.doi.org/10.5553/ELR.000253. Acesso em: 15 abr. 2025.

SANTA CLARA PRINCIPLES ON TRANSPARENCY AND ACCOUNTABILITY IN CONTENT MODERATION. **Santa Clara Principles**, [*S. l.*], dez. 2021. Disponível em: https://santaclaraprinciples.org. Acesso em: 7 jan. 2025.

SANTINI, R. Marie *et al.* Software power as soft power: a literature review on computational propaganda effects in public opinion and political process. **Partecipazione e Conflitto**, [*S. l.*], v. 11, n. 2, 2018. Disponível em: http://siba-ese.unisalento.it/index.php/paco/article/view/19546. Acesso em: 20 jun. 2022.

SANTINI, R. Marie. **O algoritmo do gosto**: tecnologias de controle, contágio e curadoria de

si. Curitiba: Appris Editora, 2020.

SANTINI, R. Marie *et al.* Do you believe in fake after all? WhatsApp disinformation campaign during the Brazilian 2018 presidential election. *In*: LÓPEZ-GARCÍA, Guillermo *et al.* (Org.). **Politics of disinformation**: the influence of fake news on the public sphere. Hoboken: Wiley, 2021. p. 49–66. Disponível em: https://onlinelibrary.wiley.com/doi/10.1002/9781119743347.ch4. Acesso em: 23 jan. 2024.

SANTINI, R. Marie; SALLES, Débora; MATTOS, Bruno. Recommending instead of taking down: YouTube hyperpartisan content promotion amid the Brazilian general elections. **Policy & Internet**, [*S. l.*], v. 15, n. 4, p. 512–527, 2023. Disponível em: https://doi.org/10.1002/poi3.380. Acesso em: 5 abr. 2025.

SANTINI, R. Marie; SALLES, Débora; BELIN, Luciane; *et al.* "Aprenda a evitar 'esse tipo' de mulher": estratégias discursivas e monetização da misoginia no YouTube. **NetLab UFRJ**, Rio de Janeiro, 13 dez. 2024. Disponível em: https://netlab.eco.ufrj.br/post/aprenda-a-evitar-esse-tipo-de-mulher-estratégias-discursivas-e-monetização-da-misoginia-no-yout. Acesso em: 26 jun. 2025.

SANTINI, R. Marie; SALLES, Débora; MATTOS, Bruno; CANAVARRO, Marcela; BARROS, Carlos E.; MOREIRA, Alékis; GRAEL, Felipe; *et al.* Índice de transparência da publicidade nas plataformas de redes sociais. **NetLab UFRJ**, Rio de Janeiro, 2024. Disponível em: https://netlab.eco.ufrj.br/itp. Acesso em: 25 jan. 2025.

SANTINI, R. Marie; SALLES, Débora; MATTOS, Bruno; CANAVARRO, Marcela; BARROS, Carlos E.; MOREIRA, Alékis; MEDEIROS, Priscila; *et al.* Índice de transparência de dados das plataformas de redes sociais. **NetLab UFRJ**, Rio de Janeiro, 2024. Disponível em: https://netlab.eco.ufrj.br/itd. Acesso em: 25 jan. 2025.

SANTINI, R. Marie *et al.* (Org.). **Atingidos pelas redes sociais**: os impactos da indústria da desinformação nos consumidores brasileiros. Porto Alegre: Editora Sulina, 2025.

SCHLESINGER, Philip. After the post-public sphere. **Media, Culture & Society**, [*S. l.*], v. 42, n. 7–8, p. 1545–1563, ago. 2020. Disponível em: https://doi.org/10.1177/0163443720948003. Acesso em: 17 abr. 2025.

SCHREIBER, Anderson. Civil Rights Framework of the Internet (BCRFI; Marco Civil da Internet): advance or setback? Civil liability for damage derived from content generated by third party. *In*: ALBERS, Marion; SARLET, Ingo Wolfgang (Org.). **Personality and data protection rights on the internet**: Brazilian and German approaches. Cham: Springer International Publishing, 2022. p. 241–266. Disponível em: https://doi.org/10.1007/978-3-030-90331-2_10. Acesso em: 28 jan. 2025.

SCHWEMER, Sebastian Felix. Digital Services Act: a reform of the e-Commerce Directive and much more. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 10 out. 2022. Disponível em: https://papers.ssrn.com/abstract=4213014. Acesso em: 29 maio 2025.

SHAHID, Farhana. Colonialism in content moderation research: the struggles of scholars in the majority world. **Center for Democracy and Technology**, Washington, D.C., 26 ago. 2024. Disponível em: https://cdt.org/insights/colonialism-in-content-moderation-research-the-struggles-of-scholars-i

n-the-majority-world/. Acesso em: 26 fev. 2025.

SHATTOCK, Ethan. Self-regulation 2:0? A critical reflection of the European fight against disinformation. **Harvard Kennedy School Misinformation Review**, [*S. l.*], v. 2, n. 3, p. 1–8, maio 2021. Disponível em: https://doi.org/10.37016/mr-2020-73. Acesso em: 1 jun. 2025.

SINGH, Nandini. Advertisers abandon Elon Musk's X amid concerns over content and trust. **Business Standard**, Nova Déli, 5 set. 2024. Disponível em: https://www.business-standard.com/world-news/advertisers-abandon-elon-musk-s-x-amid-concerns-over-content-and-trust-124090500163_1.html. Acesso em: 6 jan. 2025.

