Autorregulação de conteúdos eleitorais: aspectos legais, riscos e desafios de implementação

Rodrigo Gurgel Fernandes, Raquel Cavalcanti Ramos

Resumo

Narrativas falsas sempre foram uma arma na estratégia de persuasão do eleitorado. Entretanto, a popularização da Internet elevou exponencialmente a difusão de conteúdos falsos. Diante dessa realidade, tem-se buscado abordagens jurídicas para o controle da difusão de notícias fraudulentas. O emprego da autorregulação regulada de conteúdo eleitoral foi proposto no âmbito do Projeto de Lei no 2630/2020. Este estudo investiga modelos potenciais de regulamentação e iniciativas internacionais já em vigor, além de aspectos práticos sobre a viabilidade do controle, como a acurácia e a aplicabilidade dos métodos automáticos de detecção de *fake-news* e *deep-fakes*, ressaltando os riscos e possíveis ataques sistêmicos. Como resultado, são formuladas recomendações tanto para a elaboração de normas, quanto para a implementação de sistemas de moderação. Além disso, são destacadas assimetrias entre a regulação de conteúdo com teor eleitoral e de conteúdo abusivo, como racismo, evidenciando a necessidade de normas específicas para regulação de conteúdo eleitoral.

Palavras-chaves: *fake-news*; autorregulação; *deep-fakes*; eleitoral; PL 2630/2020;

Sobre os autores

1- Possui graduação em Engenharia de Comunicações pelo Instituto Militar de Engenharia (1996) e mestrado em Engenharia Elétrica pela Universidade Federal de Pernambuco (2001). Está cursando o Doutorado em Engenharia Elétrica na Universidade de Brasília.

2- Possui graduação em Direito pela Universidade Federal do Ceará (2001), mestrado em Direito (Direito e Desenvolvimento) pela Universidade Federal do Ceará (2006) e doutorado em Direito pela Universidade de São Paulo (2013). É professora da Universidade Federal do Ceará e advogada

Abstract
False narratives have always been a weapon in the electorate persuasion strategy. However, the popularization of the Internet has exponentially increased the dissemination of false content. Faced with this reality, legal approaches have been sought for controlling the spread of fake news. The use of regulated self-regulation of electoral content was proposed within the scope of PL 2630/2020. This study investigates potential models of regulation and international initiatives already in place, as well as practical aspects regarding the feasibility of control, such as the accuracy and applicability of automatic methods for detecting fake news and deepfakes, highlighting the risks and potential systemic attacks. As a result, recommendations are formulated both for the elaboration of norms and for the implementation of moderation systems. Additionally, asymmetries between the regulation of content with electoral content and abusive content, such as racism, are highlighted, emphasizing the need for specific norms for the regulation of electoral content.
Keywords: fake-news; self-regulation; deep-fakes; electoral; PL 2630/2020;

1. Introdução

A disseminação de conteúdo de natureza eleitoral total ou parcialmente falso não é novidade no cenário político. Segundo (CASTELLS, 2018), o assassinato do caráter é o modo mais eficiente de conquista do poder na nossa sociedade, pela criação de escândalos políticos que, embora sejam tão antigos quanto a própria história, vêm nos últimos tempos sendo ampliados pela evolução das tecnologias. Até recentemente, a divulgação de um conteúdo acessível por um universo aberto de indivíduos era feita pela grande mídia, como jornal, rádio e televisão. A confiabilidade da informação estava ligada a padrões éticos profissionais de jornalistas e à responsabilização civil ou penal. Entretanto, a popularização da Internet, passando por *blogs* até as inúmeras redes sociais atuais, pulverizou o acesso a canais de publicação. Como resultado, acompanhamos um aumento exponencial de conteúdos eleitorais falsos, representando um desafio significativo aos mecanismos judiciais de controle do processo eleitoral.

Diante da impossibilidade operacional de tratar um volume crescente de notícias falsas de natureza eleitoral, o emprego da autorregulação de conteúdo surgiu como uma resposta alternativa. Naturalmente, o interesse pelo assunto, por parte de governos e da academia, cresceu na mesma velocidade. Inúmeras pesquisas,

projetos e iniciativas procuram processar dados e ideias para tentar solucionar este problema, que é agravado pelo dilema, aparentemente inafastável, entre a manutenção da liberdade de expressão e a aplicação de medidas de moderação para prevenir e reprimir a divulgação de *fake-news*.

O presente trabalho aborda a autorregulação, analisando suas formas e preenchendo uma lacuna observada entre as iniciativas legislativas e a discussão de aspectos práticos dos sistemas de autorregulação. O relativo desconhecimento, no âmbito jurídico e legislativo, de assuntos como aprendizado de máquina e sistemas de classificação automática, induz à falácia da confiabilidade plena nos algoritmos de reconhecimento automático de notícias falsas. Essa lacuna, por si só, representa um risco, que pode levar à aprovação de soluções legislativas inócuas ou mesmo prejudiciais ao processo eleitoral. Esta pesquisa visa contribuir com informações úteis, tanto para a elaboração de normas quanto para a implementação de sistemas de moderação, analisando aspectos práticos como modelos de sistemas de revisão, desempenho de algoritmos, riscos sistêmicos e avaliação de iniciativas internacionais. A pesquisa tenta também, especificamente, abordar aspectos práticos como a acurácia e a aplicabilidade dos métodos automáticos de detecção de *fake-news* e *deep-fakes* no contexto da autorregulação de conteúdo de natureza eleitoral, e possíveis ataques sistêmicos.

No segundo capítulo são estudadas as formas de regulação. Algumas iniciativas legislativas internacionais de autorregulação regulada são descritas. É feita uma análise dos riscos e desafios, cobrindo aspectos como desempenho de métodos automáticos de detecção de *fake-news* textuais e *deep-fakes*. No terceiro capítulo, é realizada uma análise da regulação de conteúdo eleitoral prevista no PL 2.630/2020, sobretudo quanto à responsabilização dos provedores e à aplicação de sanções administrativas. Também é feito um ensaio sobre possíveis ataques sistêmicos na moderação de conteúdo notificado pelo usuário. Embora a aprovação da proposta de lei seja incerta, a sua análise pode ser eventualmente útil no debate de uma futura regulamentação da matéria, já que ela traz em si a semente de muitos debates que permanecem abertos. Neste sentido, no último capítulo, é feita uma síntese dos resultados onde são sugeridas recomendações.

2. Autorregulação de conteúdo: necessidade, formas e desafios

O conteúdo divulgado na internet admite as seguintes formas de disciplinamento: regulação por terceiro, autorregulação e a autorregulação regulada. Cabe destacar que as diferentes formas podem coexistir, sobretudo a regulação por terceiro e a autorregulação.

A *regulação por terceiro*, realizada pelo Estado, tem a vantagem de ser definida com base no interesse público, permitindo o devido processo de revisão. A desvantagem deste modelo é que o Estado não possui tantos recursos tecnológicos quanto as plataformas para sua implementação. No Brasil, a regulação estatal de conteúdo associados a tipos penais é feita na esfera criminal. A atribuição para persecução penal do crime de injúria racial (CP, art. 140, § 3º), e.g., é da justiça estadual, mesmo quando praticado pela rede mundial de computadores com potencial de transnacionalidade, salvo quando for conexo com crime federal, conforme jurisprudência do STJ (AgRg no CC 118.394/DF, DJe 22/08/2016). Já a regulação estatal de conteúdo eleitoral é feita pela Justiça Eleitoral, por meio de diversas ferramentas legais, como o direito de resposta, a Ação de Investigação Judicial Eleitoral, a Representação Eleitoral e o poder de polícia.

A *autorregulação* é realizada pelas próprias empresas, através de normas internas que regulam o uso das redes, os "Termos de Uso". Este modelo permite o uso de algoritmos de detecção e a estrutura tecnológica das plataformas, de forma mais dinâmica. Entretanto, a autorregulação é comumente alvo de críticas por priorizar interesses privados ou negociais, sem compromisso com valores sociais, além da falta de transparência. A autorregulação pode resultar na aplicação de medidas de moderação, tais como: 1) rotulação do conteúdo, e.g. pela indicação de conteúdo suspeito, inverídico ou sob análise; 2) restrição de visualização, limitando o acesso por idade; 3) suspensão provisória do conteúdo; 4) remoção definitiva do conteúdo; 5) desmonetização do usuário; 6) suspensão provisória da conta de usuário; 7) remoção definitiva da conta de usuário.

