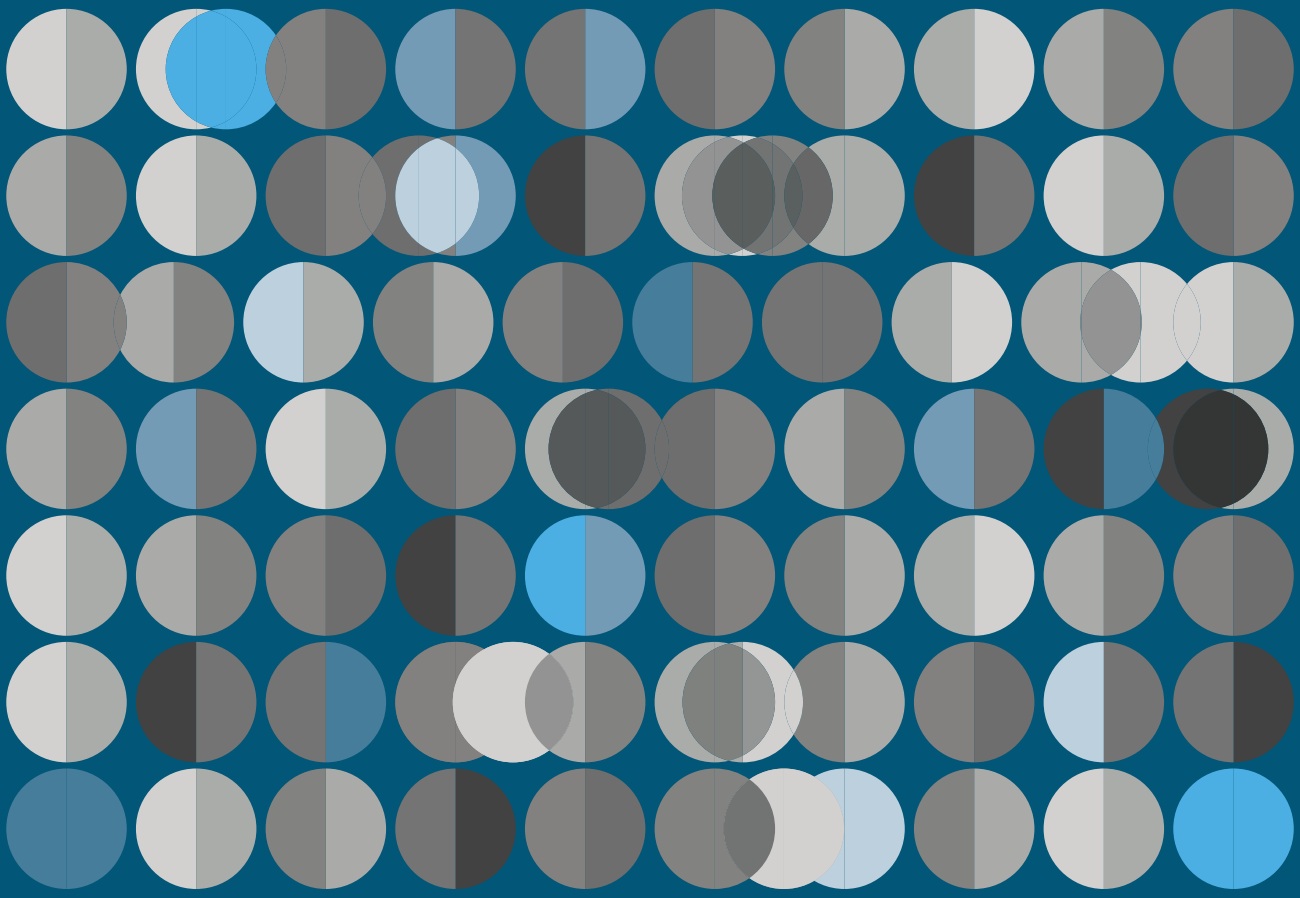
# COMPROMISSO COM A DEMOCRACIA

## Integridade eleitoral e o Estado Democrático de Direito nas políticas de plataformas digitais

**Diagnósticos e recomendações nº 10**
Julho 2023



**Ester Borges (coordenação)**
**Francisco Brito Cruz**
**Anna Martha Cintra**

INTERNETLAB

# COMPROMISSO COM A DEMOCRACIA

Julho 2023



**COMO CITAR**

Borges, E. (coord); Brito Cruz, F e Cintra, A. M, "Compromisso com a democracia: Integridade eleitoral e o Estado Democrático de Direito nas políticas de plataformas digitais", Diagnósticos & Recomendações nº10 (São Paulo: InternetLab, 2023)

**PESQUISA E REDAÇÃO**

Ester Borges (coordenação)
Francisco Brito Cruz
Anna Martha Cintra

**REVISÃO**

Daniele Kleiner

**COLABORAÇÃO**

Iná Jost
Heloisa Massaro
Fernanda Martins Sousa

**DIAGRAMAÇÃO E DESIGN**

Joana Resek

**COMUNICAÇÃO**

João Vitor Araújo

**APOIO**

International Republican Institute (IRI)

INTERNETLAB

# SUMÁRIO

# APRESENTAÇÃO

## SOBRE O INTERNETLAB

O InternetLab é um centro independente de pesquisa interdisciplinar que promove o debate acadêmico e a produção de conhecimento nas áreas de direito e tecnologia, sobretudo no campo da Internet. Uma entidade sem fins lucrativos, a organização atua como ponto de articulação entre acadêmicos e representantes dos setores público, privado e da sociedade civil, incentivando o desenvolvimento de projetos que abordam os desafios de elaboração e implementação de políticas públicas em novas tecnologias, como privacidade, liberdade de expressão e questões ligadas a gênero, raça, sexualidade e outros marcadores sociais da diferença.

## QUAL O OBJETIVO DESTE DOCUMENTO?

Esta é mais uma iniciativa do InternetLab para contribuir com o debate público sobre o papel das plataformas digitais e seus sistemas de "moderação de conteúdo", com especial atenção à sua relação com a preservação da integridade das eleições e do Estado Democrático de Direito no ambiente digital.

A consolidação de plataformas de redes sociais e outros serviços como infraestruturas que abrigam as conversas sobre processos eleitorais e democrático, como campanhas eleitorais e processos de transição pacífica de poder em sociedades democráticas, traz uma série de desafios. Exemplos recentes tornaram visível como agentes políticos e econômicos fizeram uso destas plataformas para a disseminação de propaganda antidemocrática que mobiliza conteúdos desinformativos, teorias da conspiração, e incitação à violência para viabilizar movimentos políticos orientados a deslegitimar ou mesmo abolir pilares do Estado Democrático de Direito. Estes exemplos, nacionais[1] e internacionais[2], levam à uma necessária reflexão sobre qual o compromisso político democrático que deve ser assumido pelas empresas que controlam tais serviços independente de qualquer regulação estatal.

Tendo isso em mente, este documento, ao navegar pelas políticas de conteúdo definidas por esses intermediários e identificar os seus pontos fortes e suas lacunas, pretende avançar nas discussões sobre a chamada "moderação de conteúdo", visando produzir parâmetros de compromisso com a integridade eleitoral e com o Estado Democrático de Direito **independente do que é exigido pela regulação estatal.**

**Mais do que uma "receita para a regulação", o objetivo é pontuar um caminho para um compromisso político forte das plataformas com a democracia que poderá possivelmente inspirar o debate multissetorial sobre o tema.**

1 Em 8 de janeiro de 2023, manifestantes invadiram a sede dos Três Poderes da República em Brasília, causando danos físicos aos prédios e destruindo obras de arte. Estima-se que o prejuízo do ataque chegou a quase R$ 3 milhões. Disponível em <https://www12.senado.leg.br/tv/programas/tela-brasil/2023/02/8-de-janeiro-um-ataque-a-democracia-do-brasil>.

2 Em 6 de janeiro de 2021 houve a invasão do Capitólio dos Estados Unidos por apoiadores do então presidente Donald Trump, quando ocorreria a sessão conjunta do Congresso que confirmaria a vitória de Joe Biden, adversário de Trump, nas eleições presidenciais de 2020. Sob a alegação de fraude nas votações, centenas de pessoas invadiram o espaço, causando a destruição de diversos objetos e ameaçando de morte os congressistas. Disponível em <https://www.cnnbrasil.com.br/internacional/invasao-ao-capitolio-completa-um-ano-relembre-o-ataque-a-democracia-dos-eua/>.

# PRINCIPAIS PONTOS

1

**As plataformas digitais (ou *provedores de aplicação de internet*) são verdadeiras infraestruturas do debate público.** Se, por um lado, elas carregam enorme porção das conversas entre as(os) cidadãs(os) sobre assuntos de interesse da vida cívica, por outro, também trazem preocupações relacionadas aos fenômenos da desinformação e da violência política. Esse cenário, assim, traz questionamentos sobre como as empresas responsáveis por essas plataformas arquitetam e moldam a expressão no meio digital, isto é, como elas moderam conteúdo em seus espaços digitais.

2

**Por arbitrar decisões sobre a expressão humana em escala industrial e global, a moderação de conteúdo oferecida por plataformas de internet é uma atividade que afeta continuamente uma série de valores democráticos e direitos humanos.** Ela envolve uma vasta quantidade de desafios, que vão desde utilização de ferramentas de inteligência artificial a discussões sobre os limites da liberdade de expressão, e que passam a ganhar uma nova roupagem à medida que esses serviços se concretizam como infraestruturas de debate público a respeito de processos democráticos e eleitorais.

3

**O aumento de narrativas que visam deslegitimar o regime democrático e suas instituições suscita questões sobre a necessidade de um compromisso mínimo que as empresas devem assumir** em suas atividades de moderação de conteúdo para a proteção do debate público. Já há, no direito internacional, o entendimento de que as empresas devem assumir compromissos com os direitos humanos, de modo que, na elaboração dos modelos de negócio, deve-se levar em consideração os riscos criados a direitos humanos consolidados. endo em vista que, atualmente, a maioria dos conteúdos que podem representar ameaças ao debate democrático acabam sendo regulados apenas por parâmetros gerais, entendemos que há espaço para se exigir compromissos das plataformas para a elaboração de políticas de integridade cívica eleitoral.

4

Ao mesmo tempo, **reconhece-se a impossibilidade de detalhar cada aspecto sobre eventuais políticas**, em razão das diferenças de cada plataforma quanto aos seus modelos de negócio, funcionalidades e arquiteturas e sob o risco de criação de regras muito rígidas que ameacem o direito à liberdade de expressão. A ideia que se propõe é a implementação de um olhar sistêmico na moderação de conteúdo, que foque em padrões de mudança mínimos ao invés de situações estáticas, tendo sempre em mente que o debate sobre a integridade cívica eleitoral na sociedade brasileira vai muito além das plataformas, perpassando pelo Poder Público, atores da sociedade civil, organizações internacionais e as (os) próprias(os) cidadãs(ãos).

5

**A proteção da integridade eleitoral requer um compromisso com elaborar políticas de conteúdo completas, precisas e funcionais.** Através da análise de dados do contexto eleitoral brasileiro, identificou-se quatro desafios principais a serem lidados pelas plataformas:

- **(i)** conteúdos sobre alegação de fraudes e questionamentos quanto à integridade eleitoral
- **(ii)** anúncios eleitorais
- **(iii)** perfis de candidatas(os) e de figuras públicas
- **(iv)** conteúdos de insurreição e rompimento com o processo democrático

A partir disso, foram indicados alguns parâmetros que devem ser abordados por políticas de moderação de conteúdo para cada um destes desafios:

### ALEGAÇÕES INFUNDADAS DE FRAUDE.

**Elaboração de termos de uso, políticas e protocolos específicos para discursos diretos e indiretos sobre alegações infundadas de fraudes e questionamentos quanto à integridade eleitoral**, tendo em vista que as narrativas não se restringem a períodos eleitorais específicos. O ideal é que a política seja um guarda-chuva capaz de proteger tanto a proteção mais ampla dos processos democráticos quanto períodos eleitorais específicos, com atenção para marcos fáticos e institucionais relevantes que se colocam poucos meses antes ou depois do período eleitoral oficial.

### ANÚNCIOS ELEITORAIS.

**Elaboração de medidas que regulam a circulação de anúncios eleitorais** e que sejam capazes de evitar aqueles que contenham desinformação eleitoral ou conteúdos que questionem a integridade eleitoral e a democracia; além de uma biblioteca de anúncios que permita rastreá-los.

### FIGURAS PÚBLICAS.

**Elaboração de regras que permitam uma moderação mais ágil e precisa de conteúdos que violem políticas sobre integridade eleitoral e cívica em perfis de candidatas(os) e figuras públicas em geral**, em razão do potencial de alcance e de legitimação de discursos não verídicos sobre eleições que esses perfis possuem. Ressalta-se que os protocolos não devem fazer diferenciação entre candidatas(os), sob o risco de prejudicar a igualdade de chances. Ao mesmo tempo, os perfis de figuras políticas e públicas pertencentes a grupos historicamente marginalizados devem receber atenção especial para assegurar sua segurança e proteção, considerando a probabilidade de violência e discurso de ódio direcionados a essas pessoas.

---

6

### ATAQUE À ORDEM DEMOCRÁTICA.

**Elaboração de regras que proíbam conteúdos que contenham declarações de incitação ou defesa de violência contra a ordem democrática ou contra a transmissão pacífica de poder,** levando em consideração contexto e risco de dano que a declaração possa ocasionar, sendo também sensíveis a cenários de pré-violência.

**Não basta se comprometer com boas regras, é necessário compromisso com a sua efetividade a partir de transparência e prestação de contas. Essa nova abordagem para a moderação de conteúdo só será possível se vier acompanhada de um sistema de transparência capaz de monitorar a implementação e a efetividade desse compromisso assumido pelas plataformas** por meio de métricas claras e definidas e pela apresentação acessível e organizada das políticas implementadas.

# INTRODUÇÃO: PLATAFORMAS PARA A CONVERSA SOBRE A VIDA CÍVICA

As eleições de 2022 evidenciaram que as plataformas digitais (chamadas, na linguagem regulatória, de *provedores de aplicação de internet*) se concretizaram como verdadeiras infraestruturas do debate público, carregando enorme porção das conversas entre as(os) cidadãs(os) sobre assuntos de interesse da vida cívica. Refletindo o contexto de intensa crise política que se constrói com força no país desde pelo menos 2013, tais intermediários da expressão digital, consequentemente, passaram a ser "plataformas" para uma série de campanhas e narrativas controversas sobre tais temas, frequentemente conectadas aos fenômenos da desinformação[3] e da violência política[4] dirigida a grupos socialmente vulnerabilizados. Estes acontecimentos tem sido subsídio para visões que apontam relação direta de causa e efeito entre o meio digital e a crise democrática que vivenciamos[5]. Por mais que existam conexões importantes, essa visão passa ao largo de explicar uma série de dinâmicas importantes[6].