SOLON, Olivia. To censor or sanction extreme content? Either way, Facebook can't win. **The Guardian**, San Francisco, 23 maio 2017. Disponível em: https://www.theguardian.com/news/2017/may/22/facebook-moderator-guidelines-extreme-content-analysis. Acesso em: 16 jun. 2025.

STINSON, Catherine. Algorithms are not neutral. **AI and Ethics**, [*S. l.*], v. 2, n. 4, p. 763–770, jan. 2022. Disponível em: https://doi.org/10.1007/s43681-022-00136-w. Acesso em: 16 jan. 2025.

STOCKMANN, Daniela. Tech companies and the public interest: the role of the state in governing social media platforms. **Information, Communication & Society**, [*S. l.*], v. 26, n. 1, p. 1–15, fev. 2022. Disponível em: https://doi.org/10.1080/1369118X.2022.2032796. Acesso em: 5 abr. 2025.

STROPPA, Tatiana *et al.* A seção 230 do CDA e o artigo 19 do Marco Civil da Internet. **Consultor Jurídico**, [*S. l.*], 4 maio 2022. Disponível em: https://www.conjur.com.br/2022-mai-04/direito-digital-secao-230-cda-artigo-19-marco-civil-internet/. Acesso em: 8 jan. 2025.

STROWEL, Alain; DE MEYERE, Jean. The Digital Services Act: transparency as an efficient tool to curb the spread of disinformation on online platforms? **Journal of Intellectual Property, Information Technology and Electronic Commerce Law**, [*S. l.*], v. 14, n. 1, p. 66–83, jun. 2023. Disponível em: https://www.jipitec.eu/jipitec/article/view/366. Acesso em: 5 maio 2025.

SUPREMO TRIBUNAL FEDERAL. **STF define parâmetros para responsabilização de plataformas por conteúdos de terceiros**. Brasília, DF: Supremo Tribunal Federal, 26 jun. 2025. Disponível em: https://noticias.stf.jus.br/postsnoticias/stf-define-parametros-para-responsabilizacao-de-plataformas-por-conteudos-de-terceiros/. Acesso em: 1 ago. 2025.

SUZOR, Nicolas. **Lawless**: the secret rules that govern our digital lives. Cambridge: Cambridge University Press, 2019.

SUZOR, Nicolas *et al.* What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. **International Journal of Communication**, [*S. l.*], v. 13, p. 1526–1543, mar. 2019. Disponível em: https://ijoc.org/index.php/ijoc/article/view/9736. Acesso em: 5 abr. 2025.

SUZOR, Nicolas; GILLETT, Rosalie. Self-regulation and discretion. *In*: FLEW, Terry; MARTIN, Fiona (Org.). **Digital platform regulation**: global perspectives on internet

governance. Cham: Springer International Publishing, 2022. p. 259–279. Disponível em: https://link.springer.com/10.1007/978-3-030-95220-4_13. Acesso em: 9 jan. 2025.

TAYLOR, Bryn. Letter: Imposing fees to access the Twitter API threatens public-interest research. **Coalition for Independent Technology Research**, [*S. l.*], 6 fev. 2023. Disponível em: https://independenttechresearch.org/letter-twitter-api-access-threatens-public-interest-research/. Acesso em: 4 abr. 2024.

TOKOJIMA MACHADO, Dayane F. *et al.* It-which-must-not-be-named: Covid-19 misinformation, tactics to profit from it and to evade content moderation on YouTube. **Frontiers in Communication**, [*S. l.*], v. 7, p. 1–14, nov. 2022. Disponível em: https://doi.org/10.3389/fcomm.2022.1037432. Acesso em: 10 fev. 2025.

TOMAZ, Tales. Brazilian fake news bill: strong content moderation accountability but limited hold on platform market power. **Javnost - The Public**, [*S. l.*], v. 30, n. 2, p. 253–267, maio 2023. Disponível em: https://doi.org/10.1080/13183222.2023.2201801. Acesso em: 4 abr. 2025.

TRANS, Martin *et al.* APIcalypse now: redefining data access regimes in the face of the Digital Services Act. **Digital Methods Initiative**, Amsterdã, 20 fev. 2024. Disponível em: https://www.digitalmethods.net/Dmi/WinterSchool2024APIcalypse. Acesso em: 4 abr. 2025.

TUCKER, Joshua A. *et al.* From liberation to turmoil: social media and democracy. **Journal of Democracy**, [*S. l.*], v. 28, n. 4, p. 46–59, out. 2017. Disponível em: https://doi.org/10.1353/jod.2017.0064. Acesso em: 5 abr. 2025.

TUCKER, Joshua A. *et al.* Social media, political polarization, and political disinformation: a review of the scientific literature. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2018. Disponível em: https://www.ssrn.com/abstract=3144139. Acesso em: 30 abr. 2024.

URMAN, Aleksandra; MAKHORTYKH, Mykola. How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. **Telecommunications Policy**, [*S. l.*], v. 47, n. 3, p. 1–15, abr. 2023. Disponível em: https://doi.org/10.1016/j.telpol.2022.102477. Acesso em: 5 abr. 2025.

VAN DIJCK, José; POELL, Thomas; DE WAAL, Martijn. **The platform society**: public values in a connective world. Nova York: Oxford University Press, 2018. Disponível em: https://academic.oup.com/book/12378. Acesso em: 3 abr. 2023.