Em (MARANHÃO *et al*, 2021) são descritas diretrizes para a regulação das plataformas, que buscam estimular a responsabilidade e a transparência. Desde 2018, os Princípios de Santa Clara, conjunto de normas pensado por acadêmicos e integrantes da sociedade civil preocupados com o tema, definem níveis mínimos de

transparência e responsabilidade da moderação por autorregulação, como: 1) Publicação regular dos números de postagens removidas e contas permanente ou temporariamente suspensas devido a violações de suas diretrizes de conteúdo; 2) Notificação do usuário da remoção de conteúdo ou suspensão de conta, com motivação e registro do processo de moderação; 3) Possibilidade de recurso contra qualquer remoção de conteúdo ou suspensão de conta, com revisão externa independente. Tais princípios, apesar de não levarem à coação jurídica, estabelecem parâmetros críticos para refletir sobre o controle de conteúdo.

Na *autorregulação regulada*, a moderação é realizada pela plataforma, cabendo ao Estado a definição de diretrizes e a fiscalização do cumprimento das obrigações. Esta forma permite conciliar vantagens das duas abordagens anteriores, pois preserva o interesse público, aproveitando a dinamicidade e o *know-how* das plataformas.

A *Lei Alemã de Aplicação da Rede (NetzDG)*, conhecida como "Lei do Discurso de Ódio", em vigor desde 2018, visa responsabilizar as plataformas de redes sociais e combater o discurso considerado ilegal pela lei penal interna. A NetzDG impôs a elaboração de relatórios de transparência, a criação de um sistema de gerenciamento de reclamações e a designação de um representante legal no país. Embora tenha elevado a transparência das plataformas, a NetzDG levantou polêmicas. Como descrito em (TWOREK e LEERSSEN, 2019), ativistas de direitos civis, acadêmicos, e associações, como a Associação Alemã de Jornalistas, assinaram uma declaração conjunta alertando sobre os riscos à liberdade de expressão. Esta preocupação estava associada à possibilidade de uma filtragem exagerada de conteúdo, mesmo legal, para mitigar o risco de multa. Elevados custos, associados a prazos apertados e pesadas multas poderiam produzir um viés favorável ao acolhimento das reclamações. Também se alertou que as plataformas não teriam experiência e tempo para analisar adequadamente as notificações, o que demandaria conhecimento significativo da jurisprudência alemã, além de investigações complexas de fatos. Definições vagas de conteúdos proibidos, e.g. "discurso de ódio", difamação e injúria, foram alvo de críticas. Conforme descrito em (TWOREK; LEERSSEN, 2019), outro alerta apontava o risco de casos proeminentes de exclusão induzirem a um sentimento antigovernamental

e a um efeito prático inverso, da divulgação ampla do material excluído, fenômeno conhecido como Efeito Streisand. A NetzDG, que não criou novos tipos penais, obrigou as plataformas a cumprir 22 dispositivos penais existentes, que incluem: "incitação ao ódio", "disseminação de representações de violência", "formação de organizações terroristas", "uso de símbolos de organizações inconstitucionais", "distribuição de pornografia infantil", "injúria", "difamação pessoal", "ataque a religiões", "violação da privacidade íntima por fotografias", "ameaça à prática de crime" e "falsificação de dados destinados a fornecer provas".

Conforme o *Regulamento de Serviços Digitais, Digital Service Act (DSA)*, do Parlamento Europeu, Plataformas Online de Grandes Dimensões, *Very Large Online Platforms* (VLOPs), são obrigadas a avaliar e mitigar riscos sistêmicos pelo menos uma vez por ano. O art 26.º prevê que a avaliação deve considerar os riscos sistêmicos que incluem a manipulação do seu serviço, a ação de contas inautênticas ou a exploração automatizada do serviço, com um efeito negativo real ou previsível relacionados, entre outros, a processos eleitorais. Uma vez identificados os riscos sistêmicos, os VLOPs devem implementar medidas "razoáveis, proporcionais e medidas de mitigação eficazes" que incluem a adaptação dos sistemas de moderação ou recomendação de conteúdos, os seus processos de tomada de decisão; medidas específicas de limitação da exibição de anúncios; reforço dos processos internos ou a supervisão de qualquer uma das suas atividades relativas à detecção de risco sistêmico; cooperação com sinalizadores de confiança (*Trust Flaggers*) ou com outras plataformas online através dos códigos de conduta e dos protocolos de crise. Portanto, a obrigação de remoção não está prevista no tratamento de risco sistêmico, e só deve ser aplicada a conteúdos ilegais, como conteúdos terroristas, material de abuso infantil e discurso de ódio ilegal, entre outros.

Como visto, a autorregulação regulada possui algumas vantagens, contudo há uma grande polêmica acerca dos riscos. Por um lado, alega-se que somente as plataformas de redes sociais teriam condições de implantar ferramentas de detecção de *fake news e deep fake*, rotulando e moderando o conteúdo. Por outro, há uma preocupação com o efeito da regulamentação, pela imposição de uma carga excessiva às plataformas, levando a restrições inconstitucionais da liberdade de expressão.

A responsabilização das plataformas pelos conteúdos postados por usuários é polêmica. Alguns apontam que esta responsabilização objetiva pode acarretar o aumento indevido das medidas de moderação, alegando a ausência de nexo causal entre o dano gerado e o ato de disponibilizar o acesso do usuário à plataforma. Pelo receio de qualquer punição, indivíduos podem se autoconter de publicar aquilo que legalmente podiam, e plataformas podem censurar o que não deveriam. Este efeito de silenciamento é considerado uma censura colateral. Por outro lado, parte da doutrina defende que a responsabilidade dos provedores deveria ser objetiva, com base na teoria do risco-proveito. Este debate já ocorreu no âmbito do Marco Civil da Internet (Lei nº 12.965/14), para definir a responsabilidade subjetiva, condicionada ao descumprimento de ordem judicial específica por parte dos provedores. Conforme Diogo Rais (RAIS, 2018, p.100), as regras definidas na Lei nº 12.965 devem ser aplicadas também para fins eleitorais, não havendo razão para se criar uma disciplina paralela.

Em (ALVIM *et al*, 2023, p. 50) é apresentada uma divisão teórica dicotômica da visão sobre regulação. O *Modelo de inclinação liberal* (Livre mercado de ideias) caracteriza-se pela ausência de entraves à comunicação pública, sobretudo quanto aos conteúdos, e pela rejeição da tutela do pelo Estado sobre a verdade. O *Modelo de inclinação garantista* legitima a adoção de medidas regulatórias em nível estrutural, a fim de assegurar que o debate público seja plural, assegurando o equilíbrio de influências e promovendo o direito à informação em uma dimensão substancial. Todavia, vale destacar que existe uma gradação entre esses modelos de visão, na qual se defende que a autorregulação deve ser considerada, mas a sua aplicação deve ser bastante criteriosa, mitigando os riscos de censura indevida.

A liberdade de expressão é imprescindível na democracia. Além da participação em eleições, a democracia requer, por definição, que os cidadãos sejam capazes de influenciar as decisões estatais. Segundo Aline Osório (OSÓRIO, 2017, p. 129):

> Os cidadãos precisam de plena liberdade não só para acessarem tais informações, mas para manifestarem livremente as suas próprias ideias, críticas e pontos de vista na arena pública. (...) o livre fluxo de ideias e informações é essencial ao autogoverno democrático. Igualmente, as múltiplas teorias sobre a liberdade de

expressão convergem ao lhe atribuir a função de 'guardiã da democracia' ainda que reconheçam que há outros fundamentos relevantes.

Na visão liberal, rejeita-se qualquer regulação pelo Estado, inclusive a autorregulação regulada. Em (GRAÇA, 2019, p. 56), é apontado o risco do efeito silenciador pelo receio da sanção penal ou administrativa, citando a professora Clarissa Piterman Gross:

> A imprecisão de critérios acerca do que é ou não verdadeiro ou cerca de como distinguir entre um juízo de valor, ou uma opinião, de uma proposição de fato, levaria muitas pessoas a se calar. E isso, todas as variáveis consideradas, levaria a mais prejuízos do que vantagens para o debate público de qualidade.