3 O termo "*fake news*" ou desinformação é uma alcunha simplificadora para um fenômeno que reflete a mudança como as sociedades produzem, circulam e consomem informação política. O advento da internet como meio de comunicação política permitiu que qualquer indivíduo com conectividade se tornasse potencialmente um emissor de comunicação em massa, o que faz com que os novos conteúdos produzidos se distanciem daqueles elaborados sob o imperativo do jornalismo profissional, de modo que as informações emitidas nem sempre têm como compromisso a busca por objetividade. Esse novo ambiente acaba por enfraquecer "ciclos de checagem da realidade", impulsionados pelo jornalismo profissional, em favor de ciclos de "retroalimentação de propaganda", no qual a informação circula a partir de uma lógica político-partidária. A dinâmica da desinformação, assim, extrapola componentes de manipulação e de crise entre verdade e mentira, e envolve a própria forma como os indivíduos se relacionam com informação, em um processo de comunicação em rede no qual a autonomia desse indivíduo de produzir e compartilhar conteúdo ganha escala significativa. Francisco B. Cruz, Heloisa Massaro, Thiago Oliva, Ester Borges. Internet e eleições no Brasil: diagnósticos e recomendações. InternetLab, São Paulo, 2019. Disponível em <http://www.internetlab.org.br/wp-content/uploads/2019/09/policy-infopol-26919_4.pdf>.

4 Durante as eleições de 2020 e 2022, o projeto MonitorA monitorou perfis de candidatas no Twitter, Youtube, Instagram e Facebook, avaliando postagens, comentários de usuários, e outras interações. Na edição de 2020, identificou-se que a violência política na internet se direciona de forma específica a determinados grupos sociais marcados por gênero, raça, sexualidade, geração, etc., impactando especialmente o exercício da vida política de mulheres, pessoas negras, idosas e LGBTQIAP+. Na edição de 2022, por sua vez, o projeto voltou-se ao questionamento de como garantir que eleitoras e eleitores possam demonstrar descontentamento em relação às candidaturas na internet e, ao mesmo tempo, garantir a segurança e integridade das(os) candidatas(os) e do processo democrático, elaborando recomendações para diversos atores, dentre eles, as plataformas. Os dois relatórios do projeto MonitorA estão disponíveis em: https://monitora.org.br/relatorios/

5 BRITO CRUZ, Francisco (coord.); MASSARO, Heloisa; OLIVA, Thiago; BORGES, Ester. Internet e eleições no Brasil: diagnósticos e recomendações. InternetLab, São Paulo, 2019.

6 A crescente centralidade da comunicação digital e, sobretudo, das mídias sociais, modificou, de fato, a dinâmica da comunicação em massa. Não é possível afirmar, todavia, que problemas como a desinformação surgiram desse novo cenário. Muito antes do advento da internet a circulação de informações falsas já era instrumento de influência política ou ganho econômico. Mans, M. (junho, 2018). A Era da Pós Verdade. Revista .BR, ed. 14, ano 9, pp. 5-11. Disponível em <https://www.nic.br/media/docs/publicacoes/3/revista-br-ano-09-2018-edicao14.pdf>. ORTELLADO, Pablo; RIBEIRO, Marcio Moretto. Polarização e desinformação online no Brasil. Democracia Abierta, 23 out. 2018. Disponível em: <https://www.opendemocracy.net/pt/polariza-o-e-desinforma-o-online-no-brasil/>.

Essa não é apenas uma preocupação da opinião pública, mas de instituições dedicadas a cuidar do processo eleitoral. Como uma forma de contornar essas situações, por exemplo, o Tribunal Superior Eleitoral (TSE) passou a dedicar uma assessoria específica sobre o tema e a buscar cooperação com tais empresas com vistas a implementar mecanismos de monitoramento e combate de situações de abuso[7]. Sob outra perspectiva, o Congresso Nacional tem reiteradamente se dedicado ao tema, tanto a partir de sua proposta para a regulação de grandes plataformas digitais[8] como na sua abordagem de reformas na legislação eleitoral.

Todas estas preocupações trazem para o centro das discussões como essas empresas projetam a arquitetura de seus espaços digitais e, ainda, como suas escolhas delimitam a forma como a expressão terá lugar no meio digital. O conjunto de tais escolhas é o que o jargão técnico tem apresentado como o debate sobre "moderação de conteúdo"[9]. Esse é o termo empregado para nos referirmos às regras e aos procedimentos e sistemas usados pelas plataformas para remover, limitar alcance e rotular conteúdo, assim como suspender ou remover contas. São escolhas empresariais no âmbito da "moderação de conteúdo" que poderão privilegiar ou mitigar determinadas formas de expressão, elevando ou reduzindo a exposição de usuários a esta ou aquela "narrativa", por exemplo. **Desta forma, nosso objetivo com este documento é debater e propor sobre quais compromissos públicos sociedades democráticas como a brasileira devem demandar de tais empresas no âmbito da moderação de conteúdo independentemente do que está definido pela regulação.**

Dentro desse contexto, propomos aqui discutir como a moderação de conteúdo aciona uma série de desafios que precisam ser visitados como premissa deste e de outros trabalhos sobre o assunto. Seis grandes desafios já foram previamente identificados em outros trabalhos do InternetLab, quais sejam:

1

O mais básico deles é que a moderação que será realizada é **parte constitutiva dos modelos de negócios e da arquitetura das plataformas digitais** que, por sua vez, são muitas vezes projetadas para a maximização do compartilhamento de conteúdos e interação entre usuários(as). Desta forma, entrar nessa seara pode significar colocar em discussão a viabilidade econômica de determinados "produtos" ou de inovações de empresas de tecnologia.

7 Em 2022, por exemplo, o Twitter formulou a "Política de Informações enganosas e Integridade Cívica" como uma forma de contingenciar postagens que insinuavam fraudes eleitorais. O TSE, por sua vez, firmou memorandos de cooperação com diversas empresas, como o Google, o Facebook, o Whatsapp e o TikTok, a partir do Programa Permanente de Enfrentamento à Desinformação da Justiça Eleitoral (PPED), de modo a estabelecer um patamar para uma atuação mais ativa das plataformas no acesso de informações oficiais e verídicas.

8 O Projeto de Lei nº 2630/2020, conhecido como "PL das Fake News" visa, dentre outras coisas, regular a questão da disseminação da desinformação nas redes sociais. Disponível em <https://www.camara.leg.br/propostas-legislativas/2256735>.

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

2

Por serem serviços globais e oferecidos a uma massa de usuários(as) demográfica, cultural e politicamente diversa, tais medidas acabam afetando pessoas em jurisdições diferentes, nas quais as estruturas jurídicas e regulatórias nacionais divergem, de forma que a moderação de conteúdo implica inevitavelmente desacordos sobre os limites do discurso e as providências que lhes é direcionada. **A tarefa de discutir tais parâmetros é, portanto, complexa e implica soluções de compromisso entre formas de expressão, valores e crenças que por vezes parecem inegociáveis sob determinadas visões.**

3

Por arbitrar decisões sobre a expressão humana em escala industrial e global, a moderação de conteúdo oferecida por plataformas de internet é uma atividade que afeta continuamente uma série de valores democráticos e direitos humanos. Em razão do grande volume de conteúdos nas plataformas, torna-se impossível emitir decisões de moderação de conteúdo que sempre interpretem com precisão contextos, emoções e intenções dos usuários, seja porque é impossível coletar mais informações de cada caso em escala humana, seja porque há a implementação, pelas empresas, de ferramentas de inteligência artificial que são determinantes na decisão sobre visibilidade e disponibilidade de conteúdo. Nesse cenário, a moderação de conteúdo, muitas vezes, acaba por cometer erros e abusos à liberdade de expressão de seus usuários ou reforçar discursos discriminatórios e violentos ilegais, relacionados, inclusive, à democracia e às eleições. Por outro lado, ditar cada aspecto das atividades de moderação de conteúdo também é implausível, pois, além de não ser possível prever todas as eventuais manifestações violadoras de direitos humanos e valores democráticos, poderia restringir excessivamente o discurso de usuários na plataforma. **O desafio se põe, assim, justamente em encontrar soluções que garantam, ao mesmo tempo, a manutenção de um ecossistema sadio para comunicação e interação e os direitos dos usuários nesses espaços públicos.**

4

A construção de um ecossistema digital aberto e democrático também esbarra na questão do *spam*, fraudes e conteúdo enganoso. O *spam* - sigla em inglês para "envio de publicidade em massa" - e outros conteúdos fraudulentos ou que representem insegurança aos usuários têm uma grande presença no ambiente digital. No contexto eleitoral, por exemplo, **pesquisas** indicaram que, em 2018, 26% dos brasileiros com mais de 16 anos receberam mensagens em aplicativos de mensageria sobre política de números desconhecidos. Também se **identificou** que, para além de peças publicitárias simples com o intuito de tornar o candidato conhecido, houve também o disparo de mensagens vinculadas à polarização política subjacente à disputa eleitoral. Ainda que em várias situações tais conteúdos possam ser indesejados, não necessariamente eles se configuram como ilegais. Visto como conteúdos individuais, peças de *spam* podem ser totalmente inofensivas ou protegidas pelo direito à liberdade de expressão, por exemplo. Todavia, nos casos em que esses materiais dominam a experiência de um usuário em

uma rede social ou são enquadrados como fraudulentos, é necessário que as plataformas atuem de maneira ágil na sua remoção, de maneira a assegurar serviços minimamente seguros (em relação a fraudes e crimes) e úteis nas finalidades que se propõe. Para isso, geralmente os provedores de redes sociais aplicam políticas que lhes dão alguma discricionariedade sobre o que pode ser considerado spam ou conteúdo enganoso - especialmente para que agentes maliciosos ou fraudadores busquem utilizar brechas nas políticas para continuar a espalhar *spam*. **Assim, propostas sobre moderação de conteúdo devem ter em mente essa característica presente em ambientes digitais, de modo que deve-se evitar aquelas que visam engessar as suas atividades, sob pena de tornarem os serviços de provedores de aplicação inviáveis e inúteis**.

5

A moderação de conteúdo, desse modo, não se restringe a uma série de decisões individuais sobre retirada ou manutenção de conteúdos em plataformas de redes sociais, mas envolve vasta quantidade de fatores, como utilização de ferramentas de inteligência artificial, sinalização de publicações, regulação de spams, discussão sobre os limites da liberdade de expressão, a dinâmica dos usuários nas plataformas, parcerias com agências de checagem de fatos e assim por diante. Para que a moderação de conteúdo consiga abordar todas essas temáticas, é necessário que ela seja visualizada não mais como um modo de correção de erros em decisões individuais sobre remoção ou conservação de discursos, mas como um sistema que volta o seu olhar para o desenho institucional das plataformas que administram conteúdos em massa. **A moderação de conteúdo, assim, deve focar em sistemas, não em casos individuais; em padrões de mudança ao invés de situações estáticas.**

6

O olhar sistêmico sob a moderação de conteúdo só é possível se ele vier acompanhado, também, da transparência. **O debate público sobre moderação de conteúdo carrega uma tensão constante entre transparência e segurança de serviços e usuários**(as). Se, por um lado, a transparência dos modelos de moderação de conteúdo mostra-se insuficiente para legitimar as decisões privadas das empresas sobre o conteúdo produzido por seus usuários e usuárias, redundando frequentemente em situações de insatisfação e irresignação; por outro, existem motivos razoáveis para que determinadas informações sobre a moderação sejam reservadas para que sua eficiência seja maximizada em face do uso estratégico por aqueles(as) que tem como objetivo burlar a forma como as regras são aplicadas. O argumento da segurança, todavia, não deve ser utilizado como justificativa para a divulgação de relatórios de transparência pouco detalhados sobre as providências tomadas em resposta à violação dos termos de uso das plataformas e que não possuem informações que permitam compreender o cenário em cada país ou região nacional. Dessa maneira, a discussão sobre transparência encontra o embate sobre a escolha de métricas que sejam, ao mesmo tempo, inteligíveis aos indivíduos, mas que não divulguem as estratégias adotadas pelas empresas para tornarem as suas plataformas ambientes seguros de ataques fraudulentos e criminosos.

Estes desafios ganham nova roupagem a partir do posicionamento destes serviços como infraestrutura do debate público a respeito de processos democráticos e eleitorais. O aumento do uso estratégico da comunicação digital para a propagação de campanhas que visam a deslegitimação do regime democrático e de suas instituições, como alegações de fraude sem comprovação e incitações de insurreição ou rompimento com o processo democrático, por exemplo, suscitam a questão de como esta atividade de elaborar e aplicar regras de conteúdo e comportamento em comunidades online deve lidar com estes conteúdos e comportamento.

Para além de uma relação de causa e efeito entre a expressão de usuários na internet e a erosão democrática, a questão chave é refletir se de saída há um compromisso mínimo que tais empresas devem assumir em suas atividades de moderação de conteúdo para contribuir para que o debate público sobre a vida cívica e democrática ocorrido em suas plataformas não seja instrumentalizado por operações de propaganda antidemocrática e violência política, por exemplo. **É uma questão, portanto, sobre como esta atividade já existente e tão constitutiva do negócio de empresas detentoras de plataformas digitais deve lidar com discursos enganosos, conspiratórios ou violentos que são utilizados para fins da abolição do Estado Democrático de Direito, dos direitos políticos de cidadãos e cidadãs em sociedades democráticas ou da deslegitimação de processos eleitorais e cívicos.**

A discussão sobre a necessidade de um compromisso de empresas com direitos humanos tem amparo em instrumentos de direito internacional que devem ser considerados, como os *Princípios Orientadores da ONU sobre Empresas e Direitos Humanos*[10]. Inclusive, segundo o enfoque dado por tais princípios orientadores, a própria elaboração dos modelos de negócio deve levar em consideração os riscos criados a direitos consolidados perante direito internacional dos direitos humanos, por exemplo.

Na prática, a discussão sobre este tipo de compromisso se ramifica em muitas outras: haveria a necessidade de adaptação ou de criação de novos componentes em padrões de comunidades de redes sociais para lidar com essas problemáticas relacionadas à integridade eleitoral ou a proteção da democracia? Se sim, o que cabe às plataformas no período eleitoral e fora dele? Até que ponto se espera que os(as) próprios(as) usuários(as) decifrem a credibilidade do conteúdo online ou sejam responsáveis pelas denúncias que poderão (ou não) levar à remoção de conteúdos? Quais podem ser elementos para detectar discursos que devem ser permitidos ou não? Se já existem regimes distintos para pensar em pessoas comuns e figuras públicas, cabe o mesmo regime e regras para candidatos(as) e figuras públicas?

**Este estudo busca apontar um caminho para começar a responder tais perguntas, constituindo um ponto de partida sobre a necessidade de um compromisso com a democracia por parte das grandes plataformas de internet que realizam moderação de conteúdo e sobre quais os pontos tais devem ser enfrentados na elaboração deste compromisso.**

10 Os Princípios Orientadores da ONU sobre Empresas e Direitos Humanos entendem que, para que as empresas cumpram a sua responsabilidade de respeitar os direitos humanos, é necessário que (i) elas tenham um compromisso político que respeitem os direitos humanos; (ii) um processo de due diligence de direitos humanos para identificar, prevenir, mitigar e prestar contas de como elas lidam com seus impactos sobre os direitos humanos e (iii) processos para permitir a reparação de quaisquer direitos humanos que sejam prejudicados por suas atividades. Disponível em <https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf>. O tema é abordado sob a lente das plataformas digitais por Zuleta, Lumi; Jørgensen, Rikke Frank. Private Governance of Freedom of Expression on Social Media Platforms: EU content regulation through the lens of human rights standards. Sciendo, Volume 41 (2020) - Issue 1 (January 2020), p. 51-67. Disponível em <https://sciendo.com/article/10.2478/nor-2020-0003>.