VAN DIJCK, José; NIEBORG, David; POELL, Thomas. Reframing platform power. **Internet Policy Review**, [*S. l.*], v. 8, n. 2, p. 1–18, jun. 2019. Disponível em: https://policyreview.info/node/1414. Acesso em: 2 jan. 2022.

VAN DIJCK, José. Governing digital societies: private platforms, public values. **Computer Law & Security Review**, [*S. l.*], v. 36, 1–4, abr. 2020. Disponível em: https://doi.org/10.1016/j.clsr.2019.105377. Acesso em: 5 abr. 2025.

VERGARA, Caitlyn; JAIN, Raghav; MEHTA, Swapneel. A history of transparency regulations: interdisciplinary strategies for shaping social media regulation and self-governance. *In:* DG.O 2024: 25TH ANNUAL INTERNATIONAL CONFERENCE ON

DIGITAL GOVERNMENT RESEARCH, 2024, Taipei. **Anais** [...]. Nova York: Association for Computing Machinery, 11 jun. 2024. p. 875–883. Disponível em: https://dl.acm.org/doi/10.1145/3657054.3657157. Acesso em: 3 jan. 2025.

VESE, Donato. Governing fake news: the regulation of social media and the right to freedom of expression in the era of emergency. **European Journal of Risk Regulation**, [*S. l.*], v. 13, n. 3, p. 477–513, set. 2022. Disponível em: https://doi.org/10.1017/err.2021.48. Acesso em: 1 jul. 2025.

VIVAS, Fernanda. Por 8 votos a 3, STF decide que redes podem ser responsabilizadas por posts de usuários. **G1**, Brasília, 26 jun. 2025. Disponível em: https://g1.globo.com/politica/noticia/2025/06/26/stf-retoma-julgamento-sobre-responsabilizacao-das-redes-sociais-por-posts-de-usuarios.ghtml. Acesso em: 1 jul. 2025.

VOSOUGHI, Soroush; ROY, Deb; ARAL, Sinan. The spread of true and false news online. **Science**, [*S. l.*], v. 359, n. 6380, p. 1146–1151, mar. 2018. Disponível em: https://doi.org/10.1126/science.aap9559. Acesso em: 15 jun. 2025.

WAGNER, Ben *et al.* Regulating transparency? Facebook, Twitter and the German Network Enforcement Act. *In:* FAT '20: CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2020, Barcelona. **Anais** [...]. Nova York: Association for Computing Machinery, 27 jan. 2020. p. 261–271. Disponível em: https://doi.org/10.1145/3351095.3372856. Acesso em: 3 jan. 2025.

WALDRON, Patricia. One-size-fits-all content moderation fails the Global South. **Cornell Chronicle**, Ithaca, 13 abr. 2023. Disponível em: https://news.cornell.edu/stories/2023/04/one-size-fits-all-content-moderation-fails-global-south. Acesso em: 24 fev. 2025.

WALKER, Shawn; MERCEA, Dan; BASTOS, Marco. The disinformation landscape and the lockdown of social platforms. **Information, Communication & Society**, [*S. l.*], v. 22, n. 11, p. 1531–1543, ago. 2019. Disponível em: https://doi.org/10.1080/1369118X.2019.1648536. Acesso em: 17 abr. 2025.

WANG, Rui *et al.* Fake news or bad news? Toward an emotion-driven cognitive dissonance model of misinformation diffusion. **Asian Journal of Communication**, [*S. l.*], v. 30, n. 5, p. 317–342, ago. 2020. Disponível em: https://doi.org/10.1080/01292986.2020.1811737. Acesso em: 14 jun. 2025.

WARDLE, Claire; DERAKHSHAN, Hossein. Information disorder: toward an interdisciplinary framework for research and policy making. **Council of Europe**, União Europeia, out. 2017. Disponível em: https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html. Acesso em: 20 jun. 2022.

WARNKE, Lina; MAIER, Anna-Lena; GILBERT, Dirk Ulrich. Social media platforms' responses to Covid-19-related mis- and disinformation: the insufficiency of self-governance. **Journal of Management and Governance**, [*S. l.*], v. 28, n. 4, p. 1079–1115, fev. 2024. Disponível em: https://doi.org/10.1007/s10997-023-09694-5. Acesso em: 27 dez. 2024.

WIGGERS, Kyle. Twitter reveals some of its source code, including its recommendation algorithm. **TechCrunch**, [*S. l.*], 31 mar. 2023. Disponível em:

https://techcrunch.com/2023/03/31/twitter-reveals-some-of-its-source-code-including-its-recommendation-algorithm/. Acesso em: 7 abr. 2025.

WOOLLEY, Samuel. Bots and computational propaganda: automation for communication and control. *In*: PERSILY, Nathaniel; TUCKER, Joshua A. (Org.). **Social media and democracy**: the state of the field, prospects for reform. Cambridge: Cambridge University Press, 2020. p. 89–110. Disponível em: https://www.cambridge.org/core/books/social-media-and-democracy/bots-and-computational-propaganda-automation-for-communication-and-control/A15EE25C278B442EF00199AA660BFADD. Acesso em: 14 jun. 2025.

X SAFETY. Freedom of speech, not reach: an update on our enforcement philosophy. **X Blog**, [*S. l.*], 17 abr. 2023. Disponível em: https://blog.x.com/en_us/topics/product/2023/freedom-of-speech-not-reach-an-update-on-our-enforcement-philosophy. Acesso em: 12 maio 2025.