Nesta linha, defende-se que o foco deveria ser a educação da sociedade, alertando sobre os riscos das *fake-news*, e incentivando a participação na checagem e filtragem das notícias. Neste sentido, as agências de checagem de fatos desempenham um papel central, sendo essencial a independência destas. Por outro lado, o próprio TSE pode atuar proativamente, tanto para educar a sociedade, quanto para facilitar a atuação de entidades públicas e privadas, como é feito no Programa Permanente de Enfrentamento à Desinformação da Justiça Eleitoral. Alguns especialistas defendem que a legislação e os mecanismos de controle atuais da regulação estatal são suficientes para o controle das *fake-news*. Nesse sentido se pronunciou o Min. Luiz Fux, em sede de *obiter dictum* na ADI 4451, ao ressaltar a relevância da educação, do jornalismo profissional e da cidadania responsável.

Por outro lado, como assinala Frederico Alvim (ALVIM, 2023, p. 232), a atual horizontalidade da comunicação digital tem permitido a deturpação do sentido da liberdade de expressão, o que reforça a necessidade da regulação:

> Tendo em vista essa amálgama de fatores negativos que transformou a internet em terra sem lei, torna-se indispensável a criação de regras de controle e limitação de uso desse espaço, (...) torna recomendável que a regulação se desenvolva de modo concomitante entre plataformas e Estado.

A situação se agrava ainda mais com o surgimento de tecnologias. A evolução dos modelos LLM (*Large Languade Models*) tais como ChatGPT da OpenAI, expandiu enormemente a fronteira de

aplicações e o alcance da IA. A possibilidade de criação de notícias de forma estocástica, em escala massiva e com alto nível de persuasão elevou o risco de geração e difusão de *fake-news*, como descrito em (HOES *et al*., 2023). O CHatGPT, e.g, pode ser usado como arma na geração de falsas narrativas em fraudes ou regimes autoritários, conforme anotado em (BREWSTER *et al*,2023). Algumas iniciativas vêm sendo feitas para identificar textos criados por chatbots. A OpenAI desenvolve um classificador de texto gerado por IA, com um desempenho ainda baixo. Entretanto, o uso da autenticação ativa para identificação mais precisa de textos gerados por IA, com técnicas de marca d`água digital, foi proposto, e empresas como OpenAI, Microsoft, Google, Meta, Amazon, Anthropic e Inflection, se uniram em um projeto de desenvolvimento dessas técnicas. Por ora, estes textos, dificilmente identificados pelo usuário, conforme (PENNYCOOK *et al*, 2020), podem fomentar o chamado 'efeito de verdade implícita' onde conteúdos não rotulados ou validados são considerados verdadeiros. Na outra ponta, pode produzir um efeito denominado de "Dividendo do Mentiroso" (CHESNEY e CITRON, 2018), onde até conteúdos genuínos são refutados.

Temos, portanto, um desafio de coibir a disseminação de notícias falsas que afetem a isonomia do processo eleitoral, resguardando o direito essencial da liberdade de expressão. Como assevera Frederico Alvim (ALVIM *et al*, 2023, p. 234):

> (...) toda tentativa e regulação do espaço público democrático encerra, em si, o desafio de preservar a liberdades individuais sem colocar em risco a proteção de dos direitos fundamentais e a própria estabilidade do regime democrático. Consequentemente, os controles normativos e judiciais não podem convolar-se em práticas de censura.

2.1. Desafios e riscos da autorregulação regulada

Pelas vantagens aparentes, iniciativas legislativas têm surgido para a implantação da autorregulação regulada. Entretanto, para evitar riscos, como o uso político, o efeito silenciador ou ataques sistêmicos, o sucesso da autorregulação regulada depende de uma discussão profunda e de uma especificação detalhada. Como descrito em (MARANHÃO *et al*, 2021), a *Electronic Frontier*

Foundation (EFF) emitiu o documento *Article* 19 com princípios a serem considerados por legisladores ao desenvolver, adotar e analisar normas, políticas e práticas que tratam da responsabilidade dos intermediários por conteúdo de terceiros: a vedação legal da responsabilização por conteúdos produzidos por terceiros; a necessidade de ordem judicial para a remoção de conteúdos; a clareza das requisições de restrição de conteúdos; a observância dos princípios da necessidade e proporcionalidade e do devido processo em todas leis, políticas e práticas de restrição de conteúdo; além da previsão legal de transparência e prestação de contas.

Um desafio para a implementação da autorregulação regulada decorre do custo da infraestrutura necessária. Ao contrário do que se pode pensar, o uso de máquinas não é suficiente para realizar essa tarefa. Mesmo na identificação de material de abuso sexual infantil, em inglês *Child Sexual Abuse Material* (CSAM), é necessária a intervenção e revisão humana. Conforme descrito em (JASPER, 2022), a identificação de CSAM na Google é feita em duas etapas, pela aplicação automática de tecnologias de correspondência de HASH[1] e Inteligência Artificial (IA), seguida de uma revisão por equipe treinada, feita por amostragem, com critérios de prioridade. A revisão de todo o conteúdo detectado seria inviável.

1. HASH é a transformação de uma grande quantidade de dados em uma pequena quantidade de informações, que permite a identificação rápida e a verificação de integridade.

Erros de identificação do conteúdo são um problema grave. Em (MULLIN, 2022), são descritas duas verificações errôneas de CSAM pela Google, em que algoritmos sinalizaram erroneamente fotos tiradas por pais como imagens de abuso infantil. As consequências de erros desse tipo são graves. Em um estudo da Facebook sobre 150 contas denunciadas por CSAM, descobriu-se que 75% enviaram imagens não maliciosas. O LinkedIn analisou 75 contas denunciadas pelo PhotoDNA, software de CSAM, e apenas em 31 dos casos confirmou-se o CSAM. Segundo a Comissária da União Europeia, Ylva Johnasson, a precisão dos algoritmos de CSAM é de 88%, antes da revisão humana. Essas taxas não são seguras, considerando o volume de conteúdo publicado.

O receio dos efeitos de erros deste tipo é maior ainda em autocracias ou em democracias em retrocesso. Para a regulação de conteúdo eleitoral, a dificuldade de treinar e padronizar a análise de revisão é agravada pela subjetividade da classificação de conteúdo com teor político, muitas vezes ambíguo. Em (LIM,2018), foram avaliados o desempenho e a confiabilidade das classificações de conteúdo político por agências de "checagem de fato". Para isso, foi feita a análise da convergência dos vereditos de duas agências, Fact Checker e Polifact. A classificação da Polifact é refinada em 6 níveis, variando entre "Verdade" e "Nitidamente falso" (*Pants on fire*). A análise usou somente declarações verificadas por ambas as agências, apenas 10% do conjunto de verificações. As classificações convergiram para 49 notícias, de um total de 77. O desempenho observado foi bom para falsidades absolutas ou verdades óbvias. Entretanto, a taxa de concordância foi baixa para afirmações ambíguas. Os resultados demonstram que a verificação de fatos políticos é um desafio.

Essa ambiguidade de classificação de conteúdo dificulta a análise de desempenho do sistema de autorregulação. A contabilização dos acertos e erros de classificação de conteúdo com teor político é muito difícil, uma vez que é impossível revisar todo conteúdo publicado e inexiste uma "classificação absoluta". Qualquer revisão tem um grau de subjetividade. Dessa forma, existe o risco da legislação impor sanções descabidas a provedores, caso não reconheça que nenhum sistema de classificação é infalível, ou aplique requisitos muito elevados de desempenho.

Este risco de sanção pode provocar um efeito silenciador, pelo ajuste dos parâmetros de classificação de conteúdo, elevando os casos de remoção indevida de conteúdo verídico. Podemos visualizar como o efeito silenciador seria operado na etapa automática de detecção, com base na análise da Curva Característica de Operação do Receptor[2]. Para reduzir os Falsos Negativos (conteúdo ilegal não detectado), a sensibilidade seria elevada pelo ajuste de parâmetros dos algoritmos, tendo como efeito inevitável o aumento dos Falsos Positivos (conteúdo legal erroneamente classificado como ilegal).