# NECESSIDADE DE POLÍTICAS ESPECÍFICAS SOBRE INTEGRIDADE ELEITORAL

Ressalvadas as diferenças entre empresas e seus serviços, parte fundamental da atividade das plataformas de internet consiste na organização de seus usuários e usuárias e do conteúdo por eles(as) produzido. Isto, por sua vez, possui impacto em direitos humanos, como a liberdade de expressão, o livre desenvolvimento da personalidade e a soberania popular,[11] de modo que a primeira pergunta que surge, nesse cenário, é se tais serviços **devem ou não possuir políticas de conteúdo específicas abordando questões sobre processos cívicos e democráticos, como eleições e a transmissão pacífica de poder entre representantes eleitos(as).**

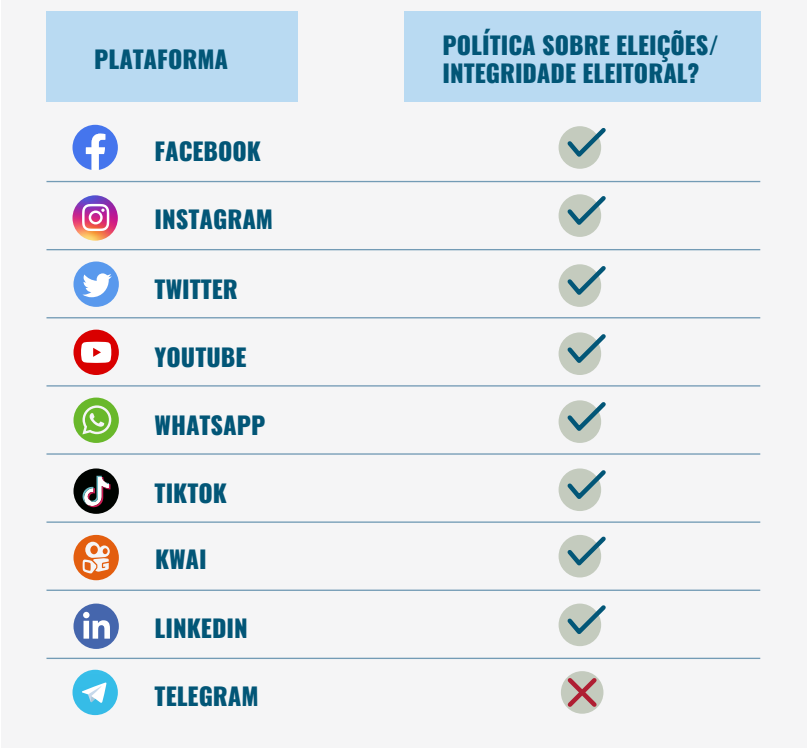
A defesa do ponto de vista de que as plataformas devem possuir políticas específicas sobre processos cívicos e democráticos compreende estas como atores importantes na dinâmica comunicacional sobre política, capazes, portanto, de serem mais do que apenas ferramentas para que grupos e indivíduos expressam-se, tomando decisões de arquitetura e design capazes de favorecer ou arrefecer determinadas narrativas.

## QUAIS SÃO AS POLÍTICAS DAS PLATAFORMAS SOBRE INTEGRIDADE ELEITORAL?

Como uma forma de compreender melhor como as plataformas têm lidado com essas questões, o InternetLab, durante as eleições de 2022, **mapeou e categorizou políticas existentes de moderação de conteúdo** e identificou as plataformas que já possuíam algum tipo de previsão sobre integridade eleitoral, conforme a tabela abaixo[12]:

11 Para a ONU, são parâmetros mínimos aqueles expressos na Declaração Universal dos Direitos Humanos e na Declaração da Organização Internacional do Trabalho (OIT) sobre os Princípios e Direitos Fundamentais no Trabalho.

12 A pesquisa foi realizada durante o primeiro semestre de 2022, com revisão entre os meses de janeiro e maio de 2023. A lista de redes sociais escolhidas para esse levantamento se baseia em duas pesquisas: uma de autoria da Mobile Time, Uso de Apps no Brasil - Jun 2022, e outra de autoria do InternetLab, Vetores da Comunicação Política em Apps de mensagem. Ambas as pesquisas consideram Facebook, Instagram, Twitter, Youtube, WhatsApp, TikTok, Kwai, Linkedin e Telegram algumas das redes sociais mais utilizadas pelos brasileiros.




À exceção do Telegram, as outras plataformas mencionadas possuíam algum tipo de política direcionada à integridade eleitoral.

Verifica-se, dentre as políticas, uma predominância de regras relacionadas à desinformação e eleições. Plataformas como **Twitter**, **Youtube**, **Facebook**, **TikTok**, **Kwai** e **Linkedin** possuíam em comum a proibição de conteúdos enganosos sobre local e métodos de votação, participação de candidatos em uma eleição e registro de eleitores. Em alguns casos, como no Twitter, TikTok, Kwai e Youtube, as políticas também previam regras contra alegações de fraude eleitoral. O detalhamento quanto a regulação desse tipo de discurso, no entanto, varia de rede social para rede social. No caso do Twitter e do Kwai, por exemplo, alegações de fraude eleitoral estavam inseridas em uma categoria específica de condutas proibidas, relacionada à informações falsas sobre processos cívicos. No Youtube, por sua vez, a fraude eleitoral também era categorizada como uma conduta proibida específica. No TikTok, por fim, alegações de fraude eleitoral apareciam como um exemplo na categoria de "informações enganosas e prejudiciais". Na Meta, conteúdos sobre fraude eleitoral apareciam tanto como uma categoria específica na política de coordenação de danos e **incitação ao crime**, quanto na política de violência e **incitação**. Por fim, no LinkedIn, afirmações sobre fraude eleitoral apareciam como um exemplo dentro da política sobre conteúdo falso ou **enganoso**.

Outro assunto bastante abordado nas políticas das plataformas é a questão dos anúncios eleitorais. O **Facebook**, o **Instagram**, o **Twitter**, o **TikTok**, o **Youtube**, o **Kawai** e o **Linkedin** possuíam políticas direcionadas à propaganda eleitoral. O Youtube limitava-se a indicar as condições que devem ser preenchidas para que seja possível a veiculação de um anúncio sobre um candidato ou partido político. No caso do Facebook e do Instagram, para além do estabelecimento de um processo de autorização para veiculação de anúncios, também havia a previsão

(i) de uma obrigação de rotulagem do anúncio como "propaganda eleitoral" e

(ii) de uma Biblioteca de Anúncios, na qual os anúncios eleitorais ficariam armazenados por 7 anos.

O Twitter, o TikTok e o Linkedin, se distanciando das demais plataformas, proibiam completamente a propaganda eleitoral em suas plataformas.

Por outro lado, foi possível perceber na análise que quase não há políticas direcionadas aos períodos antes e após as eleições. O **Twitter**, em sua política de integridade cívica, foi o único que apresentou uma previsão no sentido de proibir "informações enganosas sobre desfechos de atos cívicos" e "afirmações contestadas que podem minar a própria fé no processo" durante o período cívico e enquanto a plataforma considerar necessário.

Na mesma linha, no tocante à manutenção de conteúdos de interesse público na plataforma, o que acaba por englobar, também, a discussão da diferenciação de discursos de usuários comuns e de candidatos e políticos, e o TikTok possuíam previsões no sentido: o Twitter indicava que, por questões de interesse público, poderia manter discursos especificamente de funcionários eleitos do governo; já o TikTok possuía uma política específica sobre contas de governos, políticos e partidos políticos (denominada, em inglês, pela plataforma, como GPPA), na qual aplicava diferentes restrições de conta e penalidades em razão do "papel que essas contas de interesse público desempenham nos processos cívicos e na sociedade civil".

Destaca-se, também, que as políticas do WhatsApp e Telegram, por serem aplicativos de mensageria, tendem a divergir das políticas das demais plataformas. Ainda assim, no WhatsApp havia a previsão de regras relacionadas ao combate ao uso abusivo da plataforma em contexto eleitoral, como banimento de mensagens em massa; limites mais rígidos para mensagens virais, acompanhadas de etiquetas de encaminhamento; ferramentas para verificar as informações recebidas por encaminhamento e parcerias com agências de checagem de fatos. Já o Telegram, possuía apenas em seus termos de uso a proibição do envio de spam, não havendo políticas específicas relacionadas a desinformação ou a integridade cívica e eleitoral.

O que se observou de semelhante entre as plataformas é que, apesar da existência de uma ou outra categoria específica para processos cívicos e, em especial, para as eleições, a maioria das publicações que podem representar ameaças ao debate democrático ou aos direitos humanos- até por estarem em diálogo com assuntos mais gerais, como violência, discurso de ódio e bullying - acabam sendo regulados por moderadores de conteúdo através de parâmetros mínimos estabelecidos pelas plataformas para balizar outros comportamentos. Nesse sentido, é possível, por exemplo, que uma publicação que incita violência a uma instituição democrática seja removida pela aplicação de regras mais genéricas relacionadas a ameaças e incitação de violências e não pela proteção do ambiente cívico seguro.

Todavia, estas regras mais gerais não abordam todos os comportamentos que podem causar danos a processos eleitorais ou ao Estado Democrático de Direito. De alguma forma, tais regras não apresentam todas as ferramentas necessárias para que os controladores das plataformas e seus moderadores consigam eficientemente e de maneira transparente desempenhar a proteção destes processos e valores. Por isso, é importante em primeiro lugar reconhecer que se faz necessário o estabelecimento de políticas e regras específicas para esse quesito.

## CUIDADOS COM A COMPARAÇÃO ENTRE PLATAFORMAS E NA RIGIDEZ DE SUAS REGRAS

Reconhecemos que é implausível cristalizar cada aspecto sobre essas políticas. Ainda que cada uma das plataformas carregue a característica semelhante de dar contornos ao debate público, sublinhamos que cada uma delas têm arquiteturas, modelos de negócio e funcionalidades diferentes. Enquanto redes sociais como o Facebook e o Twitter, por exemplo, detém espaços que se propõem a ser abertos para abrigar discussões, nas quais a participação de usuários(as) é pública e plenamente visível para as empresas administradoras desses espaços, aplicativos de mensagem, como o WhatsApp, possuem uma dinâmica de funcionamento diversa, devido à tecnologia da criptografia e seu caráter privado. Assim, se, de um lado, certas medidas podem parecer adequadas e factíveis de serem aplicadas por determinadas plataformas, como, por exemplo, a moderação do discurso publicado por usuários(as); estas mesmas medidas não são possíveis de serem implementadas em outros contextos em razão da lógica do serviço prestado pela plataforma, sob o risco de desnaturalizar a particularidade de cada provedor, o que seria danoso, inclusive, às e aos usuários em geral. O exemplo do WhatsApp e serviços de mensageria é o mais latente nesse sentido, pois uma política de moderação de conteúdo nesses casos tiraria a privacidade dos seus (suas) usuários(as). Dessa forma, a esses aplicativos são recomendadas políticas de moderação de comportamento

Ao considerarmos, ainda, a volatilidade da intersecção entre eleições e internet, indicar a inclusão de regulação de conteúdos específicos nas políticas pode levar a uma rigidez nas regras, ameaçando o direito à liberdade de expressão. Essa ameaça em um país marcado pela desigualdade econômica e social como o Brasil, facilmente, pode se tornar também a fragilização da liberdade de organização e manifestação de grupos que são historicamente marginalizados. O que significa que tomar decisões e ditar mudanças no modo como ocorre o funcionamento de plataformas, que são também constitutivas do modo como a população brasileira se comunica e faz política, necessita levar em consideração os diferentes grupos sociais, como se localizam e como foram historicamente postos dentro ou fora do cenário político institucional. É crucial, desse modo, que, ao pensar na sustentação da democracia, tracemos também formas de garantir a pluralidade de discursos, ampliando o leque do que se compreende como comportamentos nocivos à integridade do debate público, não frisando, portanto, conteúdos em si, mas o sistema no qual os conteúdos se produzem.

## O QUE É FUNDAMENTAL?

No contexto em que as redes sociais desempenham um papel fundamental como um primeiro filtro para o debate político, e diante da crescente troca de ideias, informações e opiniões políticas nesses espaços virtuais, **torna-se imprescindível que as plataformas estabeleçam regras internas específicas sobre processos cívicos e democráticos**. O objetivo dessas regras deve ser garantir que as conversas e campanhas ocorridas em suas plataformas estejam em consonância com o respeito aos direitos fundamentais.

**Além disso, é importante ressaltar que essas normas específicas, assim como as regras gerais já existentes na maioria das plataformas que tratam de discurso de ódio,**

**desinformação, manipulação de informações e comportamento abusivo, não devem ser aplicadas apenas durante períodos de grandes eventos como eleições e plebiscitos. Pelo contrário, é essencial criar um ambiente permanentemente propício ao debate político íntegro, protegendo os direitos e a dignidade de todos os usuários.**

Ao estabelecer tais regras, as plataformas devem buscar alcançar um equilíbrio delicado entre a liberdade de expressão e a necessidade de proteção contra abusos. Além disso, a transparência é um valor fundamental que as plataformas devem adotar em relação às suas políticas e regras, fornecendo informações claras sobre como são tomadas as decisões relacionadas à moderação de conteúdo político-eleitoral. Essa transparência é essencial para que os usuários possam compreender como as plataformas estão moldando o debate político e para que tenham autonomia e proteção, independentemente de serem figuras públicas, políticos ou cidadãos comuns.

Portanto, é essencial que as plataformas atuem como catalisadores do debate público íntegro, promovendo a diversidade de opiniões e garantindo que as conversas ocorram em um ambiente respeitoso e seguro. Ao estabelecer regras claras e transparentes, essas plataformas podem desempenhar um papel significativo na construção de uma esfera pública digital democrática e inclusiva.

# COMO FAZER? DESAFIOS A SEREM ENFRENTADOS EM POLÍTICAS DE INTEGRIDADE ELEITORAL E COMPROMISSO DEMOCRÁTICO

A relação entre plataformas, campanhas políticas, direito eleitoral e manutenção da integridade do debate cívico abre alguns tópicos que pretendemos observar nas políticas que já existem hoje. Já brevemente indicados acima, trazemos, na tabela abaixo, a visualização destes:

## POLÍTICAS DAS PLATAFORMAS DURANTE O PERÍODO PRÉ-ELEITORAL E ELEITORAL DE 2022

| PLATAFORMA | ALEGAÇÃO DE FRAUDES | ANÚNCIOS | MODERAÇÃO DIFERENCIADA DE CONTEÚDO DE CANDIDATOS E FIGURAS PÚBLICAS | INSURREIÇÃO E ROMPIMENTO COM A ORDEM DEMOCRÁTICA |
|---|---|---|---|---|
| FACEBOOK | ✓ | ✓ | ✗ | ✓ |
| INSTAGRAM | ✓ | ✓ | ✗ | ✓ |
| TWITTER | ✓ | ✓ | ✓ | ✓ |
| YOUTUBE | ✓ | ✓ | ✗ | não especifica o suficiente* |
| TIKTOK | ✓ | ✓ | ✓ | não especifica o suficiente* |
| KWAI | ✓ | ✓ | ✗ | não especifica o suficiente* |
| LINKEDIN | ✓ | ✓ | ✗ | não especifica o suficiente* |

✓ (sim)
✗ (não)

* As plataformas que tiveram classificação como "Não especifica o suficiente" possuem políticas relacionas a manutenção da integridade do debate público durante as eleições e a inibição de atos que possam impedir ou dificultar o processo eleitoral, porém não possuíam em 2022 previsões relacionadas a insurreições após a divulgação dos resultados.