X/TWITTER. Global transparency report: H2 2024. **X Transparency**, [*S. l.*], 2025. Disponível em: https://transparency.x.com/en/reports/global-reports/2025-transparency-report. Acesso em: 28 abr. 2025.

YATES, Luke. How platform businesses mobilize their users and allies: corporate grassroots lobbying and the Airbnb 'movement' for deregulation. **Socio-Economic Review**, [*S. l.*], v. 21, n. 4, p. 1917–1943, jun. 2023. Disponível em: https://doi.org/10.1093/ser/mwad028. Acesso em: 19 abr. 2025.

ZALNIERIUTE, Monika. "Transparency-washing" in the digital age: a corporate agenda of procedural fetishism. **Critical Analysis of Law**, Sydney, v. 8 n. 1, p. 39–53, 2021. Disponível em: https://papers.ssrn.com/abstract=3805492. Acesso em: 1 ago. 2024.

ZIPURSKY, Rebecca. Nuts about NETZ: the Network Enforcement Act and freedom of expression. **Fordham International Law Journal**, [*S. l.*], v. 42, p. 1325–1374, 2019. Disponível em: https://ir.lawnet.fordham.edu/ilj/vol42/iss4/7. Acesso em: 5 mar. 2025.

ZITTRAIN, Jonathan L. Three eras of digital governance. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2019. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3458435. Acesso em: 18 fev. 2025.

ZUBOFF, Shoshana. Big other: surveillance capitalism and the prospects of an information civilization. **Journal of Information Technology**, [*S. l.*], v. 30, n. 1, p. 75–89, mar. 2015. Disponível em: https://doi.org/10.1057/jit.2015.5. Acesso em: 5 abr. 2025.

ZUBOFF, Shoshana. "We make them dance": surveillance capitalism, the rise of instrumentarian power, and the threat to human rights. *In*: JØRGENSEN, Rikke Frank (Org.). **Human rights in the age of platforms**. Cambridge, Massachusetts: The MIT Press, 2019. p. 3–52. Disponível em: https://direct.mit.edu/books/book/4531/chapter/202528/We-Make-Them-Dance-Surveillance-Capitalism-the. Acesso em: 25 jan. 2025.

ZUBOFF, Shoshana. **A era do capitalismo de vigilância**: a luta por um futuro humano na nova fronteira do poder. Rio de Janeiro: Intrínseca, 2020.

ZUCKERMAN, Ethan. Why study media ecosystems? **Information, Communication &**

Society, [*S. l.*], v. 24, n. 10, p. 1495–1513, jul. 2021. Disponível em: https://doi.org/10.1080/1369118X.2021.1942513. Acesso em: 1 jun. 2025.

APÊNDICE – QUADRO ANALÍTICO PARA COMPARAÇÃO DOS RELATÓRIOS DE TRANSPARÊNCIA DE MODERAÇÃO DE CONTEÚDO SELECIONADOS

Disposições gerais

1 O relatório de transparência é publicado com uma periodicidade fixa, no mínimo semestral?

Embasamento: Critério inspirado no Índice de Transparência de Dados publicado pelo NetLab UFRJ (Santini *et al.*, 2024) e na determinação do *Digital Services Act* para *very large online platforms*, em seu artigo 42 (European Parliament, 2022). O fato de as plataformas seguirem ou não uma periodicidade para a publicação desse tipo de documento impacta diretamente a capacidade de partes interessadas de compará-los entre si.

2 O relatório de transparência informa a data em que foi publicado, além do período coberto por seus dados?

Embasamento: Critério inspirado na prática adotada por algumas plataformas de indicar a data de publicação dos relatórios, bem como de eventuais atualizações de seu conteúdo (ver Google, 2025; Meta, 2024a, b). Embora essa exigência não conste de forma explícita no *Digital Services Act* (European Parliament, 2022), a disponibilização da data de publicação contribui para uma aplicação mais eficaz da legislação, que prevê sanções em casos de descumprimento de prazos.

3 É possível extrair os dados do relatório de transparência em formato estruturado e legível por máquinas?

Embasamento: Recomendação de Radsch (2022) e prática já adotada por algumas das plataformas analisadas (ver Meta, 2025, [*S. d.*]), essa medida facilita a análise automatizada e em larga escala dos dados divulgados nos relatórios.

4 O relatório de transparência apresenta um glossário ou seções explicativas que descrevem de forma clara as dimensões dos dados apresentados?

Embasamento: Recomendação de Radsch (2022) e de Urman e Makhortykh (2023), essa medida contribui para uma compreensão mais clara e objetiva das informações apresentadas nos documentos por qualquer parte interessada.

5 O relatório de transparência apresenta o número médio mensal de usuários da plataforma em cada país de atuação?

Embasamento: Determinação do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), contribui para uma contextualização mais precisa das informações apresentadas nos relatórios.

6 O relatório de transparência apresenta o número de moderadores especializados em cada idioma e/ou dedicados a cada país de atuação?

Embasamento: Determinação do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), contribui para uma contextualização mais precisa das informações apresentadas nos relatórios.

7 O relatório de transparência traz informações sobre as responsabilidades dos moderadores de conteúdo e os recursos de apoio e treinamento fornecidos pela plataforma?

Embasamento: Determinação do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), contribui para um maior entendimento das atribuições dos moderadores de conteúdo, que, geralmente, são pouco visíveis ou acessíveis ao público em geral.