A complexidade computacional e o desempenho dos métodos de deteção automática de *fakes-news* também são aspectos críticos. Em (HAMED *et al*, 2023), foi feita uma revisão dos dois modelos de detecção de *fake-news*: 1) A checagem de fatos baseada no conhecimento (*knowledge-based*), onde para cada consulta é feita uma pesquisa em um universo aberto de informações e a classificação por software é feita de forma não supervisionada, sem treinamento prévio; e 2) A checagem baseada em atributos (*feature-based*), onde os métodos supervisionados que empregam IA são treinados a partir de bases de dados com conteúdo rotulado.

Podem ser usados atributos de conteúdo da mensagem, como título, texto e imagens, ou baseados em características de contexto social, que são categorizados em dois tipos: 1) Atributos de rede, que são informações de redes de propagação, interação e difusão, como padrões de evolução temporal e espacial da difusão pelas redes sociais, compartilhamentos, visualizações, comentários e curtidas; 2) Atributos de usuário, envolvendo a análise dos perfis dos usuários, como credibilidade, número de seguidores; características de comportamento, nome, data de criação, dados de localização e descrição da conta.

2. É uma representação gráfica do desempenho de um sistema classificador binário, cujos valores de FP (Falso Positivo) e FN (Falso Negativo) variam à medida que parâmetros ou limiares de discriminação são ajustados.

Os melhores resultados são observados na abordagem de análise baseada em atributos de conteúdo, ou seja, o título e o corpo da notícia. Para o desenvolvimento desses sistemas, no caso de notícias com teor político, são utilizados conjuntos rotulados de dados de treinamento e teste que empregam informações de agências de checagem, como a Polifact.com.

Nesses modelos existe forte correlação positiva entre o tamanho da base de dados e o desempenho do sistema, o que reforça a necessidade de emprego de bases com muitas amostras. Segundo (HAMED *et al*, 2023), a falta de dados é o principal problema na identificação de notícias falsas. Bases pequenas e desatualizadas não refletem todo o universo de notícias. Outro problema é a diversidade da forma de classificação. Algumas bases usam rótulos binários (reais ou falsos), enquanto outras usam uma classificação refinada com vários graus de desinformação, o que dificulta a fusão de bases distintas para a formação de uma base maior de treinamento e teste. Uma alternativa seria usar conjuntos de dados em idiomas diferentes. Em (DEMENTIEVA *et al*, 2023), foi proposto um modelo capaz de identificar, com neutralidade de idioma, notícias acerca de conteúdos globais. Essa abordagem seria possível para a identificação de alguns conteúdos abusivos, que apresentam uma semelhança global, entretanto não funcionaria na identificação de notícias políticas de interesse local. Portanto, modelos de checagem de conteúdo com teor político devem ser treinados com conjuntos de dados de agências de checagem de fato locais, nacionais. Neste ponto, observamos a grande relevância das agências de checagem locais no desenvolvimento de modelos automáticos de identificação de *fake-news*.

Outro aspecto relevante é a escolha de bases confiáveis para o treinamento dos sistemas de checagem automática. O emprego de checagens feitas por agências ligadas a partidos políticos ou ao governo fatalmente produziria um viés nos sistemas de verificação. Nesse sentido, seria recomendável que iniciativas legislativas de autorregulação de conteúdo previssem requisitos para as bases de treinamento. Cabe citar como exemplo, o *Article* 19 da DSA, que impõe requisitos de transparência de financiamento, competência e experiência para as *Trusted Flaggers*, entidades notificadoras de conteúdo abusivo, que podem ser públicas ou privadas. No caso das bases de treinamento para

sistemas de autorregulação de conteúdo político, a vinculação das agências ao governo seria um risco. Portanto, seriam recomendáveis requisitos como apartidarismo comprovado, financiamento independente, desvinculação com o governo, participação de entidades de jornalismo e equipes de análise com perfil profissional, comprometidas com a ética, além da acreditação de organismos internacionais de transparência. A Tabela 1 lista as maiores agências de checagem de fatos nacionais, cobrindo aspectos como número de checagens, assuntos, financiamento, código de conduta e rótulos usados. Observa-se que as bases, bem menores que a base americana Polifact, com mais de 20 mil amostras, são insuficientes para o treinamento de modelos automáticos, sobretudo se forem selecionadas apenas notícias políticas. A junção das bases é um desafio, uma vez que utilizam rótulos distintos.

Na análise de desempenho dos classificadores, foram selecionados os trabalhos que usaram bases de checagem de notícias políticas internacionais, como a FakeNewsNet e PolitiFact.com. Na análise da precisão, foi observada uma média de intervalos de 67%. Em (HOES *et al*, 2023), o uso do ChatGPT 3.5 na identificação de fake-news foi avaliado e observou-se uma precisão baixa de 69%. Os resultados são considerados ruins, ficando pouco acima do desempenho de 50% do classificador aleatório.

Tabela 1: Agências de checagem de notícias nacionais, dados dos próprios sites em 10/10/2023. Elaboração própria

Agência	Rótulos	Tamanho	Assuntos	Financiamento	Código de conduta
LUPA[3]	Falso, Verdadeiro, Contraditório, Ainda é cedo pra dizer, Insustentável, Exagerado Subestimado, Verdadeiro, mas; De olho	5.742	Política. Eleições, Saúde. Ciência, Meio Ambiente	-Contribuições voluntárias; -Parcerias com Facebook Instagram, Whatsapp; Editora Alvinegra, revista Piauí; - Hospedagem no UOL	-IFCN ; The Trust Project; Third-Party Fact-Checking Program
Aos Fatos[5]	Falso, Não é bem assim, Verdadeiro	3.240	Política. Eleições, Saúde. Ciência, Meio Ambiente	-Remuneradas por parceiras com Meta, Telegram, Kwai e portal Terra; Escriba, serviço proprietário de transcrição automática de áudios e vídeos; - Contribuições voluntárias	-IFCN
Comprova[6]	Enganoso, Falso, Comprovado, Sátira	Em torno de 900	Política. Eleições, Saúde. Ciência, Meio Ambiente	-Coalizão de 41 veículos de comunicação; apoio de estudantes da FAAP, sem fins lucrativos, liderada pela Abraji; Google News Initiative, Whastapp e o Meta Journalism Project	-Conselho editorial e decisões coletivas; Termo de Compromisso dos Jornalistas; Revisão por pares; Cross-checking
Agence France-Presse[7]	Falso, Enganoso, Checado	1.970	Política. Eleições, Saúde. Ciência, Meio Ambiente, Esporte	- Governo francês, dezenas de parcerias com grupos de mídia e plataformas.	-Estatuto e normas de boas práticas próprios, sujeito a leis francesas.

3. Disponível em https://lupa.uol.com.br/jornalismo/, acessado em 20/10/2023.
4. Código de conduta internacional proposto pela IFCN (International Fact-Checking Network).
5. Disponível em https://www.aosfatos.org/, acessado em 20/10/2023.
6. Disponível em https://projetocomprova.com.br/, acessado em 20/10/2023.
7. Disponível em https://checamos.afp.com/, acessado em 20/10/2023.

Fakes News podem ser multimodais, contendo áudios e vídeos além de texto. O termo *Deep Fake* (DF), junção de *Deep Learning* (DL) com *Fake*, foi usado inicialmente para descrever conteúdos realistas de vídeo criados com uso de DL. Com o surgimento das DF, o interesse para desenvolver um padrão aberto de autenticação ativa de mídias ressurgiu, como na iniciativa conjunta de empresas como Adobe, Microsoft, Intel e BBC, denominada "*The Coalition for Content Provenance and Authentication* (C2PA)". Entretanto, por ora, os métodos de identificação aplicam a autenticação passiva, com base na análise dos conteúdos das mídias.