Como dito na legenda da tabela, as plataformas que tiveram classificação como "Não específica o suficiente" possuíam entre junho de 2022 e janeiro de 2023 políticas relacionas a manutenção da integridade do debate público durante as eleições e a inibição de atos que possam impedir ou dificultar o processo eleitoral, porém não possuem previsões relacionadas a insurreições após a divulgação dos resultados. A política contra desinformação em eleições do **Youtube**, por exemplo, proíbe conteúdos que inicitem "o público a interferir em processos democráticos". Isso inclui "a obstrução ou interrupção do processo eleitoral"; já a política eleitoral do **Kwai** proíbe conteúdos de "intimidação e incitação ao boicote às eleições" e define esse tipo de conteúdo como aqueles que "visem difundir e/ou gerar um clima de violência no processo eleitoral, fazendo com que as pessoas se abstenham de exercerem seu direito de voto" e que "promovam e incitem ações violentas para obstruir as atividades eleitorais". O LinkedIn se limita a dizer em sua **política para comunidades profissionais**) que usuários não devem compartilhar "conteúdo para interferir ou influenciar indevidamente uma eleição ou outro processo cívico". E o TikTok não possuía políticas relacionadas a insurreição e rompimento com a ordem democrática durante as eleições. Na **política de integridade cívica e eleitoral**, diz que não permite "conteúdo impreciso, enganoso ou falso que possa causar danos significativos a indivíduos ou à sociedade, independentemente da intenção". Em atualização em 2023, contudo, a plataforma acrescentou que a política inclui: "informações enganosas sobre como votar, registrar-se para votar, requisitos de elegibilidade dos candidatos, sobre os processos para contar votos e validar eleições e sobre o resultado final de uma eleição.".

Os aplicativos de mensagem, como mencionado anteriormente, lidam com comunicações privadas, e devido à sua própria arquitetura, possuem menos poder para moderar conteúdo, apesar de possuírem poder para modelar comportamentos. Portanto, existem outras políticas que abordam desafios relacionados à integridade cívica e eleitoral, conforme listado na tabela abaixo:

| PLATAFORMA | ANTI-SPAM | ENVIO DE MENSAGEM EM MASSA | LIMITES DE ENCAMINHAMENTO DE MENSAGENS | ENVIO DE MENSAGENS MEDIANTE A PAGAMENTO |
|---|---|---|---|---|
| WHATSAPP | ✓ | ✓ | ✓ | ✗ |
| TELEGRAM | ✓ | ✗ | ✗ | ✗ |

✓ (sim)
✗ (não)

Percebe-se que

(i) algumas plataformas não tratam de nenhum desses temas e que

(ii) nenhuma delas possui previsão sobre todos os desafios mapeados.

Mesmo nos casos em que há alguma previsão, foram observadas lacunas, o que pode estar relacionado a diversos fatores, como surgimento de novas dinâmicas nos períodos eleitorais recentes, a ausência de adaptação ao contexto eleitoral brasileiro ou mesmo falta de medidas que impeçam que determinados movimentos nas plataformas evoluam para situações graves. Esse cenário demanda, assim, a elaboração de políticas de integridade cívica e eleitoral que tenham um olhar mais atencioso para essas dificuldades.

## ESTUDO DE CASO: O CONTEXTO ELEITORAL BRASILEIRO DE 2022

No contexto eleitoral brasileiro, as plataformas de redes sociais tiveram papéis diferentes em situações políticas que se desenharam a partir de conjunturas diversas. O mesmo aconteceu com as medidas tomadas pelo poder público e pelo Tribunal Superior Eleitoral no intuito de regulamentar o uso das plataformas para fins de campanhas eleitorais.

Até 2018, considerando a experiência norte-americana, o foco de atenção da regulamentação das autoridades eleitorais estava no micro direcionamento de peças de publicidade eleitoral a partir do impulsionamento de conteúdo e a construção oficial de campanhas online. A partir do que ocorreu nas eleições daquele ano, contudo, observa-se que, apesar de existirem investimentos oficiais altos em impulsionamento de conteúdo por campanhas eleitorais e uma relevância de candidatos(as) em espaços *online*, havia uma outra novidade relacionada a redes sociais: a grande capacidade da arquitetura de comunicação das plataformas de aproximar voluntários(as) e apoiadores(as) independentes que, por seu turno, geram engajamentos espontâneos para candidaturas[13]. Essa comunicação orgânica e descentralizada entre eleitores é legalizada e não se caracteriza como propaganda política, conforme a Resolução n° 23.610/2022 do Tribunal Superior Eleitoral (TSE)[14].

Em geral, a descentralização traz mais dinamismo para a comunicação política, expandindo narrativas de mobilização, porém complexifica a sua regulamentação e a responsabilização de comportamentos e/ou práticas de candidatos(as) e eleitores(as) que possam se tornar ameaças à integridade do debate público e ao bom andamento das eleições. Após o período eleitoral de 2018, **foram encontrados**, por exemplo, materiais que questionavam os resultados do primeiro turno da disputa eleitoral com base em informações falsas e que, apesar de não estarem explicitamente vinculados a um partido ou candidato(a), se posicionavam e mencionavam diretamente os(as) principais candidatos(as) à presidência da república e ao governo do estado, além de questionar o TSE.

A partir desses acontecimentos, constataram-se algumas lacunas tanto na legislação eleitoral, quanto nas políticas internas das plataformas de redes sociais e aplicativos de mensagem, uma vez que essas são, muitas vezes, o primeiro filtro pelo quais discursos se proliferam online. Citando especificamente o que cabe às políticas das plataformas, destacavam-se a ausência de

**(i)** um canal direto para denúncias, com categorias específicas relacionadas à integridade cívica eleitoral e

**(ii)** fiscalização de spam eleitoral.

13 O segundo relatório do projeto "Você na Mira", na qual se analisou as campanhas dos presidenciáveis durante o primeiro mês do período eleitoral de 2018, identificou que, apesar da existência de despesas com impulsionamento, sobretudo no Facebook e no Google, estas representavam apenas uma fração pequena dos gastos declarados à justiça eleitoral. Em face da desproporção entre os investimentos e a relevância de candidatos políticos nesses espaços, levantaram-se duas hipóteses: (i) a presença de "impulsionamento cruzado" entre candidatos e (ii) a existência de um engajamento orgânico por parte de apoiadores da candidatura.

14 "A publicação com elogios ou críticas a candidatas e candidatos, feitos por uma eleitora ou eleitor em página pessoal, não será considerada propaganda eleitoral. A repercussão desse conteúdo está autorizada, desde que não ocorra impulsionamento pago de publicações por parte do eleitor com a finalidade de obter maior engajamento". Disponível em <https://www.tse.jus.br/comunicacao/noticias/2022/Agosto/eleicoes-2022-confira-o-que-pode-e-nao-pode-na-propaganda-eleitoral>.

Tais lacunas não passaram, todavia, despercebidas pelas plataformas. Como uma forma de combatê-las, ou ao menos mitigá-las, algumas delas adotaram medidas como a **proibição total de anúncios durante o período eleitoral** ou, então, a **proibição de anúncios questionando a legitimidade das eleições**, além de terem estruturado canais de denúncias em parceria com o TSE[15].

Por mais que as novas regras e iniciativas sobre conteúdos e comportamentos tenham auxiliado na manutenção do debate público durante o período eleitoral[16], o ano de 2022 foi marcado por tensões online e offline. A partir de meados de novembro, após a derrota de Jair Bolsonaro nas urnas, apoiadores de sua campanha passaram a protestar em frente a quartéis do exército em diferentes regiões do país pedindo por uma intervenção militar que se baseava em uma interpretação do Artigo 142 da Constituição brasileira[17]. O dispositivo, segundo a narrativa, abriria margem para uma intervenção das forças armadas para a "restauração da ordem", sobrepondo-se aos ritos democráticos, caso houvesse apoio popular o suficiente. De acordo com pesquisa **Democracia digital: análise dos ecossistemas de desinformação no Telegram durante o processo eleitoral brasileiro de 2022**, realizada por parceiros do InternetLab, mensagens que disseminavam esse tipo de interpretação equivocada e outros tipos incitação a ataques ao Estado Democrático de Direito e à integridade do sistema eleitoral circularam em um ecossistema comunicacional multiplataforma pelo menos entre janeiro de 2022 e janeiro de 2023.

Este ecossistema comunicacional multiplataforma de propaganda e agitação antidemocrática operou a partir de ferramentas, tecnologias e serviços para a criação e disseminação de conteúdos em diferentes plataformas. Dessa forma, permitia que militantes e usuários(as) de diferentes níveis de dedicação e comprometimento com campanhas criassem e compartilhassem conteúdo de agitação antidemocrática (muitas vezes falsificado e/ou enganoso) em diferentes meios a baixo custo e de maneira tanto distribuída, como estratégica, até depois da posse do presidente recém eleito.

Notou-se, portanto, a ocorrência de dinâmicas e estratégias novas que não estão previstas em políticas claras por parte das plataformas e que os eventos de insurreição e rompimento com o processo democrático são alguns dos exemplos de narrativas que não foram e não são contempladas de forma eficiente nas políticas de uso e também na moderação de conteúdo das plataformas de mídias sociais, o que as torna, consequentemente, insuficientes para proteger a higidez e a integridade do debate público, mais do que do processo eleitoral estrito[18].

Todo esse contexto demonstra a necessidade da elaboração de políticas mais específicas e atentas para a promoção de direitos fundamentais e princípios democráticos, não restritos aos períodos eleitorais.

15 O Whatsapp firmou um memorando de entendimento com o TSE no qual se comprometeu a criar um canal de comunicação extrajudicial não vinculativo para a denúncia de conteúdos que veiculem desinformação relacionada ao processo eleitoral. O Facebook, o Instagram, o Google, o Youtube, o Telegram, o LinkedIn e o Kwai adotaram medidas semelhantes.

16 Antes do início do período eleitoral brasileiro de 2022, o Facebook informou que "atualmente, 99,7% das contas falsas que removemos do Facebook são excluídas antes mesmo de serem denunciadas, com uso de inteligência artificial. Também investigamos e interrompemos redes que, de maneira coordenada, utilizam contas falsas para influenciar o debate público". Disponível em <https://about.fb.com/br/news/2022/08/como-a-meta-esta-se-preparando-para-as-eleicoes-do-brasil-em-2022/>.

17 Segundo o artigo 142 da Constituição Federal do Brasil "[a]s Forças Armadas, constituídas pela Marinha, pelo Exército e pela Aeronáutica, são instituições nacionais permanentes e regulares, organizadas com base na hierarquia e na disciplina, sob a autoridade suprema do Presidente da República, e destinam-se à defesa da Pátria, à garantia dos poderes constitucionais e, por iniciativa de qualquer destes, da lei e da ordem".

18 Disponível em <https://www.oc.eco.br/wp-content/uploads/2022/07/Papel_das_plataformas_na_protec%CC%A7a%CC%83o_da_integridade_eleitoral_-_doc_sociedade_civil.pdf> e <https://www.conectas.org/wp-content/uploads/2023/02/Balanco-2_SAD_O-papel-das-plataformas-na-protecao-da-integridade-eleitoral.pdf>.

## ALEGAÇÃO DE FRAUDES E QUESTIONAMENTOS QUANTO À INTEGRIDADE ELEITORAL

Conteúdos que possuem narrativas de alegações de fraudes eleitorais sem nenhum tipo de comprovação e questionamentos infundados do sistema de votação têm sido registrados em plataformas digitais ao menos desde 2018. Durante as eleições daquele ano, identificou-se, no WhatsApp, uma série de conteúdos que questionavam os resultados do primeiro turno da disputa eleitoral com base em desinformações. Dentre os **exemplos de materiais encontrados**, sobressaiu-se um no qual alegava-se que não existiam votos nulos em eleições realizadas em urnas eletrônicas, de forma que o TSE deveria dar explicações à população brasileira quanto ao número de votos anulados na corrida presidencial. Na mesma linha, nos dias seguintes ao primeiro turno das eleições gerais de 2022, foram encontradas no Twitter narrativas recorrentes questionando a lisura do voto nordestino. Os tuítes analisados afirmavam que os resultados das eleições no Nordeste foram fraudados. Em muitos dos tuítes, alegava-se que uma suposta ausência de comemorações e de festas seria uma comprovação da fraude. Em **outros**, o número de deputados ou senadores bolsonaristas ou alinhados ao Bolsonaro eleitos era o suficiente para comprovar a fraude no resultado das eleições presidenciais.

Assim como no Twitter, **notou-se**, no aplicativo de mensagem Telegram, que essa temática esteve presente desde novembro de 2021, onze meses antes do primeiro turno das eleições de 2022, havendo a disseminação de variadas notícias falsas acerca das urnas eletrônicas e da legitimidade do processo eleitoral, conforme registrado na seção anterior. Esse cenário se prolongou até o segundo turno, com o surgimento de novos grupos para a organização de manifestações antidemocráticas motivadas por informações inverídicas, como ocorrência de fraude nas eleições e a possibilidade de manifestações de inconformismo com o resultado das urnas gerarem uma intervenção militar. Medidas do poder público, como um pedido do TSE de bloqueio de grupos que incitavam movimentos antidemocráticos **não impediu**, porém,

**(i)** a continuidade da circulação de mensagens negacionistas quanto ao resultado das eleições e à legitimidade da posse presidencial e

**(ii)** a articulação e mobilização para a participação ou apoio a atos antidemocráticos que ocorreram no dia 8 de janeiro de 2023 em Brasília.

Analisando-se as políticas das plataformas, é possível perceber a existência de regras que lidam com situações, por exemplo, de incitação de violência ou promoção de organizações, indivíduos ou atos de extremismo violento[19] e, também, de desinformação sobre o dia das eleições, como data, lugares, horários e métodos de votação, candidatos(as) que estão concorrendo e indivíduos que podem votar[20]. Todavia, na grande maioria delas, não há previsões específicas para narrativa de desconfiança sobre as urnas, instituições eleitorais ou contestação do pleito. Não existem também exemplos claros sobre essas situações que possam balizar a atuação de moderadores

19 As plataformas analisadas possuem proibições sobre incitação a atos violentos. Algumas delas, como as plataformas da Meta e Twitter, possuíam previsões mais abrangentes relacionadas às eleições e/ou processos cívicos que abarcavam o período pós-eleitoral. Outras se limitavam ao processo de votação de forma estrita. Destaca-se, porém, que, no caso do Twitter, a política de atividade prejudicial coordenada foi excluída durante a escrita deste trabalho, restando apenas as previsões da própria política de integridade cívica.

20 Todas as plataformas analisadas possuem previsões a respeito de desinformações e eleições.

de conteúdo, gerando qualquer tipo de previsibilidade para usuários(as) quanto a se o conteúdo passará por qualquer tipo de cerceamento, o que é prejudicial à integridade eleitoral.

Destaca-se, ainda, que, no geral, grande parte das políticas sobre eleições voltam-se somente para o período eleitoral, não havendo transparência e clareza sobre qual é a postura das plataformas no período pré e pós-eleições, que também são essenciais para a definição de como se dará a disputa entre candidatos e candidatas na corrida eleitoral.