8 O relatório de transparência apresenta exemplos ilustrativos de casos relevantes de moderação de conteúdo?

Embasamento: Recomendação de MacCarthy (2020), Radsch (2022) e Urman e Makhortykh (2023), permite que os usuários compreendam de forma mais concreta os tipos de conteúdo que violam as regras das plataformas e, portanto, estão sujeitos à moderação.

9 O relatório de transparência descreve a utilização de sistemas automatizados para moderação de conteúdo, incluindo critérios utilizados, cenários de uso e limitações?

Embasamento: Critério adaptado do artigo 15 do *Digital Services Act* (European Parliament, 2022), das recomendações de Radsch (2022) e dos *Santa Clara Principles* (2021), amplia a compreensão sobre o uso de sistemas automatizados na moderação de conteúdo.

10 O relatório de transparência apresenta métricas de desempenho e acurácia dos sistemas automatizados para moderação de conteúdo, discriminadas por idioma?

Embasamento: Determinação do *Digital Services Act*, em seus artigos 15 e 42 (European Parliament, 2022), e recomendação dos *Santa Clara Principles* (2021) e de MacCarthy (2020), permite a identificação de possíveis vieses linguísticos e/ou regionais na incorporação de sistemas automatizados à moderação de conteúdo.

Moderação de conteúdo por determinação da plataforma

11 O relatório de transparência apresenta o número de publicações orgânicas removidas por determinação da plataforma, discriminadas por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem particularidades regionais nas ações de moderação das plataformas e nas categorias de publicações mais afetadas, oferecendo indícios sobre a prevalência de determinados conteúdos problemáticos.

11.1 O relatório de transparência especifica o número de publicações orgânicas removidas proativa e reativamente por determinação da plataforma, discriminadas por país e tipo de violação?

Embasamento: Critério inspirado diretamente nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Meta, 2025), permite que as partes interessadas analisem o quão dependentes das denúncias de usuários são as ações de moderação de conteúdo das plataformas.

11.2 O relatório de transparência especifica o número de publicações orgânicas removidas por determinação dos sistemas automatizados da plataforma, discriminadas por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021) e uma determinação do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), permite que as partes interessadas analisem o quão dependentes de sistemas automatizados são as ações de moderação de conteúdo das plataformas.

11.3 O relatório de transparência apresenta informações agregadas ou médias do engajamento e/ou alcance das publicações orgânicas removidas no momento da moderação, discriminados por país e tipo de violação?

Embasamento: Critério inspirado diretamente em MacCarthy (2020) e nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Google, [*S. d.*]), que informam o número de visualizações de publicações no momento de sua remoção, permitindo que as partes interessadas quantifiquem o alcance e o impacto de conteúdos problemáticos antes da moderação.

12 O relatório de transparência apresenta o número de publicações orgânicas com alcance reduzido por determinação da plataforma, discriminadas por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite maior visibilidade dessas ações, geralmente ofuscadas pelas plataformas.

13 O relatório de transparência apresenta o número de anúncios removidos por determinação da plataforma, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem particularidades regionais nas ações de moderação das plataformas e nas categorias de anúncios mais impactadas, fornecendo indícios sobre a prevalência de determinadas formas de publicidade problemática.

13.1 O relatório de transparência especifica o número de anúncios removidos proativa e reativamente por determinação da plataforma, discriminados por país e tipo de violação?

Embasamento: Critério inspirado diretamente nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Meta, 2025), embora estes não especifiquem ações tomadas em relação a anúncios, permite que as partes interessadas analisem o quão dependentes das denúncias de usuários são as ações de moderação de publicidade das plataformas.

13.2 O relatório de transparência especifica o número de anúncios removidos por determinação dos sistemas automatizados da plataforma, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021) e uma determinação do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), embora esta não seja específica a anúncios, permite que as partes interessadas analisem o quão dependentes de sistemas automatizados são as ações de moderação de publicidade das plataformas.

13.3 O relatório de transparência apresenta informações agregadas ou médias do alcance dos anúncios removidos no momento da moderação, discriminados por país e tipo de violação?

Embasamento: Critério inspirado diretamente em MacCarthy (2020) e nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Google, [*S. d.*]), que apresentam informações sobre as visualizações de vídeos removidos no momento da remoção, mas não sobre o alcance de conteúdo publicitário. Permite que as partes interessadas quantifiquem o alcance e o impacto de anúncios problemáticos antes da moderação.

14 O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos por determinação da plataforma, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem particularidades regionais nas ações de moderação das plataformas, oferecendo indícios sobre a prevalência de comportamentos problemáticos entre seus usuários.

14.1 O relatório de transparência especifica o número de usuários restritos, suspensos e/ou banidos proativa e reativamente por determinação da plataforma, discriminados por país e tipo de violação?

Embasamento: Critério inspirado diretamente nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Meta, 2025), embora estes não especifiquem ações tomadas em relação a usuários. Permite que as partes interessadas analisem o quão dependentes das denúncias são as ações de moderação de usuários das plataformas.

14.2 O relatório de transparência especifica o número de usuários suspensos e/ou banidos por determinação dos sistemas automatizados da plataforma, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021) e uma determinação do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), embora esta não seja específica a usuários, permite que as partes interessadas analisem o quão dependentes de sistemas automatizados são as ações de moderação de usuários das plataformas.