O principal problema dos modelos de detecção de DF, treinados e testados em uma mesma base, é a incapacidade de generalização. Desempenhos satisfatórios para uma base de teste específica não necessariamente seriam observados em uma aplicação real, como a autorregulação de conteúdo. Uma forma mais realista de avaliação é feita por meio de torneios, nos quais competidores geram mídias falsas para construção das bases, e na sequência, testam reciprocamente seus modelos pré-treinados. Em (ALTUNCU *et al*, 2022), foi conduzida uma meta revisão de métodos propostos em torneios para detecção de vídeos falsos. No torneio *Deepfake Detection Challenge* (DFDC) de 2020, dentre os 2.114 participantes, o melhor modelo teve precisão de apenas 65,18%, considerada muito baixa. Para a detecção de *Deep fakes* em áudios, analisou-se o desempenho dos métodos no torneio *Audio Deep synthesis Detection challenge* (ADD) que inclui trilhas contendo áudios mais reais, com inserção de ruído ou com apenas um trecho falso. Conforme a revisão feita em (Yi, 2022), na trilha de desafio com áudio ruidoso, a melhor Taxa de Erro Equivalente (EER) foi de 21,7%, indicando taxas de Falso Positivo e Falso Negativo em elevado patamar. Estes resultados mostram as dificuldades dos sistemas atuais de detecção de DF em vídeo ou áudio, com desempenhos ainda baixos para aplicações práticas.

3. Análise da regulação de conteúdo eleitoral prevista no PL 2630/2020

O Projeto de Lei no 2.630/2020, aprovado no Senado em 30/06/2020, não incluía previsão de moderação específica de conteúdo eleitoral, mas previa a moderação de conteúdos indefinidos, que

implicassem em "dano imediato de difícil reparação". Como destacado em (CAMURÇA, 2021), a Coalizão Direitos na Rede se opôs à responsabilização objetiva das plataformas por conteúdos produzidos por terceiro, contrária ao Marco Civil da Internet, e alertou sobre os riscos de cerceamento da liberdade de expressão decorrentes da subjetividade da classificação da informação, associada ao receio de sanção e à indefinição de critérios de aplicação de medidas de moderação. Em 2021, foi criado um grupo de trabalho na Câmara dos Deputados para analisar e elaborar parecer ao PL. A proposta inicial foi totalmente modificada, ampliando significativamente o seu escopo e incluindo inúmeros dispositivos de regulação de plataformas digitais, como observado em (RAIS, 2022), como as restrições operacionais aos serviços de mensagem instantânea, que embora visem dificultar o intercâmbio de conteúdos inverídicos, limitam também a difusão de informações verídicas.

O texto previa a criação de uma entidade autônoma de regulação, responsável por fiscalizar o cumprimento das regras previstas no PL. Muita preocupação surgiu em torno da isenção, composição e atribuições desta entidade. Entretanto, a solução adotada (POMPEU, 2023) foi retirar a previsão da entidade de regulação do PL. Dessa forma, o Comitê Gestor da Internet (CGI.br) restou como a única entidade prevista no PL, responsável apenas por atividades de gestão e planejamento, como elaboração de estudos e diretrizes. Na ocasião, o próprio CGI.br alegou que não teria a capacitação necessária para a fiscalização.

Esta análise se baseou na versão final do PL, apresentada em 27/04/2023 (PL 2.630-Câmara, 2020), com 47 páginas e 60 artigos distribuídos em 16 capítulos. Apesar de ter ficado conhecido como "PL das *fake-news*", o PL trata de temas menos polêmicos, como a exclusão de conteúdos abusivos associados a racismo, induzimento ao suicídio e automutilação, terrorismo, crimes cometidos contra mulheres, crianças e adolescentes, e distribuição de pornografia infantil. Neste aspecto, a norma se assemelha à NetzDG. O escopo deste trabalho é limitado à análise da moderação de conteúdo relativo ao processo eleitoral. Portanto, apenas os capítulos II, III, XIII e XV do PL, diretamente associados à regulação de conteúdo, foram tratados.

A moderação de conteúdo pode ocorrer por iniciativa da própria plataforma, ou seja, uma autorregulação interna, ou decorrer de uma notificação externa. Ao prever a moderação por notificação do usuário, a norma implicitamente admite que a autorregulação interna não é infalível, e que conteúdos falsos não detectados podem vir a ser publicados. Para a autorregulação, a definição dos tipos de conteúdo proscritos é feita no Capítulo II (Da Responsabilização dos Provedores), que trata das responsabilizações dos provedores. A obrigação de tratamento de notificações do usuário, apesar de também impor responsabilidades ao provedor, é tratada no Capítulo III.

3.1. Da responsabilização dos provedores (Capítulo II do PL 2630/2020)

O caput do art 7º do PL (PL 2.630-Câmara, 2023) associa risco sistêmico ao tratamento e prevenção de ataques a algoritmos e protocolos do sistema de regulação. O § 2º do mesmo artigo vai além e define os tipos de conteúdo proscritos associados aos riscos. O inciso I proíbe os conteúdos ilegais listados no art. 11, que trata da autorregulação decorrente do Dever de Cuidado. O inciso III, de forma redundante, censura conteúdos abusivos já previstos no art. 11. Esses conteúdos abusivos proscritos englobam atos de terrorismo, relativos à violência contra a mulher, racismo, incitação à prática de crimes contra idosos, crianças e adolescentes, oposição a medidas sanitárias de emergência pública, e induzimento ou auxílio a suicídio e automutilação.

O inciso IV do § 2º do art. 7º, entretanto, proscreve conteúdos que representem riscos ao Estado democrático de direito e à higidez do processo eleitoral[8], que pode ser maculada de várias formas. No contexto do PL, de forma implícita, a ameaça está associada à divulgação de notícias falsas. O próprio TSE entende que notícias inverídicas têm o potencial de macular a higidez do processo eleitoral (TSE, 2023). Portanto, a norma objetiva a prevenção e repressão da difusão de notícias falsas que afetem o discernimento do eleitor, e, consequentemente, os resultados das eleições. O inciso III do art. 8o, também de forma implícita, prevê a possibilidade de remoção de conteúdo, inclusive, como foi visto, de notícias eleitorais falsas, que afetem a higidez do processo eleitoral[9].

A Seção IV aborda o tratamento de situações de Risco Iminente de Danos. Conforme o caput do art. 12[10], a iminência se configura quando ocorrem os riscos listados no art. 7º, ou quando há negligência ou insuficiência da ação do provedor. Nessa situação é instaurado um Protocolo de Segurança. Como observado, os riscos definidos no art. 7º incluem a ameaça à higidez do processo eleitoral, como a difusão de *fake-news*. A norma não define regras para a aplicação de medidas

8. Art. 7º Os provedores devem identificar, analisar e avaliar diligentemente os riscos sistêmicos decorrentes da concepção ou do funcionamento dos seus serviços e dos seus sistemas relacionados, incluindo os sistemas algorítmicos. (...) § 2º A avaliação abrangerá especificamente em cada um dos serviços dos provedores e considerará os riscos sistêmicos, tendo em conta a sua gravidade e probabilidade de ocorrência, e incluirá, no mínimo, a análise dos seguintes riscos: (...) IV – ao Estado democrático de direito e à higidez do processo eleitoral. (Grifo nosso).
9. Art. 8º Os provedores adotarão medidas de atenuação razoáveis, proporcionais e eficazes, direcionadas aos riscos sistêmicos de que trata o art. 7o : (...) III - adaptar os processos de moderação de conteúdos, incluindo a rapidez e a qualidade do processamento de notificações e quando necessário aplicar remoção de conteúdo, garantidos os procedimentos previstos no Capítulo III. (Grifo nosso).
10. Art. 12. Quando configurada a iminência de riscos descritos no art. 7º, ou a negligência ou insuficiência da ação do provedor, poderá ser instaurado, na forma da regulamentação e por decisão fundamentada, protocolo de segurança pelo prazo de até 30 (trinta) dias, sem prejuízo de outras medidas legais cabíveis, procedimento de natureza administrativa cujas etapas e objetivos deverão ser objeto de regulamentação.

do protocolo de segurança. O §1º do art. 15[11] prevê a possibilidade de remoção de conteúdo, e o § 4º prevê a aplicação de sanção no caso de abuso na aplicação das medidas. Portanto, no tratamento de Risco Iminente de Danos, há possibilidade de remoção de conteúdo durante a vigência do Protocolo de Segurança, decorrente de ameaça à higidez do processo eleitoral.