Tendo em vista que os discursos sobre fraude eleitoral no Brasil têm fortalecido estratégias organizadas de descredibilização da democracia que podem levar a consequências materiais, como a deslegitimação de resultados de eleições livres e a organização de protestos pró-golpes de Estado, é fundamental que as políticas passem a abarcar previsões sobre alegações de fraudes e ataques à integridade eleitoral, uma vez que não são apenas uma disputa de narrativas normais ao jogo democrático. Ressalta-se, aqui, que o intuito na inclusão dessas previsões não é proibir a produção de conteúdos ou a organização de manifestações contra figuras políticas, atividades que são permitidas pelos direitos de liberdade de expressão e de manifestação, mas sim regular comportamentos que contestam a própria integridade democrática e incitam atos de violência, seja contra indivíduos, populações historicamente marginalizadas ou instituições. Com este quadro contextual em tela, apresentamos abaixo duas reflexões sobre como as políticas de conteúdo das plataformas podem tentar abordar diferentes facetas destes fenômenos tendo sensibilidade a estas diferenciações e proteções necessárias à liberdade de expressão.

## A) PONTOS FUNDAMENTAIS PARA POLÍTICAS DE INTEGRIDADE ELEITORAL E COMPROMISSO DEMOCRÁTICO

**As plataformas devem ter duas características em mente como pontos de partida para a elaboração de políticas sobre integridade eleitoral:**

**(i)** as narrativas sobre fraudes não são inerentes a uma eleição específica, mas persistem ao longo do tempo e

**(ii)** ganham mais força no período entre os turnos de votação e no período pós-eleitoral.

Vamos considerar um caso ocorrido nos serviços do Facebook, discutido no Comitê de Supervisão sob o número 2023-001-FB-UA[21]: em 3 de janeiro de 2023, após a posse de Luiz Inácio Lula da Silva como presidente do Brasil e quase três meses após o término das eleições nacionais, um vídeo foi publicado no Facebook com uma legenda em português. A legenda pedia para "sitiar" o Congresso Brasileiro como "a última alternativa" contra os adversários políticos. O vídeo mostrava parte de um discurso de um general brasileiro conhecido e apoiador do adversário eleitoral de Lula, convocando as pessoas a irem para as ruas e se dirigirem ao Congresso Nacional e ao Supremo Tribunal Federal. O vídeo também continha imagens de um incêndio na Praça dos Três

21 Disponível em: <https://www.oversightboard.com/news/539069631694711-oversight-board-announces-two-new-cases-and-upholds-meta-s-decision-in-the-sri-lanka-pharmaceuticals-case/>

Poderes, em Brasília, onde estão localizados a Presidência da República, o Congresso e o Supremo Tribunal Federal. O texto nas imagens incentivava a invasão e o cerco aos três poderes, além de exigir a divulgação do código-fonte das urnas eletrônicas do Brasil.

O vídeo recebeu mais de 18 mil reproduções e foi denunciado sete vezes entre 3 e 4 de janeiro, mas permaneceu na plataforma. Um moderador humano analisou o conteúdo após a primeira denúncia e considerou que estava de acordo com as políticas do Facebook. Após uma apelação do usuário, outro moderador manteve a decisão. No dia seguinte, cinco moderadores diferentes analisaram as outras seis denúncias e todos concluíram que o conteúdo estava de acordo com as políticas do Facebook. Posteriormente, o conteúdo foi escalonado para especialistas em políticas para revisão adicional.

Um usuário que denunciou o conteúdo recorreu ao Comitê de Supervisão da Meta, alegando que o conteúdo poderia incitar a violência, especialmente considerando o movimento de pessoas no Brasil que não aceitavam os resultados das eleições em 8 de janeiro de 2023. O Facebook considerou que as várias decisões de manter o conteúdo foram um equívoco e, em 20 de janeiro de 2023, removeu o conteúdo, aplicou uma advertência à conta do usuário que o publicou e limitou certos recursos da conta, impedindo a pessoa de criar conteúdo novo.[22]

O caso foi selecionado pelo Comitê de Supervisão da Meta sob a justificativa de analisar como a Meta modera conteúdo relacionado a eleições e como está aplicando o **Protocolo de Política de Crise em um "local designado temporariamente de alto risco"**, desenvolvido em resposta à recomendação do Comitê no caso **"Suspensão do ex-presidente Trump"**.

**O InternetLab contribuiu com informações sobre o caso**, considerando que a abordagem da plataforma se adequou à realidade brasileira. Em 16 de agosto de 2022, quando começou oficialmente o período de campanha eleitoral, a Meta divulgou medidas adotadas para as eleições de outubro, além das ações gerais voltadas para períodos eleitorais, como a biblioteca pública de anúncios e os rótulos em postagens pagas. Foram destacadas medidas específicas para a votação no Brasil, como a proibição de discursos de ódio e incitação à violência, a remoção de conteúdos que interferissem na votação, como informações incorretas sobre a data e o número dos candidatos, e a ampliação da colaboração com agências de checagem.

No entanto, surgiram complicações na implementação dessas medidas, pois cada país precisa de ajustes nos parâmetros e termos de uso de acordo com seu contexto, havendo até contradições dentro desses contextos.  Por um lado, a data escolhida para implementar e divulgar as novas políticas estava de acordo com os tempos do processo eleitoral brasileiro regulado por legislação federal. Por outro lado, manifestações políticas nas redes sociais não dependem do marco temporal institucional das campanhas, ocorrendo antes, durante e depois da votação.

22 Em 22.06.2023,  o Comitê de Supervisão da Meta (Oversight Board) divulgou sua decisão acerca desse tema. Na presente decisão, o Comitê revogou a resolução original da Meta, que mantinha a divulgação de um vídeo no Facebook. Tal desdobramento evidencia a importância de estabelecer uma estrutura abrangente de avaliação dos esforços de integridade eleitoral por parte da Meta. A decisão está disponível em: <https://www.oversightboard.com/news/539069631694711-oversight-board-announces-two-new-cases-and-upholds-meta-s-decision-in-the-sri-lanka-pharmaceuticals-case/>

Apesar de ter anunciado essas políticas no início das campanhas, a Meta não explicou por quanto tempo elas seriam aplicadas, levantando dúvidas sobre sua excepcionalidade e validade. Após as eleições, as políticas voltariam ao estado anterior? E qual seria o momento apropriado para essa transição, o anúncio oficial dos resultados, a diplomação dos eleitos ou a posse?

Essas questões destacam que, embora as novas normas contribuam para um debate público eleitoral mais íntegro, existem dinâmicas que não são abordadas pelas políticas. Por exemplo, conteúdos prejudiciais aos valores protegidos pelas políticas podem surgir após as eleições. Para lidar com esse dilema, é fundamental que as políticas se estendam além do período eleitoral legalmente regulado.

Nesse sentido, em primeiro lugar, entende-se que as políticas não devem ser estruturadas de maneira a serem aplicadas a eventos eleitorais específicos, sob o risco de se tornarem ineficazes em relação, por exemplo, a eventuais postagens que surjam alegando fraudes e desinformações a respeito de eleições anteriores ou futuras[23]. A ideia é que haja a criação e desmembramento de uma política de integridade cívica e democrática em dois "ramos" que lidariam tanto com a proteção mais ampla dos processos democráticos como, subsidiariamente, com as especificidades do momento eleitoral:

**(i)** O primeiro ramo, permanente, consideraria que o funcionamento da democracia não se restringe ao período eleitoral e estaria destinado a proteger tais processos independentemente do momento.

**(ii)** O segundo ramo seria circunscrito a um recorte temporal do período eleitoral em razão inclusive do tipo de conteúdo potencialmente violador próprio deste período e que decai após o pleito, por exemplo. Neste segundo caso, ao mesmo tempo, fica apontada a necessidade de adoção de período estendido em relação ao período eleitoral fixado em lei, acrescido da introdução de um novo conceito de "intervalos de amortecimento" nos quais a política poderia ser operada para além do período oficial.

No tocante especificamente à previsão de moderação de conteúdos sobre alegação de fraude, é importante que as plataformas **não se restrinjam a elaborar políticas genéricas que apenas repetem políticas de moderação de conteúdo básicas, como proibição de utilização de serviços para envio de spam ou proibição de promoção de violência ou de informações enganosas, sem adaptação para o contexto eleitoral**. Dentro desse contexto, algumas plataformas já possuem **políticas mais complexas** relacionadas a postagens que insinuam fraudes eleitorais no Brasil. Nesses casos, as plataformas passaram a considerar como proibidas informações enganosas sobre como participar de atos cívicos e seus desfechos e afirmações contestadas que podem prejudicar e deslegitimar o próprio processo eleitoral, indicando, inclusive, **exemplos** para os moderadores de conteúdo.

Para aplicativos de mensageria que utilizam a tecnologia de criptografia de ponta a ponta,

23 O Youtube, apesar de possuir uma política bem estruturada e com exemplos sobre desinformação em eleições, indica que a proibição contra conteúdos com alegações de fraudes, erros ou problemas técnicos em eleições se aplica somente as eleições presidenciais do Brasil de 2014, 2018 e 2022. Apenas nos Estados Unidos há a previsão de aplicação da política sobre qualquer eleição presidencial. Disponível em <https://support.google.com/youtube/answer/10835034?hl=pt-BR>.

dado que não é possível realizar a moderação do conteúdo em si, a inclusão de **ferramentas de pesquisa de fácil**[24] **acesso que permitem verificar a veracidade de mensagens encaminhadas com frequência pode ser uma alternativa para a mitigação de riscos quanto ao compartilhamento de narrativas falaciosas sobre as eleições e sobre a democracia.**

## B) DISCURSO INDIRETO SOBRE FRAUDES ELEITORAIS

Para além de conteúdos que explicitamente alegam sobre a ocorrência de fraudes eleitorais, é possível a existência, também, de conteúdos que não diretamente violem políticas sobre integridade eleitoral, mas que, de alguma forma, compartilham informações enganosas ou sensacionalistas de modo a desencorajar pessoas a votarem ou se posicionarem, por exemplo.[25] Esse tipo de discurso, que não fala diretamente sobre fraude, também merece atenção das plataformas. **A questão que se pretende levantar aqui é que nem sempre o problema está relacionado à remoção de conteúdo - que, muitas vezes, possui uma maior eficácia quando é evidente a sua violação às diretrizes da comunidade - mas pode estar ligado, também, a declarações implícitas envolvendo recomendação de conteúdo, buscas e hashtags, que são mais difíceis de serem identificadas.** Nesse sentido, é essencial que as políticas englobem regras sobre esses discursos indiretos e equipes treinadas a contextos locais sejam envolvidas na moderação desse tipo de conteúdo, pois, por vezes, grupos antidemocráticos se aproveitam da literalidade de determinadas políticas e passam a burlar o sistema com mensagem prejudiciais, mas indetectáveis por meio de técnicas automatizadas.

Uma problemática importante, também, relacionada ao discurso indireto sobre fraudes relaciona-se ao conceito de ecossistema multiplataforma de desinformação, que consiste na utilização de ferramentas, tecnologias e serviços para a criação e disseminação de desinformação em diferentes plataformas *online*. Essa dinâmica permite que usuários criem e compartilhem conteúdo falsificado e/ou enganoso em diferentes plataformas sem precisar recriar o conteúdo para cada mídia social. Levando em consideração que não há como uma plataforma exercer controle sobre a moderação de conteúdo de outra, mas que não é possível, ao mesmo tempo, ignorar a interconexão entre plataformas, medidas, como a disponibilização de um aviso aos usuários sobre a ausência de veracidade do conteúdo antes que se realize o compartilhamento para outra rede social e o trabalho conjunto entre plataformas, são formas de amenizar esse problema sem o estabelecimento de parâmetros inalcançáveis[26].

24 O WhatsApp oferece uma maneira simplificada para que usuários verifiquem a veracidade de mensagens que o aplicativo detecta como "encaminhadas com frequência". Ao tocar ou clicar em uma lupa exibida ao lado dessas mensagens, os usuários podem carregá-las em um navegador e encontrar notícias ou outras fontes de informação sobre o conteúdo recebido. Disponível em: <https://faq.whatsapp.com/518562649771533/?helpref=uf_share>

25 Exemplos dessas modalidades de mensagens e publicações podem ser encontradas na série "Democracia Digital – Análise dos ecossistemas de desinformação no Telegram durante o processo eleitoral brasileiro de 2022". No terceiro relatório deste projeto, que analisou o período pré-eleitoral de 2022, foram encontradas mensagens que ilustram esse discurso indireto: eram mensagens que levantavam suspeitas em relação às urnas, informações falsas sobre o seu funcionamento e com instruções para que o leitor supostamente não anulasse o seu voto. Indicando uma estratégia organizada, ainda que indireta, de descredibilização do processo eleitoral e, consequentemente, de deslegitimação dos seus resultados.

26 Em sua política de mídia sintética e manipulada, o Twitter indica, como uma possível sanção à sua violação, a colocação de um aviso antes que o usuário compartilhe ou curta o Tweet. Disponível em <https://help.twitter.com/en/rules-and-policies/manipulated-media>.

## ANÚNCIOS ELEITORAIS

Ligado a essas narrativas que prejudicam a integridade do debate público, encontra-se um segundo desafio, referente a anúncios eleitorais[27]. Desde 2017, é possível a veiculação de propaganda eleitoral em plataformas de mídia social, sendo proibida, porém, "a divulgação ou compartilhamento de fatos sabidamente inverídicos ou gravemente descontextualizados que atinja a integridade do processo eleitoral". Apesar dessa previsão, as políticas sobre anúncios eleitorais em plataformas regulam apenas sobre o procedimento para que haja autorização para um anúncio circular, não possuindo medidas para os cenários nos quais esses anúncios contenham desinformação eleitoral ou conteúdos que questionem a integridade eleitoral. A ausência de parâmetros, nesse caso, é ainda mais sensível, dado que as plataformas lucram com essas publicações monetizadas.

Por mais que não seja possível controlar toda e qualquer postagem relacionada às eleições em redes sociais, é necessário que as políticas prevejam fiscalizações e restrições mais intensas em datas marcantes para eleições. Para as eleições de 2022 nos Estados Unidos, por exemplo, a Meta **anunciou** que passaria a proibir anúncios políticos, eleitorais e sociais durante a última semana da campanha eleitoral, sob a justificativa de que não haveria tempo suficiente para as e os candidatos contestarem as novas reivindicações feitas durante os dias finais de uma eleição. **Na mesma linha, políticas que impeçam a circulação de anúncios sobre fraude eleitoral em vésperas de eleições, ou que restrinjam o seu alcance de forma transparente, informando tanto aquelas(es) que o publicaram, quanto que o visualizam sobre essa medida e como ela opera, parecem ser interessantes para evitar ataques ao processo eleitoral democrático e já vem sendo adotado por algumas empresas.**[28] A previsão de uma Biblioteca de Anúncios, na qual as publicidades são armazenadas durante um período de tempo, é fundamental nesse último caso, pois permite um maior nível de *accountability*, tanto pelo Estado, quanto por partidos e pela sociedade civil, permitindo também a possibilidade de denúncia de conteúdos políticos impulsionados contrários às regras previamente anunciadas pelas plataformas.