14.3 O relatório de transparência apresenta informações agregadas ou médias do engajamento e/ou alcance dos usuários restritos, suspensos e/ou banidos no momento da moderação, discriminados por país e tipo de violação?

Embasamento: Critério inspirado diretamente em MacCarthy (2020) e nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Google, [*S. d.*]), que apresentam informações sobre as visualizações de vídeos removidos no momento da remoção, mas não sobre o alcance total de usuários. Permite que as partes interessadas quantifiquem o alcance e o impacto de usuários problemáticos antes da moderação.

15 O relatório de transparência apresenta informações sobre outros tipos de conteúdos e interações orgânicas removidos e/ou restritos por determinação da plataforma (por exemplo, comentários, classificados de *marketplace*)?

Embasamento: Critério inspirado diretamente nos relatórios de transparência voluntários publicados por algumas das plataformas analisadas (ver Google, [*S. d.*]), que apresentam, por exemplo, estatísticas sobre a moderação de comentários. Já o artigo 15 do *Digital Services Act* determina que as plataformas sejam transparentes quanto a "qualquer ação de moderação de conteúdo que tenham realizado" (European Parliament, 2022, n.p., tradução do autor). O critério é intencionalmente mais amplo e genérico, de modo a não exigir o detalhamento completo da arquitetura digital das plataformas, permitindo avaliar de forma mais abrangente a transparência de suas diversas ações de moderação de conteúdo.

Denúncias realizadas por usuários

16 O relatório de transparência apresenta o número de denúncias de publicações orgânicas feitas por usuários, discriminadas por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*). Permite que as partes interessadas avaliem, a partir das percepções dos usuários, a prevalência de conteúdos e comportamentos problemáticos e/ou ilegais nas plataformas analisadas.

16.1 O relatório de transparência apresenta o número de publicações orgânicas removidas após denúncias de usuários, discriminadas por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*). Permite que as partes interessadas avaliem o grau de alinhamento entre a avaliação das plataformas e a percepção dos usuários sobre conteúdos e comportamentos problemáticos e/ou ilegais.

16.2 O relatório de transparência apresenta o número de denúncias de publicações orgânicas feitas por usuários, discriminadas por tipo de processamento (automático ou manual)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*). Permite que as partes interessadas compreendam como as plataformas tratam conteúdos e comportamentos percebidos como problemáticos e/ou ilegais por seus usuários.

16.3 O relatório de transparência apresenta o número de denúncias de publicações orgânicas, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*). Permite que as partes interessadas analisem o papel dos diversos atores no processo de moderação, identificando aqueles que colaboram mais ativamente com a moderação de conteúdo.

16.4 O relatório de transparência apresenta o tempo médio para lidar com denúncias de publicações orgânicas feitas por usuários, discriminado por país e tipo de violação?

Embasamento: O critério é inspirado nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*). Permite que as partes interessadas avaliem a agilidade das plataformas na resposta às denúncias enviadas por seus usuários.

17 O relatório de transparência apresenta o número de denúncias de anúncios feitas por usuários, discriminadas por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas avaliem, a partir das percepções dos usuários, a prevalência de anúncios problemáticos e/ou ilegais nas plataformas analisadas.

17.1 O relatório de transparência apresenta o número de anúncios removidos após denúncias de usuários, discriminados por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas avaliem o grau de alinhamento entre a avaliação das plataformas e a percepção dos usuários sobre anúncios problemáticos e/ou ilegais.

17.2 O relatório de transparência apresenta o número de denúncias de anúncios feitas por usuários, discriminadas por tipo de processamento (automático ou manual)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas compreendam como as plataformas tratam anúncios percebidos como problemáticos e/ou ilegais por seus usuários.

17.3 O relatório de transparência apresenta o número de denúncias de anúncios, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas analisem o papel dos diversos atores no processo de moderação, identificando aqueles que colaboram mais ativamente com a moderação de publicidade.

17.4 O relatório de transparência apresenta o tempo médio para lidar com denúncias de anúncios feitas por usuários, discriminado por país e tipo de violação?

Embasamento: O critério é inspirado nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas avaliem a agilidade das plataformas na resposta às denúncias enviadas por seus usuários.

18 O relatório de transparência apresenta o número de denúncias de usuários feitas por outros usuários, discriminadas por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de outros usuários. Permite que as partes interessadas avaliem, a partir das percepções dos usuários, a prevalência de comportamentos problemáticos e/ou ilegais nas plataformas analisadas.

18.1 O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos após denúncias de outros usuários, discriminados por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas

sobre denúncias de outros usuários. Permite que as partes interessadas avaliem o grau de alinhamento entre a avaliação das plataformas e a percepção dos usuários sobre comportamentos problemáticos e/ou ilegais.

18.2 O relatório de transparência apresenta o número de denúncias de usuários feitas por outros usuários, discriminadas por tipo de processamento (automático ou manual)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 42 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de outros usuários. Permite que as partes interessadas compreendam como as plataformas tratam comportamentos percebidos como problemáticos e/ou ilegais por seus usuários.

18.3 O relatório de transparência apresenta o número de denúncias de usuários, discriminadas por categoria de usuário responsável pela denúncia (por exemplo, *trusted flaggers*)?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, conforme seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de outros usuários. Permite que as partes interessadas analisem o papel dos diversos atores no processo de moderação, identificando aqueles que colaboram mais ativamente com a moderação de usuários.