Cabe destacar que a mera instalação do protocolo de segurança pode gerar um efeito silenciador, devido ao receio de sanções administrativas (art 15) ou da responsabilização civil (art. 13). Sua instauração advém da caracterização de riscos sistêmicos ou da insuficiência da ação do provedor.

A caracterização de omissão depende da necessária classificação de conteúdo. Conforme descrito em (LIM, 2018), a checagem da notícia é, em muitos casos, uma tarefa subjetiva, e até as agências de checagem de fatos se confundem. Da mesma forma, seria natural observar divergências entre a classificação feita pela plataforma e um órgão revisor. O PL também não define critérios objetivos para caracterizar insuficiência de ações. O uso de critérios baseados em fórmulas de desempenho mínimo seria inviável, devido à impossibilidade de revisão de todo conteúdo. Apenas uma fração dos Falsos Negativos poderia ser identificada pela análise das notificações procedentes. Para a contabilização dos Falsos Positivos, haveria necessidade de uma revisão de todo o conteúdo por uma entidade externa independente. A dificuldade de classificação de conteúdo político e os baixos desempenhos observados dos sistemas de detecção de fake news e deep fakes, elevam o risco do efeito silenciador da norma.

11. Art. 15. Os provedores deverão produzir relatórios específicos das suas ações envolvendo o protocolo de segurança, conforme regulamentação. §1º Conteúdos tornados indisponíveis em razão do protocolo de segurança deverão ser armazenados pelos provedores atingidos, pelo tempo determinado em regulamentação, para fins de análise posterior. (...) § 4º Configurado abuso na aplicação das medidas previstas no protocolo de segurança, os provedores ficam sujeitos às sanções previstas nesta Lei. (Grifo nosso).

3.2. Da moderação por notificação pelo usuário (Capítulo III do PL 2630/2020)

O Capítulo III aborda os procedimentos associados à notificação pelo usuário de conteúdos ilegais, conforme art. 17[12] e caput do art. 18. Os Termos de Uso são regras próprias estabelecidas pelo provedor, que incluem a aplicação de medidas de moderação. O Capítulo III não define um rol de conteúdos proscritos, o que permite que a ameaça à higidez do processo eleitoral seja incluída a posteriori, com base nas diretrizes do CGI.br.

A execução em duas etapas, uma automática e outra realizada por equipes treinadas de checadores de conteúdo, implica em um custo elevado para as plataformas. Comparada à autorregulação interna, que se aplica a todo conteúdo publicado, a moderação do conteúdo decorrente de notificação, com volume de checagens menor, é mais exequível.

A moderação por notificação está sujeita a ataques sistêmicos, que podem inviabilizá-la. O direito de revisão da moderação está previsto no Capítulo III, com prazos para notificação e resposta definidos no Código de Conduta. A norma não define requisitos de permissão, identificação ou de autenticação dos autores nas notificações de usuário. Portanto, considera-se que denúncias anônimas podem ser realizadas. Pela norma proposta, o autor do conteúdo deverá ser informado apenas após a aplicação da regra de moderação, e a norma não define se a aplicação da medida de moderação ocorre quando da notificação, ou somente após a análise do conteúdo. Nesta análise consideraremos as duas situações.

12. Art. 17. O procedimento de moderação de conteúdo e de conta deve observar o normativo vigente e ser aplicado com equidade, consistência e respeito ao direito de acesso à informação, à liberdade de expressão e à livre concorrência. Parágrafo único. Os termos de uso, quanto à moderação de conteúdo e de contas, devem sempre estar orientados pelos princípios da necessidade, proporcionalidade e não discriminação, inclusive quanto ao acesso dos usuários aos serviços dos provedores. Art. 18. Após aplicar as regras contidas nos Termos de Uso que impliquem moderação de conteúdos, incluindo aquelas envolvendo alteração de pagamento monetário ou publicidade de plataforma, os provedores de redes sociais e de mensageria instantânea devem, ao menos: (...) (Grifo nosso).

Na Situação 1, a aplicação da medida de moderação é feita em curto prazo, logo após a notificação, e a checagem do conteúdo é realizada depois. Podemos imaginar um ataque pelo qual todo conteúdo eleitoral publicado, verídico ou não, seria notificado por um ou mais adversários. A depender do volume de notificações, e considerando uma fila FIFO (*First In First Out*) de atendimento por ordem de entrada, o tempo médio para restabelecimento do conteúdo verídico pode ser tão alto que prejudique o processo eleitoral, limitando a sociedade das informações necessárias à decisão eleitoral. Uma vez restabelecida a publicação, seu resumo criptográfico (HASH) seria inserido em uma lista pública, e uma consulta rápida pelo HASH dispensaria novas checagens. Podemos antever, nesse caso, um ataque sistêmico em que quase a totalidade do conteúdo eleitoral publicado em plataformas seria denunciado, excluído, analisado e, na maioria dos casos, restabelecido, uma única vez. Haveria, portanto, um prejuízo à divulgação de informações necessárias à decisão do eleitor.

Na Situação 2, a medida de moderação somente é aplicada após a devida checagem do conteúdo. Podemos identificar uma vulnerabilidade favorável a partidos que se utilizam de *fake-news* para manipular eleitores. O ataque visa elevar ao máximo o tempo médio de checagem de conteúdos inverídicos. Considerando uma fila FIFO de atendimento por ordem de entrada, os autores de *fake-news* poderiam, de forma anônima, realizar um volume elevado de notificações, mesmo de conteúdos não eleitorais, para elevar o tempo médio de análise. O ataque maciço, do tipo Negação de Serviço (*Denial of Service* (DoS)[13] poderia até mesmo travar o sistema de checagem. Dessa forma, conteúdos inverídicos permaneceriam publicados por longos períodos, aguardando a devida verificação. Mesmo que fossem adotados critérios de prioridade distintos do FIFO, e.g. baseados em atributos de rede, o ataque poderia ser adaptado, através da notificação de notícias com este perfil.

13. DoS (*Denial of Service*): O ataque de negação de serviço é uma tentativa de indisponibilizar os recursos de um sistema para os seus usuários, pela sua invalidação por sobrecarga, geralmente forçando o sistema vítima a reinicializar ou consumir todos os recursos de forma a obstruir o seu serviço o ou meio de comunicação.

Os ataques descritos, na medida em que criam uma demanda excessiva às equipes de checagem, também afetam a autorregulação interna, que também depende da análise dessas equipes. Algumas soluções podem ser adotadas para mitigar esse risco, como a autenticação prévia do usuário para a notificação. Nesse caso, os provedores poderiam suspender ou bloquear contas de robôs ou usuários com notificações fraudulentas reincidentes. Uma alternativa seria restringir a permissão de notificação, e.g. a partidos políticos. A responsabilização legal pela notificação maliciosa, além da suspensão ou bloqueio de contas, também mitigaria o risco de ataque. Cabe destacar que o art. 326-A do CE tipifica os crimes de denunciação caluniosa por dar causa indevida à investigação, inclusive administrativa.

3.3. Das sanções administrativas (Capítulo XIII do PL 2630/2020)

O art. 47[14] prevê que o descumprimento de quaisquer de suas normas está sujeito à sanção administrativa. Em uma análise rápida da

14. Art. 47. Os provedores, em razão das infrações cometidas às normas previstas nesta Lei, ficam sujeitos às seguintes sanções administrativas, aplicáveis de forma isolada ou cumulativa:
I - advertência, com indicação de prazo para adoção de medidas corretivas;
II - multa diária, observado o limite total a que se refere o inciso III;
III – multa simples, de até 10% (dez por cento) do faturamento do grupo econômico no Brasil no seu último exercício ou, ausente o faturamento, multa de R$ 10,00 (dez reais) até R$ 1.000 (mil reais) por usuário cadastrado do provedor sancionado, limitada, no total, a R$ 50.000.000,00 (cinquenta milhões de reais), por infração;
IV - publicação da decisão pelo infrator;
V - proibição de tratamento de determinadas bases de dados; e
VI – suspensão temporária das atividades.

exceção prevista na primeira parte do caput do art. 48[15], poder-se-ia pensar que a intenção do legislador fosse blindar as atividades de autorregulação interna do risco de sanções. Nesse caso, os dispositivos previstos no Capítulo II não teriam força coercitiva. Entretanto, em uma análise detalhada, observa-se que a exceção se aplica apenas a processos de moderação de conteúdos proscritos específicos, definidos nos Termos de Uso, por iniciativa própria do provedor, ou seja, além dos conteúdos proscritos obrigatórios previstas na lei, como aqueles listados no art. 7o § 2º. Portanto, não haveria exceção para aplicação de sanção no caso de descumprimento das obrigações de análise e atenuação de riscos sistêmicos, incluindo a remoção de conteúdo prevista no inciso III do art. 8º. A segunda parte do caput do art. 48 prevê a possibilidade de aplicação da sanção pelo descumprimento sistemático das obrigações previstas no Capítulo III. Deve-se destacar novamente a indefinição acerca dos critérios para a caracterização do "descumprimento sistemático". Dependendo de um viés punitivo do órgão regulador, a mera reincidência poderia caracterizá-lo.