## PERFIS DE FIGURAS PÚBLICAS: DEVE HAVER DIFERENCIAÇÃO EM RELAÇÃO AOS DEMAIS PERFIS?

O contexto de integridade do debate público engloba também, figuras públicas, dentre elas ex-políticos, políticos(as) eleitos(as) e candidatas(os) à eleição, que podem estar envolvidas tanto na criação, quanto no compartilhamento desse tipo de narrativa. O alcance de publicações e mensagens desse tipo de perfil e a legitimidade adquirida de seu discurso pelo cargo que exercem é o que torna necessário um olhar atento sobre eles.

27 Todas as plataformas analisadas possuem políticas relacionadas a anúncios durante as eleições.

28 Em agosto de 2022, a Meta anunciou que passaria a proibir anúncios questionando a legitimidade das eleições de 2022. Disponível em <https://about.fb.com/br/news/2022/08/como-a-meta-esta-se-preparando-para-as-eleicoes-do-brasil-em-2022/>.

Algumas plataformas já tomaram decisões que geraram controvérsias a respeito de usuárias(os) com relevância política. Em 2021, por exemplo, a Meta **decidiu** suspender as contas do Facebook e do Instagram do ex-presidente dos Estados Unidos, Donald Trump, indefinidamente, em razão de publicações feitas pelo ex-presidente estadunidense demonstrando apoio aos participantes da invasão do Capitólio[29]. À época, **uma das grandes críticas feitas** em relação à decisão foi a ausência de clareza e transparência quanto às normas e sanções que seriam aplicáveis às e aos líderes políticos. Ao analisar o caso posteriormente, o Comitê de Supervisão da Meta **decidiu** que a empresa deveria "explicar publicamente as regras que usa quando impõe sanções às contas de usuários influentes", além de questionar a indefinição de um prazo para o banimento de Trump, requerendo que este fosse definido.

Apesar da expressão dessa preocupação por aqueles que lidam com assuntos relacionados à moderação de conteúdo quando a situação ocorreu, é possível dizer que, até o momento, não há consenso sobre como a moderação de conteúdo desses usuários deve ocorrer de acordo com regras gerais ou específicas. No geral, vê-se que as plataformas possuem políticas específicas para conteúdo considerados ou de "interesse público", como no Twitter, ou "interessantes"/"noticiáveis"(newsworthy) nas plataformas da Meta, sob os quais há a previsão de ainda que violem outras políticas da rede, não serão removidos em razão da sua relevância, porém não há distinções – pelo menos aparentes – sobre como as regras serão aplicadas a depender de quem realizou a publicação. Ou, ainda, mesmo que as políticas declarem expressamente que serão considerados discursos de todas as fontes[30], sistemas elaborados pelas plataformas para aplicação de tratamentos diferenciados a usuários(as) e entidades específicas, no momento de moderação de conteúdo, põem em xeque a equidade e a transparência na aplicação destas normas[31].

A primeira consequência de eventual diferenciação na forma que regras de conteúdo são aplicadas pelas plataformas em relação a figuras públicas é o fato de que candidatos e candidatas podem passar a não gozar das mesmas oportunidades para persuadir os eleitores. Se, por exemplo, candidatas(os) que buscam a reeleição em eleições presidenciais são beneficiadas(os) com exceções dos termos de uso das plataformas, os incumbentes (ou seja, os que já estão no poder) têm uma vantagem injusta. A segunda consequência, por sua vez, está relacionada à como figuras públicas podem usar sua proteção para propagar e legitimar discursos enganosos sobre o processo eleitoral, numa conexão com o necessário enfrentamento a ataques fraudulentos à integridade das eleições, discutido nos pontos acima. De modo geral, como já dito, os perfis de candidatas(os) ou figuras públicas possuem um alcance maior em relação ao perfil de demais usuárias(os) e os seus discursos também carregam mais credibilidade em razão de, muitas vezes, possuírem a

29 Durante o dia da invasão do Capitólio, Trump fez diversas publicações nas quais alegava infundadamente que tinha vencido as eleições presidenciais de 2020 e apoiava e exaltava as pessoas que estavam causando tumulto no prédio do legislativo estadunidense. Inicialmente, a suspensão de suas contas foi justificada para evitar que houvesse outros eventuais atos que impedissem uma transição de governo pacífica. Disponível em <https://www.oversightboard.com/decision/FB-691QAMHJ>.

30 A Meta, por exemplo, afirma em sua política de interesse público que não há a presunção de que o discurso de alguém, inclusive de políticos, é interessante. Disponível em <https://achearegra.internetlab.org.br/pesquisa/?plataformas%5B%5D=meta&visual=regras&categoria=interesse-publico>.

31 Em outubro de 2021, o Wall Street Journal divulgou informações sobre o programa de verificação cruzada da Meta, que dá diferentes tratamentos a usuários e entidades específicas no momento de moderação de conteúdo, com implementação de níveis adicionais de análise para essas contas específicas. Disponível em <https://oversightboard.com/news/501654971916288-oversight-board-publishes-policy-advisory-opinion-on-meta-s-cross-check-program/>.

legitimidade relacionada aos cargos que ocupam ou serem figuras de autoridade com quem as pessoas se identificam . Devido a essas características, e, considerando o papel desses indivíduos na sociedade, as suas publicações em plataformas podem possuir mais visibilidade e credibilidade aos demais usuários, de forma a influenciar de forma especial episódios de ataques à integridade eleitoral ou da ordem democrática.[32]

Por outro lado, diferenciações podem ser justificáveis não só pela perspectiva de acesso à informação, mas também a partir de necessárias camadas de segurança para que a moderação de conteúdo seja sensível aos marcadores sociais da diferença, não tolerando que a visibilidade propiciada pela condição de figura política ou pública aumente a probabilidade de que essas pessoas sejam alvo de violência e discurso de ódio. De acordo com o **MonitorA 2022: observatório político de violência de gênero**, é importante que haja por parte das plataformas a consciência de que grupos historicamente marginalizados como mulheres, negros, indigenas, LGBTQIAP+ e pessoas com deficiência, quando em posições de destaque podem ter seus direitos políticos minados por críticas e ataques massivos online e que, portanto, suas redes devam ter um monitoramento diferente em relação aos demais usuários(as). Dessa forma, é importante que a moderação de conteúdos hostis leve em conta as diferenças entre os grupos de pessoas em cargos públicos e seus contextos políticos, considerando a magnitude dos ataques e insultos, bem como as narrativas envolvidas.

Estas discussões sobre a diferenciação no tratamento a perfis de candidatas, candidatos ou figuras públicas em geral impõe uma reflexão necessária: **se, de um lado, pode-se argumentar que os discursos destas figuras são importantes para a promoção de um debate público saudável, não é possível, por outro, que as plataformas apenas observem os efeitos deletérios quanto a integridade eleitoral e cívica que esses perfis podem ter, ao conseguirem amplificar narrativas desinformativas e que ameacem instituições democráticas, sob a justificativa de interesse público.** Ainda mais quando o que é ou não interesse público não é transparente o suficiente para todas as partes interessadas na discussão.

Nesse sentido, ainda que possa haver um sistema de diferenciação entre o perfil de pessoas "comuns" e o perfil de figuras públicas (para proteger o acesso à informação de interesse público ou a participação política de grupos socialmente marginalizados), **conteúdos que violem políticas sobre integridade eleitoral e cívica devem ser tratados de maneira semelhante e moderados de forma ágil**, de modo a proteger o processo eleitoral como um todo e sua legitimidade. **O raciocínio neste ponto é que a informação de interesse público só é útil e a participação política só se viabiliza se o processo eleitoral restar preservado.**

As políticas sobre como figuras públicas devem ser tratadas pelas plataformas em momento eleitoral, assim, devem ser claras em relação ao momento em que um limite for ultrapassado,

32 Nesse cenário, setores passaram a defender a ideia da criação de uma espécie de "imunidade parlamentar" *online*, que seria aplicada às publicações de políticos em redes sociais defendida no projeto de Lei 2630/2016, conhecido como "PL das Fake News". Enquanto se entende que, no Congresso Nacional, os parlamentares têm toda a liberdade em seus pronunciamentos, pela presunção de que estão no exercício de sua função; fora dele, a imunidade é relativa, devendo ser analisada caso a caso diante das palavras proferidas. A ideia de uma disposição nesse sentido no projeto de lei supracitado, assim, é que os deputados e senadores não sejam responsabilizados, civil ou penalmente, por suas manifestações políticas em plataformas de rede social. Essa decisão tem efeitos diretos na moderação de conteúdo das plataformas, uma vez que, ainda que a previsão se direcione ao Estado, os provedores poderiam ficar receosos na remoção de atividades desses atores sociais, ainda que fossem violadoras de suas políticas, em razão da proteção concedida por lei.

para que candidatas(os) e outras pessoas estejam apropriadamente informadas(os) e de sobreaviso e possam saber quais ações de moderação de conteúdo podem ser tomadas em cada caso e a como as plataformas determinarão se um limite foi ultrapassado. Nessa linha, **a política de exceção de interesse público do Twitter** possui parâmetros interessantes para esse tipo de decisão, como o imediatismo e a gravidade do dano a partir da eventual violação de regras da plataforma; se a manutenção do conteúdo gerará alguma responsabilidade para a figura política envolvida; se existem outras fontes de informações sobre o assunto acessíveis para o público; se a remoção do conteúdo impede que as pessoas entendam uma questão de interesse público e se o conteúdo é necessário para uma discussão mais ampla. Destaca-se, ainda, que, para que a política tenha efetividade, é necessário que elas considerem como as autoridades eleitorais operam e em que momento esgotam-se todos os meios disponíveis para impugnar os resultados, além da realidade política local.

Assim, partindo de uma perspectiva mais ampla, faz sentido defender um necessário equilíbrio em casos que envolvam a moderação de conteúdos de candidaturas no momento eleitoral. De um lado, pesquisas como o *MonitorA* demonstram a necessidade de tratar diferentes como diferentes quando falamos do enfrentamento à violência política online. Perfis de candidatas e candidatos de grupos socialmente marginalizados só estarão em condições de igualdade de chances com seus demais competidores pelo voto popular se gozarem de proteções mínimas para exercer sua participação política em um mundo conectado. De outro lado, plataformas devem estar vigilantes para que seus sistemas e políticas sobre figuras públicas não desequilibrem a paridade de armas entre diferentes candidaturas, contribuindo para a equidade, como já ocorre em outros formatos de comunicação, por exemplo (emissoras de TV e de rádio, que não podem dar tratamento privilegiado a nenhum candidato(a), partido ou coligação).

Por fim, ressalta-se que, em eventuais decisões de remoção de conteúdos ou suspensão de perfis de candidatas(os) ou figuras públicas, há desafios de justificação a serem enfrentados.

Primeiramente, as plataformas devem ser transparentes em relação a quais regras da rede foram violadas e qual o período de tempo dessa suspensão de acordo com a gravidade do ocorrido. Em segundo lugar, uma vez que a suspensão ou restrição definitiva de um perfil significa o banimento permanente de uma pessoa ao espaço de debate político, o ato consiste em impedimento de sua expressão no futuro. Desta maneira, esta medida extrapola o que poderia ser considerado uma reparação ou cessação de danos já reconhecidos. A justificativa normativa é outra, com paralelo no mundo jurídico à medida "cautelar", uma ação protetiva que é tomada com base numa previsão ou iminência de danos no *futuro* que ainda não ocorreram. Assim, o bloqueio de contas deve ser pensado a partir de uma matriz diferente dos casos de mera remoção de uma postagem. Mais do que justificada por regras claras e possuir objetivo legítimo, como nesses casos mais corriqueiros, bloquear o acesso a um perfil requer que a plataforma assuma uma postura preventiva proporcional ao risco de dano, nos limites do necessário[33].

33 Disponível em <https://www.oversightboard.com/decision/FB-691QAMHJ>.

## INSURREIÇÃO E ROMPIMENTO COM O PROCESSO DEMOCRÁTICO

Assim como narrativas sobre fraudes eleitorais, conteúdos de incitação de insurreição ou rompimento com o processo democrático também merecem uma atenção especial por parte das políticas das plataformas. Como relatamos na parte contextual acima, as graves agressões às instituições democráticas brasileiras que ocorreram no dia 08 de janeiro de 2023[34] estão entrelaçadas a uma operação de influência distribuída com vários agentes engajados sem liderança central aparente, com algum nível de pactuação de ações entre estes componentes e estratégica, atingindo fins convergentes, de teor antidemocrático que estava sendo construída há meses em um nível multiplataforma. Em vésperas do primeiro turno das eleições brasileiras, materiais de intenção, incitações e declarações ligados à realização de manifestações antidemocráticas foram identificadas em aplicativos de mensagem.[35] Após o resultado eleitoral, discursos de conteúdos golpistas, notícias falsas sobre a possibilidade de ocorrência de intervenção militar e alegações de insegurança das urnas ganharam ainda mais força e adesão[36], tendo o seu ápice no dia do ataque aos prédios dos três poderes em Brasília.

Nos deparamos então com um grande entrelaçamento entre *online* e *offline*, o que é desafiador. Nessa seção, em vez de atribuir causa e efeito a este processo, argumentamos que faz sentido entender que as dinâmicas *online* compõem um quadro de retroalimentação com as dinâmicas *offline*, em um processo de afinidade e reforço. E que a força que essa categoria de manifestação possui em um possível rompimento com o processo democrático não pode ser ignorada, sendo fundamental:

**(i)** a existência de políticas voltadas a interromper esses ciclos de retroalimentação quando as dinâmicas *offline* tiverem o potencial de atingir a integridade física de pessoas e a integridade de processos democráticos;

**(ii)** a construção de uma gradação das sanções pela nocividade contextual do discurso, para que as plataformas que moderem esse tipo de conteúdo minimizem a sua interferência na circulação de conteúdos de interesse público, ainda que limítrofes em relação às políticas.

## A. POLÍTICAS PARA INTERROMPIMENTO DE CICLOS DE ROMPIMENTO DA ORDEM DEMOCRÁTICA

É necessário ter critérios, portanto, de como e quando cortar o circuito que pode implicar em insurreição e possíveis rompimentos da ordem democrática, ou seja, em golpes de Estado em regimes democráticos. Algumas plataformas já vêm tomando medidas nesse sentido, mas, além de nem todas possuírem políticas específicas para esse assunto, aquelas que possuem geralmente dispõem de regras genéricas ou que contemplam apenas o período eleitoral.

34 Em 8 de janeiro de 2023, manifestantes invadiram a sede dos Três Poderes da República em Brasília, causando danos físicos aos prédios.

35 A pesquisa Democracia digital: análise dos ecossistemas de desinformação no Telegram durante o processo eleitoral brasileiro de 2022 identificou dois grupos do Telegram, "Deputado federal José Medeiros" e "Mistérios do Mundo", respectivamente, que possuíam como maior atividade o compartilhamento de conteúdos golpistas explícitos e notícias falsas sobre a possibilidade de ocorrência de intervenção militar. Disponível em <https://internetlab.org.br/wp-content/uploads/2023/02/telegram-03-relatorio-02.pdf>.

36 A imagem com maior número de compartilhamentos neste período, com 7.931 compartilhamentos, apresentava uma convocação para manifestações na frente dos quartéis no dia 01 de novembro. Disponível em <https://internetlab.org.br/wp-content/uploads/2023/02/telegram-03-relatorio-02.pdf>.

**Em primeiro lugar, este mapeamento evidencia a importância de regras que proíbam conteúdos que contenham declarações, sejam elas explícitas ou implícitas, incitando ou defendendo a violência contra a ordem democrática ou contra a transmissão pacífica de poder antes, durante e após as eleições.** Como essas narrativas ultrapassam momentos específicos, as regras não devem se restringir somente aos períodos eleitorais, como já dito anteriormente em *Alegação de fraudes e questionamentos quanto à integridade eleitoral*.