18.4 O relatório de transparência apresenta o tempo médio para lidar com denúncias de usuários feitas por outros usuários, discriminado por país e tipo de violação?

Embasamento: O critério é inspirado nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de denúncias enviadas por usuários, com base em seu artigo 16 (*Notice and Action Mechanisms*), embora este não faça determinações específicas sobre denúncias de outros usuários. Permite que as partes interessadas avaliem a agilidade das plataformas na resposta às denúncias enviadas por seus usuários.

Restauração de conteúdo e contestações à moderação

19 O relatório de transparência apresenta o número de contestações à remoção de publicações orgânicas, discriminadas por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de contestações enviadas por usuários moderados, conforme seu artigo 20 (*Internal complaint-handling system*). Permite que as partes interessadas avaliem o grau de discordância dos usuários cujos conteúdos foram moderados pelas plataformas, indicando possíveis casos de moderação excessiva ou decisões equivocadas que impactam injustamente determinados indivíduos.

19.1 O relatório de transparência apresenta o número de publicações orgânicas restauradas, discriminadas por país e tipo de violação, após contestações de usuários?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de contestações enviadas por usuários moderados, conforme seu artigo 20 (*Internal complaint-handling system*). Permite que as partes interessadas avaliem o grau de reconhecimento reativo de erros de moderação de conteúdo por parte das plataformas.

20 O relatório de transparência apresenta o número de publicações orgânicas restauradas, discriminadas por país e tipo de violação, após a identificação proativa de erros na moderação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021) e de MacCarthy (2020), permite que as partes interessadas avaliem o grau de reconhecimento proativo de erros de moderação de conteúdo por parte das plataformas.

21 O relatório de transparência apresenta o número de contestações à restrição, suspensão e/ou ao banimento de usuários, discriminadas por país?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de contestações enviadas por usuários moderados, conforme seu artigo 20 (*Internal complaint-handling system*). Permite que as partes

interessadas avaliem o grau de discordância dos usuários cujos perfis foram moderados pelas plataformas, indicando possíveis casos de moderação excessiva ou decisões equivocadas que impactam injustamente determinados indivíduos.

21.1 O relatório de transparência apresenta o número de usuários restaurados, discriminados por país, após contestações dos próprios usuários moderados?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), sobre a disponibilização de informações de contestações enviadas por usuários moderados, conforme seu artigo 20 (*Internal complaint-handling system*). Permite que as partes interessadas avaliem o grau de reconhecimento reativo de erros de moderação de conteúdo por parte das plataformas.

22 O relatório de transparência apresenta o número de usuários restaurados discriminados por país, após a identificação proativa de erros na moderação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021) e de MacCarthy (2020), permite que as partes interessadas avaliem o grau de reconhecimento proativo de erros de moderação de conteúdo por parte das plataformas.

Demandas de autoridades públicas

23 O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição e/ou remoção de publicações orgânicas, discriminados por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022). Permite que as partes interessadas avaliem o impacto de fatores externos nos processos de moderação de conteúdo das plataformas, frequentemente relacionado à ineficácia dessas plataformas nessa área.

23.1 O relatório de transparência apresenta o número de publicações orgânicas restritas e/ou removidas por determinação de autoridades públicas, discriminadas por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem o grau de cumprimento das ordens de autoridades públicas para moderação de conteúdo por parte das plataformas.

23.2 O relatório de transparência indica se as publicações orgânicas restritas e/ou removidas após determinação de autoridades públicas foram moderadas por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem se as ordens de autoridades públicas para moderação de conteúdo foram motivadas por questões cobertas ou não pelas políticas internas das plataformas.

23.3 O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição e/ou remoção de publicações orgânicas, discriminadas por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem as autoridades responsáveis pelos pedidos de moderação de conteúdo.

23.4 O relatório de transparência informa a quantidade de pedidos de remoção de publicações orgânicas que tiveram origem em ordens judiciais, discriminados por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas possam identificar a origem clara de pedidos de moderação de conteúdo.

24 O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição e/ou remoção de anúncios, discriminados por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022), embora este não faça determinações específicas sobre denúncias de conteúdo publicitário. Permite que as partes interessadas avaliem o impacto de fatores externos nos processos de moderação de conteúdo das plataformas, frequentemente relacionado à ineficácia dessas plataformas nessa área.

24.1 O relatório de transparência apresenta o número de anúncios restritos e/ou removidos por determinação de autoridades públicas, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem o grau de cumprimento das ordens de autoridades públicas para moderação de conteúdo publicitário por parte das plataformas.

24.2 O relatório de transparência indica se os anúncios restritos e/ou removidos após determinação de autoridades públicas foram moderados por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem se as ordens de autoridades públicas para moderação de conteúdo publicitário foram motivadas por questões cobertas ou não pelas políticas internas das plataformas.

24.3 O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição e/ou remoção de anúncios, discriminadas por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem as autoridades responsáveis pelos pedidos de moderação de conteúdo publicitário.

24.4 O relatório de transparência informa a quantidade de pedidos de remoção de anúncios que tiveram origem em ordens judiciais, discriminados por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas possam identificar a origem clara de pedidos de moderação de conteúdo.

25 O relatório de transparência apresenta o número de pedidos feitos por autoridades públicas para restrição, suspensão e/ou banimento de usuários, discriminados por país e tipo de violação?