---

§ 1º Após procedimento administrativo que possibilite a oportunidade de ampla defesa, as sanções serão aplicadas de forma gradativa, isolada ou cumulativa, de acordo com as peculiaridades do caso concreto e considerados os seguintes parâmetros e critérios:I - a gravidade e a natureza das infrações e a eventual violação de direitos;
II - a boa-fé do infrator;
III - a vantagem auferida pelo infrator, quando possível estimá-la;
IV - a condição econômica do infrator;
V - a reincidência;
VI - o grau do dano;
VII - a cooperação do infrator;
VIII - a pronta adoção de medidas corretivas; e
IX - a proporcionalidade entre a gravidade da falta e a intensidade da sanção.
§ 2° Antes ou durante o processo administrativo do § 1º, poderão ser adotadas medidas preventivas, incluída multa cominatória, observado o limite total a que se refere o inciso III do caput, quando houver indício ou fundado receio de que o provedor:
I - cause ou possa causar dano irreparável ou de difícil reparação; ou
II - torne ineficaz o resultado do processo.

15. Art. 48. As sanções não serão aplicadas a processos de moderação de conteúdos específicos por iniciativa própria dos provedores e de acordo com seus termos de uso, salvo em caso de descumprimento sistemático das obrigações previstas no Capítulo III. (Grifo nosso).

O § 1º-III do art. 11 prevê que o descumprimento das normas, e.g. por omissão na aplicação de medida de moderação, é passível de sanção. No § 2º do art. 11[16], há uma salvaguarda da aplicação da sanção para casos isolados, diante dos esforços adotados. Porém a norma é vaga e não define a tolerância máxima de casos ou de reincidência.

Por fim, cabe destacar que, conforme o Marco Civil da Internet (Lei nº 12.965/14) e a Lei Eleitoral (LE, art. 57-F), somente devem ser aplicadas penalidades aos provedores por descumprimento de ordem da justiça eleitoral.

3.3.1. Da ampla defesa

A doutrina e a jurisprudência reconhecem que garantias do Direito Processual Penal, como a ampla defesa prevista em diversas normas administrativas, podem ser aplicadas às sanções administrativas. O § 1º Art. 47 do PL prevê a ampla defesa, antes da aplicação das sanções, no procedimento administrativo. Entretanto, o § 2º permite que, antes ou durante o processo administrativo de revisão previsto no § 1, seja imposta multa cominatória. Ressaltamos que a multa diária e a multa simples, previstas no art. 47, têm uma clara natureza cominatória, com o objetivo de forçar o cumprimento da norma. Dessa forma, o § 2º enfraquece a garantia de ampla defesa prévia.

16. Art. 11. Os provedores devem atuar diligentemente para prevenir e mitigar práticas ilícitas no âmbito de seus serviços, envidando esforços para aprimorar o combate à disseminação de conteúdos ilegais gerados por terceiros, que possam configurar (...) § 1º A avaliação do cumprimento do disposto no caput será feita tendo em vista: (...) III - o tratamento dado ao recebimento de notificações e reclamações. § 2º A avaliação será realizada sempre sobre o conjunto de esforços e medidas adotadas pelos provedores, não cabendo avaliação sobre casos isolados.

3.3.2. Da dosimetria das penalidades

Os critérios para a aplicação de sanção previstos no art. 47 são semelhantes aos do art. 52 da Lei nº 13709 (LGPD). Ressaltamos que o PL não define prazos e procedimentos do processo administrativo. Considerando a especificidade do processo eleitoral, o emprego de regras gerais não seria adequado. Vale destacar, a título de exemplo, que no âmbito da aplicação da LGPD, a ANPD elaborou a Resolução CD/ANPD nº1/2021, com normas específicas para o Processo Administrativo Sancionador.

Conforme descrito em (CHAVES, 2023) e (SANTOS, 2022), diversas agências reguladoras, como ANPD, ANATEL, ANAC, ANS, ANTAQ e ANTT, adotam princípios da regulação responsiva. Ao contrário da estratégia de Comando e Controle, baseada na coerção pela ameaça de sanções, a Regulação Responsiva é flexível, enfatizando a cooperação e negociação, incentivando o cumprimento voluntário em um ambiente de diálogo e interação entre regulador e regulado, combinando estratégias de persuasão e sanção. Nesse sentido, propõe-se um escalonamento gradual das intervenções, conforme o princípio do mínimo suficiente, escalando as sanções na pirâmide de constrangimentos até o cumprimento da norma pelo regulado. Dessa forma, a pirâmide regulatória deve possuir uma hierarquia de sanções e estratégias, com vários graus de intervenção e punições ameaçadoras no topo.

Vale destacar a elevada multa prevista no III do Art 47, de até 10% do faturamento anual do provedor, comparada aos percentuais usados na LGPD, de até 2%, e no Regulamento Europeu de Serviços Digitais (DSA), de até 6%.

A penalidade de advertência possui um caráter pedagógico de alerta, para que o infrator possa corrigir a conduta. A aplicação gradual das sanções, iniciando-se com a advertência, é adotada em diversas normas administrativas de direito público, que condicionam a aplicação de sanções mais graves à reincidência, após a aplicação da advertência, conforme a teoria da regulação responsiva.

As sanções previstas no art. 47 são graduadas em gravidade, indo da advertência à suspensão das atividades. O § 1º do mesmo artigo estabelece critérios para orientar a aplicação da sanção, como proporcionalidade, gravidade da infração, grau do dano, boa-fé e cooperação do infrator, vantagem auferida, condição econômica do infrator, e reincidência. Entretanto, a redação é confusa, pois enquanto o § 1º do

art. 47 prevê a aplicação de forma *gradativa*, isolada ou cumulativa; o Caput do mesmo artigo prevê apenas a aplicação isolada ou cumulativa. Ademais o uso da conjunção "ou", permite uma interpretação alternativa, pela qual a aplicação gradativa não seria necessária.

A falibilidade dos métodos automáticos de detecção de *fake-news* e *deep-fakes*, e a possibilidade de aplicação inicial da multa simples, antes mesmo da advertência, podem gerar um efeito silenciador, tendendo à censura de conteúdo. Na moderação de conteúdo eleitoral, seria recomendável a aplicação gradual das sanções, condicionadas à reincidência. Convém citar como exemplo a restrição de aplicação gradual de sanção prevista no Art 52, § 6º da LGPD[17]. Vale ainda destacar também a Resolução CD/ANPD Nº 4, que definiu a Dosimetria e Aplicação de Sanções Administrativas, de forma detalhada, onde a reincidência é avaliada na definição da sanção.

Por fim, destaca-se que a sanção pela suspensão temporária das atividades dos provedores, prevista no art. 47-VI, viola frontalmente a Constituição Federal, o Marco Civil da Internet e o art. 13 do Pacto de San José da Costa Rica. Segundo (RAIS, 2018, p. 151), o bloqueio de serviços de aplicação não preenche os requisitos de adequação e necessidade, já que outros meios coercitivos podem ser empregados. Neste sentido, a Lei 13.488, de 2017 reformou o art 57-I da Lei de Execução, limitando a sanção de suspensão apenas ao conteúdo veiculado que a descumprir.