Além disso, a discussão relacionada a conteúdos considerados golpistas e/ou de insurreição caminha lado a lado com questões de liberdade de expressão. É normal que, em uma eleição, haja descontentamento com o resultado e expressão de frustrações por pessoas de diferentes posições políticas e ideológicas. É também normal que isso ocorra de forma organizada. Ao pensarmos em políticas para impedir ameaças de abolição da ordem democrática ou à interferência na transmissão pacífica do poder, não podemos permitir que elas considerem essas manifestações automaticamente discursos contra a integridade eleitoral ou a democracia de forma mais ampla. A elaboração de políticas de baixa qualidade sobre o assunto pode levar a situações em que conteúdos são suprimidos indevidamente ou representar um risco de engessamento da moderação de conteúdos eleitorais pelas plataformas.

**Para evitar que haja a exclusão de discursos meramente descontentes quanto ao resultado eleitoral, prejudicando assim o direito à liberdade de expressão de usuárias(os), é importante reconhecer a existência de regras e definições explícitas sobre discursos antidemocráticos, com sanções específicas previstas.**

A ideia, portanto, é que as plataformas sejam capazes de avaliar, de maneira mais abrangente, os riscos que seus modelos de negócio criam para as eleições e para a democracia, bem como esses riscos podem ser mitigados nos estágios iniciais das operações da empresa e durante seu crescimento.[37] Por serem manejados, justamente, em estágios iniciais, a política de incitação a violência contra a ordem democrática ou contra a transmissão pacífica do poder seria aplicada apenas subsidiariamente, uma vez que as plataformas estariam na capacidade de impedir que os discursos não chegassem a esse nível de gravidade.

Para que essa situação seja possível, é necessário que as políticas de plataformas de redes sociais lidem, primeiramente, com a questão do que pode ser considerado como um conteúdo de ataque ao Estado Democrático de Direito. Ainda que o compartilhamento de materiais sobre golpe, intervenção militar e fraude nas urnas sejam algumas narrativas que podem servir de modelo na estruturação destas políticas, tem-se que, anteriormente a esta etapa, é necessário lidar com a linha tênue e subjetiva entre quais discursos são aceitáveis e quais não. Como reforçado anteriormente, por mais que críticas aos resultados oficiais não sejam um problema em si, manifestações que reverberam ou exponham a perigo direto ao Estado Democrático de Direito não podem ser aceitas ou ignoradas pelas plataformas. Além disso, as narrativas mencionadas foram as que surgiram nas últimas eleições, outras podem vir a surgir em outros momentos políticos.

37 Disponível em <https://gp-digital.org/wp-content/uploads/2022/04/OHCHR-UNGPs-and-Tech-Companies-Consultation_GPD-Submission.pdf>.

Em razão da dificuldade em estabelecer um conceito que considere todas as variáveis de manifestações contra a ordem democrática, é fundamental que as políticas estabeleçam balizas para se definir a representação real de perigo de cada discurso, em acordo com os padrões estabelecidos pela legislação, conforme se verá a seguir.

Ao implementar políticas sólidas que abordem as ameaças à ordem democrática, as plataformas podem interromper os ciclos de narrativas que comprometem a democracia e assegurar a preservação dos princípios democráticos em seu ambiente digital.

## B. MECANISMOS DE MINIMIZAÇÃO DA INTERFERÊNCIA NO DEBATE PÚBLICO SOBRE INTEGRIDADE ELEITORAL: GRADAÇÃO E NOCIVIDADE

A discussão relacionada a conteúdos considerados golpistas e/ou de insurreição caminha lado a lado com a questão da liberdade de expressão, trazendo pontos delicados sobre os limites deste direito e sobre censura. É normal, em uma eleição, que eleitoras e eleitores se sintam descontentes com o resultado e expressem suas frustrações de acordo com as suas posições políticas e ideológicas. Ao pensarmos em políticas para impedir ameaças de abolição da ordem democrática ou à interferência na transmissão pacífica do poder, não podemos deixar, portanto, que elas considerem essas manifestações, que fazem parte do jogo democrático, automaticamente, como discursos contra a integridade eleitoral ou a democracia de forma mais ampla.

Dessa maneira, a elaboração de políticas pouco claras sobre o assunto pode, além de levar a situações em que conteúdos são suprimidos indevidamente, representar um risco de engessamento da moderação de conteúdos eleitorais pelas plataformas. Tendo em vista essa preocupação, para evitar que haja a exclusão de discursos meramente descontentes quanto ao resultado eleitoral, prejudicando assim o direito à liberdade de expressão de usuárias(os), o que se defende é a existência de regras e definições explícitas sobre discursos antidemocráticos, com sanções específicas previstas e possibilidade de recurso em caso de erro da plataforma. **As sanções ou avisos educativos, por sua vez, devem ser gradativas perante critérios que determinam a gravidade da situação apresentada e o risco que representa ao contexto político local. Nesse sentido, é importante que haja o treinamento coerente com esses parâmetros para moderadoras(es) de conteúdo.**

Além disso, é preciso considerar, também, que, dentro do espectro de discursos considerados golpistas ou de insurreição, há aqueles que são mais graves, como os que incitam a utilização de violência em locais de zonas eleitorais ou de cerimônias de posse, e outros mais brandos, que buscam promover a realização de protestos contra os resultados das urnas, mas que ainda se moldam dentro da ideia de protestos pacíficos. Desse modo, a gradação das sanções a depender do grau de periculosidade material do discurso que se enfrenta parece ser uma boa alternativa para as plataformas que podem, em conteúdos mais graves, realizar a sua remoção e, em outros casos, indicar avisos/rótulos sobre o teor do discurso, de forma a deixar as e os usuárias(os) conscientes quanto ao tipo de conteúdo que estão consumindo, ou diminuir o alcance desse tipo de conteúdo de forma transparente. Essa gradação permitiria que assuntos de interesse público se mantivessem no ar, ainda que limítrofes em relação às políticas, sendo removidos apenas

em casos que representassem ameaças de violência ou à integridade de processos eleitorais ou democráticos de uma maneira mais ampla.

Para tornar esta análise de gravidade possível, recomendamos a elaboração de escalas de risco e proximidade com um possível cenário de insurreição e rompimento com a ordem democrática, com o auxílio de observadores locais e o treinamento coerente com esses parâmetros para sistemas e moderadoras(es) de conteúdo. Inclusive, entendemos que é esta reflexão a chave para produzir uma diferenciação entre "organização política legítima" (*legitimate political organization*) e "ação coordenada nociva" (*harmful coordinated action*). As ações poderiam ser consideradas nocivas caso contenham, conforme os exemplos mencionados, elementos que alimentam um ciclo de retroalimentação entre dinâmicas *online* (operações de influência) e *offline* (ação direta e movimentação de atores políticos relevantes) em contexto muito próximo à insurreição e/ou rompimento da ordem democrática.

Ao mesmo tempo, para a aplicação deste enquadramento, é necessário produzir definições operacionais de "insurreição" e "rompimento à ordem democrática" que, no limite, vão explicitar escolhas políticas e valores de cada empresa. Em tais definições é necessário redobrar o cuidado com a proteção à participação política e produzir salvaguardas que evitem a instrumentalização de determinados enquadramentos na moderação de conteúdo, inclusive. Não consideramos adequado, por exemplo, que se defina qualquer processo de mobilização civil de alta intensidade contra um regime estabelecido (o que poderia ser uma definição operacional de "insurreição") como algo tendente ao "rompimento à ordem democrática". Assim, é possível que haja uma mobilização do tipo que esteja posicionada em face de um regime autocrático e que não possui elementos básicos para ser enquadrado como um Estado Democrático de Direito, como a ausência de eleições livres e competitivas e algum arranjo de proteção à direitos fundamentais (que poderiam compor uma definição operacional de "ordem democrática").

# TRANSPARÊNCIA SOBRE A EFETIVIDADE DAS POLÍTICAS - A NECESSIDADE DE DESENVOLVIMENTO DE MÉTRICAS CLARAS

Uma vez destacados os pontos básicos que as plataformas devem levar em consideração para a elaboração de políticas que contribuam para a sustentação do estado democrático de direito, combatendo e mitigando efeitos de incitações à violência, desinformação quanto a processos cívicos e discursos prejudiciais à integridade eleitoral, é necessário jogar luz sobre um segundo ponto. Para que as políticas sejam efetivas, não basta que elas sejam apenas elaboradas e aplicadas, sendo fundamental a existência de métricas de avaliação e acompanhamento delas. Pois, ao final do dia, ainda que as regras das plataformas sejam as mais completas e claras possíveis, é impossível o controle sobre cada conteúdo específico emitido em uma rede social. Desse modo, para aferir o funcionamento das políticas, é imprescindível que haja um sistema capaz de acompanhar

(i) se as suas regras estão sendo suficientes para a construção de um espaço de debate compatível com valores democráticos mais saudável, e

(ii) em caso positivo, se as plataformas estão levando a cabo os seus compromissos.[38]

Esse sistema deve ser transparente para que esses objetivos se concretizem. A transparência torna o processo de moderação de conteúdo passível de *accountability* e, portanto, menos propenso a exceções ou equívocos, pois permite

38 O Comitê de Supervisão da Meta (Oversight Board) em decisão em junho de 2023 fez recomendação semelhante à Meta. Ao Comitê a empresa afirmou "não adotar nenhuma métrica específica para mensurar o sucesso de seus esforços de integridade eleitoral em geral". Disponível em: <https://oversightboard.com/decision/FB-659EAWI8/>

(i) a correção de assimetrias de informação entre aqueles(as) que estão por dentro da elaboração das políticas de uma determinada plataforma e aqueles(as) que não estão, de modo que estes(as) últimos(as) consigam compreender como as regras estão sendo aplicadas na prática;

(ii) que outros atores institucionais, como o governo e a sociedade civil, reajam ao conteúdo produzido nas redes e a sua correspondente moderação, devido a disponibilização de informações, pelos provedores, da extensão e dos tipos de publicações inflacionárias que têm circulado, sobre público alvo deste conteúdo e sobre a intervenção das plataformas para mitigá-lo;

(iii) uma atuação mais coordenada daqueles envolvidos em projetos de moderação de conteúdo e

(iv) uma maior legitimidade e razoabilidade para as decisões das plataformas.[39]

O ponto, porém, é que, atualmente, os dados fornecidos comumente pelas plataformas não parecem estar a serviço da concretização destas finalidades. Por mais que haja a divulgação periódica de relatórios de atividades, **outras iniciativas da sociedade civil** já destacaram que eles não são "completos, específicos e imediatos" e "os números, quando apresentados, não possuem denominador (ou indicativo de prevalência) ou discussão sobre a eficiência das políticas", o que os tornaria insuficientes. A falta de indicação de bases necessárias para que o(a) leitor(a) compreenda as estatísticas e números torna os relatórios unilaterais e vagos.

No geral, o que ocorre é a disponibilização de números brutos referentes às atividades de moderação de conteúdo. Tais números, entretanto, não são capazes de expressar as tendências em plataformas de rede social. Por exemplo, quando os provedores relatam um aumento na remoção de conteúdo, pode-se presumir, à primeira vista, que está ocorrendo um aperfeiçoamento da atividade de fiscalização sobre os conteúdos emitidos na plataforma. Todavia, sem a presença de métricas específicas que confirmem essa constatação, a exclusão de conteúdo pode ser justificada por diversos outros fatores: houve o aumento de conteúdo geral na plataforma; a plataforma impôs limites de tolerância mais baixos para conteúdos violadores; houve a ampliação da definição de violação de conteúdo pela plataforma; e assim por diante.

É necessário, assim, que eventuais relatórios de transparência divulgados por plataformas venham acompanhados de métricas claras e bem definidas, de modo que seja possível uma avaliação mais concreta sobre a moderação de conteúdo, que perpasse pelas causas, línguas, populações e regiões dos conteúdos presentes nas plataformas.

Ainda que o estabelecimento de quais métricas devem ser utilizadas seja um desafio contínuo para os provedores de redes sociais, já há a existência de algumas que podem ser empregadas com razoabilidade. Uma delas é a métrica de prevalência. A prevalência "estima a porcentagem

39 Douek, Evelyn, Content Moderation as Systems Thinking (Janeiro, 2022). Harvard Law Review Vol. 13. Disponível em: <https://ssrn.com/abstract=4005326>

de vezes que as pessoas veem o conteúdo violador em oposição ao seu volume bruto[40]". Dessa maneira, a métrica incentiva o provedor não só a remover postagens violadoras de suas políticas de conteúdo, mas também a reduzir a quantidade de vezes que as pessoas as visualizam.

Em razão da importância de aplicação desse tipo de métrica em relatórios de atividades, o Facebook, em 2019, compôs um **grupo de trabalho** destinado a avaliar a aplicação destas pela rede social. O trabalho deste grupo importa na medida que ele avança em imaginar quais seriam dados publicáveis em relatórios de transparência que poderiam produzir um ganho na prestação de contas das plataformas sobre seus esforços de moderação de conteúdo - mas deve ser lido com uma lente de cautela por cada plataforma funcionar de maneira diferente. Dentre as **recomendações elaboradas**, destacam-se:

**(i)** a criação, para além de uma métrica de prevalência de consumo, na qual se indica a quantidade de conteúdos violadores que são consumidos, de uma métrica de produção, que seria capaz de indicar a quantidade de conteúdo violador do total de conteúdos existentes na plataforma;

**(ii)** que o Facebook empregasse esforços para relacionar as suas métricas de prevalência aos danos causados no mundo real;

**(iii)** que o Facebook explorasse formas de contabilizar o nível de gravidade da violação nas métricas de prevalência;

**(iv)** explicações mais detalhadas das métricas já aplicadas, com a indicação da prevalência de certos tipos de violação em determinadas áreas do mundo ou quantos conteúdos são removidos em comparação aos conteúdos que são apenas sinalizados e

**(v)** métricas adicionais capazes de indicarem os esforços da plataforma para aplicar suas políticas, como a frequência com que usuários(as) discordam das decisões de moderação de conteúdo.