Embasamento: O critério é inspirado nas recomendações dos *Santa Clara Principles* (2021) e nas determinações do *Digital Services Act*, em seu artigo 15 (European Parliament, 2022). Permite que as partes interessadas avaliem o impacto de fatores externos nos processos de

moderação de conteúdo das plataformas, frequentemente relacionado à ineficácia dessas plataformas nessa área.

25.1 O relatório de transparência apresenta o número de usuários restritos, suspensos e/ou banidos por determinação de autoridades públicas, discriminados por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem o grau de cumprimento das ordens de autoridades públicas para moderação de usuários por parte das plataformas.

25.2 O relatório de transparência indica se os usuários restritos, suspensos e/ou banidos após determinação de autoridades públicas foram moderados por violações a legislações locais ou às políticas internas da plataforma, discriminando os números por país e tipo de violação?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas avaliem se as ordens de autoridades públicas para moderação de usuários foram motivadas por questões cobertas ou não pelas políticas internas das plataformas.

25.3 O relatório de transparência especifica as autoridades públicas por trás dos pedidos de restrição, suspensão e/ou banimento de usuários, discriminadas por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas identifiquem as autoridades responsáveis pelos pedidos de moderação de usuários.

25.4 O relatório de transparência informa a quantidade de pedidos de restrição, suspensão e/ou banimento de usuários que tiveram origem em ordens judiciais, discriminados por país?

Embasamento: Recomendação dos *Santa Clara Principles* (2021), permite que as partes interessadas possam identificar a origem clara de pedidos de moderação de conteúdo.

REFERÊNCIAS

EUROPEAN PARLIAMENT. **Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act)**. Bruxelas: Parlamento Europeu, 2022. Disponível em: http://data.europa.eu/eli/reg/2022/2065/oj/eng. Acesso em: 4 jan. 2025.

GOOGLE. EU DSA transparency report – July 1, 2024 to December 31, 2024. **Google**, [*S. l.*], 28 fev. 2025. Disponível em: https://storage.googleapis.com/transparencyreport/report-downloads/pdf-report-27_2024-7-1_2024-12-31_en_v1.pdf. Acesso em: 28 abr. 2025.

GOOGLE. YouTube policy removals – transparency report. **Google**, [*S. l.*], [*S. d.*]. Disponível em: https://transparencyreport.google.com/youtube-policy/removals. Acesso em: 28 abr. 2025.

MACCARTHY, Mark. Transparency requirements for digital social media platforms: recommendations for policy makers and industry. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2020. Disponível em: https://www.ssrn.com/abstract=3615726. Acesso em: 30 abr. 2024.

META. DSA transparency report – April - September 2024: Facebook. **Meta Transparency Center**, [*S. l.*], 25 out. 2024a. Disponível em: https://transparency.meta.com/sr/dsa-transparency-report-sep2024-facebook. Acesso em: 28 abr. 2025.

META. DSA transparency report – April - September 2024: Instagram. **Meta Transparency Center**, [*S. l.*], 25 out. 2024b. Disponível em: https://scontent.fsdu8-2.fna.fbcdn.net/v/t39.8562-6/466943155_1291701138400105_7867447844898917200_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=b8d81d&_nc_ohc=ApoOoCEllfAQ7kNvwFRQnAn&_nc_oc=Adm81Piz8hRSBE9lKzNsjbzyj7uOa-QNz_DKGX1JmlYSsZfuq_D_YJ-B69QCMNvSQ_KK5i1t8NUDv2N-ILhawLah&_nc_zt=14&_nc_ht=scontent.fsdu8-2.fna&_nc_gid=cVUCVFcM3VMSnrWeQ7cX3Q&oh=00_AfQF03TjrSKEtruB1B2WQcsOfs3SLtlbN5HBMdu6nIMAaQ&oe=688EDADD. Acesso em: 28 abr. 2025.

META. Community standards enforcement report. **Meta Transparency Center**, [*S. l.*], fev. 2025. Disponível em: https://transparency.meta.com/reports/community-standards-enforcement/. Acesso em: 28 abr. 2025.

META. Content restrictions based on local law. **Meta Transparency Center**, [*S. l.*], [*S. d.*]. Disponível em: https://transparency.meta.com/reports/content-restrictions/. Acesso em: 28 abr. 2025.

RADSCH, Courtney. Transparency reporting: good practices and lessons from global assessment frameworks. **SSRN Scholarly Paper**, Rochester, NY: Social Science Research Network, 2022. Disponível em: https://dx.doi.org/10.2139/ssrn.4416400. Acesso em: 28 abr. 2025.

SANTA CLARA PRINCIPLES ON TRANSPARENCY AND ACCOUNTABILITY IN CONTENT MODERATION. **Santa Clara Principles**, [*S. l.*], dez. 2021. Disponível em: https://santaclaraprinciples.org. Acesso em: 7 jan. 2025.

SANTINI, R. Marie *et al.* Índice de transparência de dados das plataformas de redes sociais. **NetLab UFRJ**, Rio de Janeiro, 2024. Disponível em: https://netlab.eco.ufrj.br/itd. Acesso em: 25 jan. 2025.

URMAN, Aleksandra; MAKHORTYKH, Mykola. How transparent are transparency reports? Comparative analysis of transparency reporting across online platforms. **Telecommunications Policy**, [*S. l.*], v. 47, n. 3, p. 1–15, abr. 2023. Disponível em: https://doi.org/10.1016/j.telpol.2022.102477. Acesso em: 5 abr. 2025.