No contexto de garantia da democracia e das eleições, as plataformas devem, assim, tornar públicas as informações relacionadas à moderação de conteúdo por violação das políticas de integridade eleitoral por meio dessas métricas de transparência, com o número total de publicações e contas marcadas (*flagged*) e removidas de forma especificada quanto ao formato - vídeo, áudio, imagem, texto ou *livestream*), à fonte – determinação governamental, algoritmos, outra determinação legal, sinalização por outros usuários ou sinalização de *trusted flaggers*, e quanto aos locais das remoções.[41]

40 Ibidem

41 Esses indicadores inspiram-se no princípio 1 dos *Operational Principles (Numbers)* dos Princípios de Santa Clara: "Companies should publish information about pieces of content and accounts actioned, broken down by country or region, if available, and category of rule violated, along each of these dimensions: Total number of pieces of content actioned and accounts suspended; Number of appeals of decisions to action content or suspend accounts; Number (or percentage) of successful appeals that resulted in pieces of content or accounts being reinstated, and the number (or percentage) of unsuccessful appeals and; Number (or percentage) of successful or unsuccessful appeals of content initially flagged by automated detection.; Number of posts or accounts reinstated by the company proactively, without any appeal, after recognition that they had been erroneously actioned or suspended. Numbers reflecting enforcement of hate speech policies, by targeted group or characteristic, where apparent, though companies should not collect data on targeted groups for this purpose (...) Special reporting requirements apply to decisions made with the involvement of state actors, which should be broken down by country: The number of demands or requests made by state actors for content or accounts to be actioned; The identity of the state actor for each request; Whether the content was flagged by a court order/judge or other type of state actor; The number of demands or requests made

Destaca-se, também, que a leitura e compreensão destes relatórios só é possível se as próprias políticas das plataformas se apresentam de maneira acessível e organizada. Atualmente, porém, as políticas referentes a um determinado assunto, como eleições, tendem a estar dispersas entre postagens nos sites da plataforma, declarações de diretores(as) da empresa e de porta-vozes para a mídia nacional e internacional, discursos públicos feitos por funcionários(as) e assim por diante. Sendo essa uma das motivações para que o InternetLab criasse o **achearegra**, observatório de termos de uso das plataformas digitais já mencionado, entendemos que é crucial que a plataforma faça o possível para tornar mais fácil o acompanhamento, por usuárias(os), das atualizações, alterações e limites das políticas.[42] Dessa maneira, é necessário, também, que as plataformas assegurem que as políticas  abordem categorização por tema, indiquem exemplos que as deem maior concretude, prevejam gradação de penas de acordo com a gravidade da conduta cometida, listem quais assuntos possuem maior urgência para serem moderados pela plataforma, indiquem critérios para a utilização de moderação humana e assinalem se um determinado tema possui relação com outras categorias de políticas ou não.

É importante ainda que esses dados sejam acessíveis a pesquisadores(as) e acadêmicos(as) de forma aberta, considerando a construção de diagnósticos, avaliações, comparações e recomendações para o enfrentamento a discursos antidemocráticos. Para além de dados quantitativos, é fundamental para a garantia da liberdade acadêmica e de produção de conhecimento que as práticas de transparência e acesso a dados estenda-se a outros aspectos da plataforma, como contato, reuniões e entrevistas com as equipes de moderação de conteúdo, de direitos humanos e de políticas públicas das plataformas e acesso a políticas internas.

**Desta forma, não é só fundamental a pactuação de um compromisso com a democracia na elaboração das regras a serem aplicadas, mas também a extensão deste compromisso às medidas de transparências que sejam suficientes para verificar se tais regras estão sendo implementadas. Políticas, decisões e relatórios de transparência de plataformas digitais devem ser instrumentos úteis para que a sociedade acompanhe o gerenciamento em escala que tais atores privados fazem da expressão - em especial em momentos eleitorais.**

by state actors that were actioned and the number of demands or requests that did not result in actioning. Whether the basis of each flag was an alleged breach of the company's rules and policies (and, if so, which rules or policies) or of local law (and, if so, which provisions of local law), or both; Whether the actions taken against content were on the basis of a violation of the company's rules and policies or a violation of local law." Disponível em: <https://santaclaraprinciples.org/>

42 O que também foi recomendado no Community Standards Enforcement Report produzido pela Meta em 2019, disponível em: <https://about.fb.com/news/2019/05/dtag-report/>

## A RECENTE DECISÃO DO COMITÊ DE SUPERVISÃO DA META E RECOMENDAÇÕES DE TRANSPARÊNCIA

Durante a escrita deste *policy paper*, como já mencionado, o Comitê de Supervisão da Meta (Oversight Board) anunciou a seleção, para julgamento, de um caso brasileiro envolvendo a manutenção de uma publicação que incitava invasões às sedes dos Três Poderes, em Brasília. De acordo com o Comitê, o objetivo seria analisar como a Meta modera conteúdo relacionado a eleições e como estaria aplicando o **Protocolo de Política de Crise em um "local designado temporariamente de alto risco"**, desenvolvido em resposta à recomendação da organização no caso **"Suspensão do ex-presidente Trump"**. O InternetLab apresentou **contribuição** ao caso.

Em sua deliberação, o Comitê revogou a resolução original da Meta de manter no Facebook a publicação. De acordo com a decisão do Comitê, a Meta esclareceu em sua contestação que não possuía dados abrangentes sobre denúncias relacionadas às eleições brasileiras em suas plataformas, tanto antes quanto durante e após o período eleitoral. Além disso, informaram que as pessoas responsáveis por moderar o conteúdo, apesar de possuírem fluência em português e conhecimento cultural e linguístico para analisar conteúdos brasileiros, estavam localizadas na Europa.

Nesse contexto, o Comitê ressaltou a importância da Meta desenvolver uma estrutura de avaliação dos esforços de integridade eleitoral que abrange tanto o processo eleitoral propriamente dito, como o período pós-eleitoral, que também é vulnerável à manipulação, à desinformação, e às ameaças de violência. Indicou-se, assim, que a definição de métricas precisas e a divulgação transparente dessas informações são passos primordiais para prevenir a incitação à violência e aprimorar a moderação de conteúdo.

Ademais, além do Protocolo de Política de Crise, o Comitê enfatizou a relevância de a Meta elucidar, em sua Central de Transparência, a adoção de outros protocolos destinados a enfrentar riscos inerentes a eventos eleitorais e situações de elevado potencial de risco. Isso inclui nomear e descrever os protocolos, os seus objetivos, as suas semelhanças e diferenças. Tais ações são essenciais para promover a prevenção contra ações organizadas que visam prejudicar os processos democráticos.

# RECOMENDAÇÕES

Pensar em recomendações para que plataformas lidem com questões de manutenção do debate eleitoral íntegro nas redes sociais ou, mais que isso, para que as redes possam se colocar como um espaço para debates cívicos acontecerem com qualidade envolve compreender a complexidade que a esfera pública tomou com a internet, buscando a garantia dos direitos fundamentais de eleitoras, eleitores, pessoas que ocupam cargos na política institucional ou que almejam ocupá-los.  Esse processo exige, assim, se atentar sempre aos perigos à liberdade de expressão, de associação e manifestação que políticas e normas que regulam o debate político podem carregar. Além disso, apesar de compreender que a responsabilidade por um debate eleitoral guiado por princípios democráticos não é apenas das plataformas, focamos nossas recomendações em suas políticas por entendermos que elas representam um primeiro filtro a ameaças modernas à democracia.

Portanto, os compromissos a serem assumidos pelas plataformas que operam em contextos eleitorais e são arena da comunicação política digital são:

## 1 ELABORAR DIRETRIZES DE CONTEÚDO ESPECÍFICAS SOBRE INTEGRIDADE CÍVICA E ELEITORAL

A moderação de conteúdo é parte fundamental da atividade das plataformas e possui impacto direto em direitos humanos, como a liberdade de expressão, o livre desenvolvimento da personalidade e a soberania popular. Por sua vez, a discussão sobre a necessidade de um compromisso de empresas com direitos humanos tem amparo em instrumentos de direito internacional, que entendem que a própria elaboração dos modelos de negócio deve levar em consideração os riscos criados a direitos fundamentais.

Com a consolidação das plataformas como atores importantes na dinâmica comunicacional sobre política e o aumento de desinformações e de atitudes que incitam a violência em discussões sobre eleições e sobre a própria democracia vinculados nas redes sociais, o estabelecimento de políticas de integridade cívica e eleitoral, de acordo com os limites de suas arquiteturas, é essencial.

Recomendamos, ainda, a criação de um grupo de trabalho da indústria, no qual as plataformas  possam coordenar parâmetros mínimos de proteção da integridade eleitoral com a finalidade de entender as similaridades e diferenças nas suas diretrizes e pensar em maneiras de operacionalizar a contenção de operações de influência coordenada que ocorram inter-plataformas.

## 2 COMPROMETER-SE COM ABORDAR PONTOS-CHAVE EM SUAS DIRETRIZES SOBRE INTEGRIDADE CÍVICA E ELEITORAL

Para além da elaboração de políticas específicas sobre integridade cívica e eleitoral, é necessário que elas sejam capazes de lidar com as novas dinâmicas dos períodos eleitorais e sejam adaptadas ao contexto brasileiro, abordando, pelo menos, os seguintes pontos:

### ALEGAÇÃO DE FRAUDES E QUESTIONAMENTO QUANTO À INTEGRIDADE ELEITORAL.

Tendo em vista o fortalecimento de narrativas de alegações de fraudes eleitorais sem nenhum tipo de comprovação e questionamentos infundados quanto à integridade eleitoral, é fundamental que as plataformas priorizem a criação de termos de uso, políticas e protocolos específicos para discursos diretos e indiretos sobre o tópico. Como essas narrativas não são inerentes a um período eleitoral específico, o ideal é que a política se desdobre em dois ramos:

(i) **o primeiro que lidaria com a proteção mais ampla dos processos democráticos, levando em consideração que o funcionamento da democracia não se restringe ao período eleitoral e**

(ii) o segundo, subsidiário, que seria circunscrito a um recorte temporal do período eleitoral, abarcando conteúdos potencialmente violadores próprios desta temporalidade e que possuiria "intervalos de amortecimento" para lidar com marcos fáticos e institucionais relevantes que se colocam poucos meses antes ou depois do período eleitoral oficial.

**Para aplicativos de mensageria, a inclusão de ferramentas de pesquisa que permitem verificar a veracidade de mensagens encaminhadas, a limitação de encaminhamentos de mensagens e, também, o controle de disparos em massa de mensagens pode ser uma alternativa para a mitigação de riscos desses discursos.**

### ANÚNCIOS ELEITORAIS.

Considerando que anúncios eleitorais em plataformas digitais também fazem parte, agora, das dinâmicas das eleições, é necessário que as políticas das plataformas regulem essa prática de marketing. Além de procedimentos gerais para a circulação do anúncio, as plataformas devem ter medidas que permitam evitar anúncios que contenham desinformação eleitoral ou conteúdos que questionem a integridade eleitoral e a democracia. A previsão de restrições quanto à circulação de anúncios eleitorais em períodos próximos aos dias de votação e quanto ao alcance de anúncios que contenham desinformações, assim

como a existência de uma biblioteca de anúncios são algumas das alternativas para uma melhor fiscalização desses conteúdos.

### PERFIS DE CANDIDATURAS E FIGURAS PÚBLICAS: EQUIDADE E PROTEÇÃO À VIOLÊNCIA POLÍTICA.

O grande alcance dos perfis de candidatas(os) e figuras públicas em geral e o seu potencial de legitimação de discursos não verídicos sobre eleições demonstram a necessidade de uma moderação rápida de conteúdos que violem políticas sobre integridade eleitoral e cívica nesse tipo de perfil. Ao mesmo tempo, é preciso cautela para que os protocolos não façam diferenciação entre candidatas(os) durante as corridas eleitorais, de modo que todas(os) tenham as mesmas chances de persuadir eleitores. O que se pretende dizer é que a equidade entre as(os) candidatas(os) deve ser uma prioridade, sendo a igualdade de chances um valor a ser respeitado também pelas plataformas, como já ocorre em outros formatos de comunicação, por exemplo emissoras de TV e de rádio, que não podem dar tratamento privilegiado a nenhum candidato(a), partido ou coligação. É por isso que as políticas relacionadas a esse tópico devem ser claras em relação ao momento em que um limite for ultrapassado, quais as punições permitidas e seu intervalo temporal.

Ainda nesse tópico, se faz necessária uma ressalva quanto a violência política sofrida por figuras públicas pertencentes a grupos historicamente marginalizados, como mulheres, negros e LGBQIAP+. Para igualdade de chances entre candidatas(os) é importante que as plataformas estejam atentas quanto a moderação de comentários e respostas que as postagens destes grupos recebem. Atuando de forma contundente contra todo tipo de violência política.

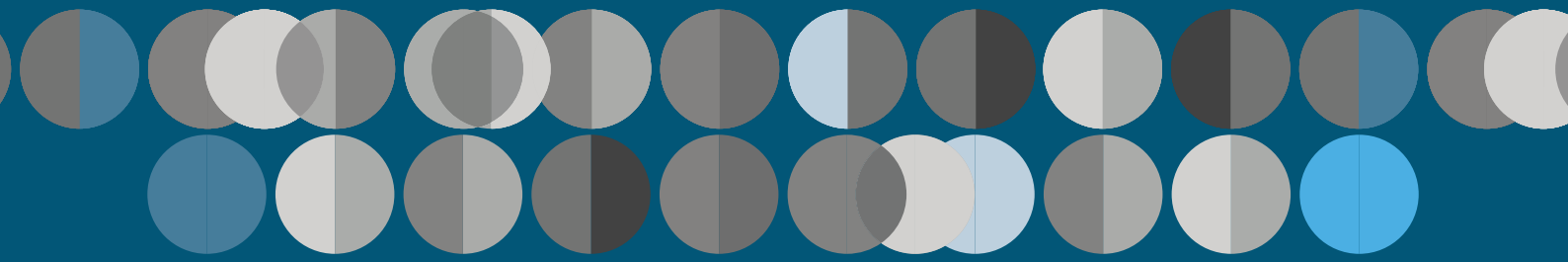
### ROMPIMENTO COM O PROCESSO DEMOCRÁTICO.

Compartilhando algumas semelhanças com as narrativas sobre fraudes eleitorais, conteúdos de incitação de insurreição ou rompimento com o processo democrático também merecem uma atenção especial por parte das políticas das plataformas. Assim como no caso de alegação de fraudes e questionamento quanto à integridade eleitoral, as narrativas contra o processo democrático também ultrapassam momentos específicos. Assim, a previsão de políticas que proíbam conteúdos que contenham declarações de incitação ou defesa de violência contra a ordem democrática ou contra a transmissão pacífica de poder, levando em consideração o seu contexto e risco de dano, não deve se restringir somente aos períodos eleitorais. Ainda, é necessário que essas políticas sejam sensíveis a contextos de pré-violência, isto é, que tornem as plataformas capazes de avaliar, previamente e de maneira abrangente, os riscos que seus modelos de negócio criam para as eleições e para a democracia. Para que essa

situação seja possível, as plataformas devem estabelecer balizas para se definir a representação real de perigo de cada discurso, com sanções gradativas perante critérios que determinem a gravidade da situação apresentada e o risco que representa ao contexto político local.

## 3 APRIMORAMENTO DE MEDIDAS DE TRANSPARÊNCIA NA MODERAÇÃO DE CONTEÚDO QUE SEJAM SUFICIENTES PARA ACOMPANHAR OS COMPROMISSOS ASSUMIDOS.

Ainda que as políticas  sejam importantes para garantir a integridade do debate público e o respeito ao pluralismo político, não basta que elas sejam apenas elaboradas. Para que seja possível avaliar o seu funcionamento, é imprescindível a criação, também, de um sistema capaz de monitorar a implementação e a efetividade desse compromisso assumido pelas plataformas. Esse sistema, por sua vez, demanda a presença de relatórios de transparência que possuam métricas claras, que ajudem a dimensionar as ameaças à democracia, através da indicação da prevalência do conteúdo violador, do número de denúncias recebidas, da quantidade e do local das publicações removidas. Ademais, para que os próprios relatórios façam sentido, também é necessária a apresentação das políticas e das ações de moderação empreendidas de uma forma acessível e organizada.  Tais práticas possibilitam não só uma avaliação das políticas, mas, também, permitem  legitimar as atitudes pelas plataformas,  de modo a mitigar acusações de enviesamento advindas de todo o espectro político.



INTERNETLAB