3.4. Da regulação dos provedores (Capítulo XV do PL 2630/2020)

O capítulo XV define as atribuições do CGI.br, órgão com composição multissetorial, já constituído anteriormente e com atribuições previstas na Lei nº 12.965 (Marco Civil da Internet), e Lei nº 13.853 (Lei Geral de Proteção de Dados). Dentre as competências atribuídas

17. Art. 52. Os agentes de tratamento de dados, em razão das infrações cometidas às normas previstas nesta Lei, ficam sujeitos às seguintes sanções administrativas aplicáveis pela autoridade nacional: (...) § 6º As sanções previstas nos incisos X, XI e XII do caput deste artigo serão aplicadas: I - somente após já ter sido imposta ao menos 1 (uma) das sanções de que tratam os incisos II, III, IV, V e VI do caput deste artigo para o mesmo caso concreto.

no art. 51 ao CGI.br, típicas de pesquisa e planejamento, destacamos a elaboração de diretrizes, inclusive do Código de Conduta, para a moderação de conteúdo inverídico, prevenção e enfrentamento da desinformação na internet e redes sociais. O Código de Conduta é a ferramenta usada pelos provedores de serviço para aplicação da moderação de conteúdo. As atribuições do CGI.br descritas no PL não englobam as atividades de moderação do conteúdo notificado, nem de fiscalização dos provedores.

Observa-se, portanto, a total indefinição dos órgãos competentes para fiscalizar a autorregulação e a moderação decorrente da notificação, bem como para aplicar sanções. A exclusão da entidade autônoma de regulação do PL das *fake-news* deixou uma grave lacuna na legislação. A omissão legal abre perigoso espaço para a intervenção imparcial do Poder Executivo, delegando por meio de decreto de mera execução, com o argumento de cumprimento da eficácia da Lei, a competência a um órgão do Poder Executivo existente, o que evitaria a necessidade de uma nova lei para sua criação. De fato, a ideia de aproveitar órgãos existentes já foi ventilada. O próprio CGI.br sugeriu a delegação da competência à ANATEL (POMPEU, 2023). Tal possibilidade representa um risco, sobretudo considerando os vícios, como abuso de poder, associados ao instituto da reeleição e ao apadrinhamento de sucessores políticos. Vale ressaltar o poder de influência do Poder Executivo, por meio de nomeações de cargos de chefia e direção. A opção por empregar órgãos total ou parcialmente vinculados ao Poder Executivo nas atividades de moderação de conteúdo e fiscalização dos provedores enfraquece a democracia.

As atividades administrativas de regulação podem demandar elevado volume de trabalho, e a composição de um conselho de tamanho fixo não seria o ideal. Vale destacar que, visando atender a elevada demanda associada à propaganda eleitoral, a Justiça Eleitoral regulamentou o recrutamento de juízes auxiliares, para atuar em ações eleitorais, e das Comissões de Fiscalização de Propaganda Eleitoral (CFPE), para exercer o controle das propagandas pelo poder de polícia. Portanto, uma alternativa natural seria o emprego de uma estrutura administrativa no âmbito da Justiça Federal, composta pelo quadro de servidores, atuando em conjunto com as CFPE. As atividades meramente administrativas seriam realizadas pelos servidores, enquanto as medidas preventivas e sanções seriam determinadas pelos Juízes Eleitorais.

Conforme o art. 47, a natureza da sanção é expressamente administrativa. Com base no Poder de Polícia, dentro da função administrativa da Justiça Eleitoral, o Juiz Eleitoral pode agir de ofício sem necessidade de provocação. Porém, devido à natureza administrativa, e conforme a Súmula-TSE nº 18, o Juiz não teria legitimidade para fixar multa no exercício do poder de polícia, uma vez que não há garantia de ampla defesa. Tal posicionamento do TSE foi reafirmado na Res. 23.608/2019. Entretanto esta posição histórica do TSE foi subitamente modificada pela Res-TSE 23714/2022, art. 2º[18], permitindo multas com natureza de astreintes na atuação do poder de polícia. A essência da posição tradicional do TSE, pela necessidade de processo jurisdicional para aplicação de sanções coercitivas, é a garantia do direito à ampla defesa. Porém, nos processos como Representação Eleitoral e AIJE, a atuação é condicionada a denúncias de partidos políticos, candidatos, coligações ou Ministério Público. Como se observa, inexiste um processo que garanta, ao mesmo tempo, a atuação ativa da Justiça Eleitoral e o direito à ampla defesa. Uma alternativa razoável seria a criação, por meio de lei, de um processo administrativo que assegure o direito à ampla defesa, dentro da função administrativa da Justiça Eleitoral. Vale citar, a título de exemplo, o Processo Administrativo Sancionador definido pela ANPD na Resolução CD/ANPD nº1/2021.

18. Art. 2º É vedada, nos termos do Código Eleitoral, a divulgação ou compartilhamento de fatos sabidamente inverídicos ou gravemente descontextualizados que atinjam a integridade do processo eleitoral, inclusive os processos de votação, apuração e totalização de votos. § 1º Verificada a hipótese prevista no caput, o Tribunal Superior Eleitoral, em decisão fundamentada, determinará às plataformas a imediata remoção da URL, URI ou URN, sob pena de multa de R$ 100.000,00 (cem mil reais) a R$ 150.000,00 (cem e cinquenta mil reais) por hora de descumprimento, a contar do término da segunda hora após o recebimento da notificação.

3.5. Da necessidade de norma específica para regulação de conteúdo eleitoral

Destacamos algumas diferenças significativas entre a moderação de conteúdo abusivo em geral e de conteúdos eleitorais falsos, que apontam para a necessidade de elaboração de uma norma específica para a regulação de conteúdo eleitoral.

A primeira diferença diz respeito à necessidade do anonimato na notificação de conteúdo. No caso de denúncias de conteúdo abusivo, como crimes, o anonimato é recomendável, pois garante a segurança do denunciante e vem sendo adotado em ouvidorias da administração pública, com base no art. 24 do Decreto 9.492/2018[19], em denúncias de racismo ao Ministério da Igualdade Racial, e em denúncias de distribuição de pornografia infantil feitas através da Central Nacional de Denúncias de Crimes Cibernéticos. No caso da notificação de conteúdo eleitoral irregular, que na maioria das vezes não configura crime, o receio de retaliação é menor. A motivação da denúncia de conteúdo abusivo está associada a valores morais e à proteção mútua da sociedade. Na notificação de conteúdo eleitoral, há um forte viés político que pode gerar um elevado engajamento dos eleitores, inclusive através da denunciação caluniosa, já tipificada como crime eleitoral. Ademais, o anonimato em notificação de conteúdo eleitoral irregular pode dar margem a ataques sistêmicos. Cabe citar, como exemplo da motivação eleitoral, a onda de representações eleitorais com a finalidade de mera censura decorrentes da vedação ao anonimato na propaganda eleitoral incluída pela Lei nº 12.034/2009, que somente foi contida pelo disciplinamento feito na Res-TSE 23.551, que limitou o escopo da vedação.

19. Art. 24. As unidades que compõem o Sistema de Ouvidoria do Poder Executivo federal assegurarão a proteção da identidade e dos elementos que permitam a identificação do usuário de serviços públicos ou do autor da manifestação, nos termos do disposto no art. 31 da Lei nº 12.527, de 18 de novembro de 2011.

A segunda assimetria está relacionada à complexidade dos métodos automáticos de verificação de mídias, como áudio e imagem. Conteúdos abusivos, como pornografia infantil, são detectados por métodos de reconhecimento de padrão, e a revisão é feita pela análise visual por equipes treinadas. No caso de conteúdo eleitoral, o desafio é ainda maior, pois a análise demanda a aplicação de métodos para a detecção de *deep-fake* em áudio, imagens e vídeos. Em média, a complexidade computacional dos métodos de checagem de *deep-fake* é superior à complexidade dos métodos de checagem de conteúdo abusivo, e.g. CSAM.

A terceira assimetria refere-se à necessidade de dados para treinamento de sistemas automáticos de checagem. Estudos, como o realizado por (DEMENTIEVA et al, 2023), mostram que um sistema já em uso em língua estrangeira poderia ser empregado no âmbito nacional, com neutralidade de idioma, na regulação de conteúdos globais, e.g. racismo, sem necessidade de novo treinamento. Entretanto, no caso de conteúdo eleitoral e notícias locais, essa abordagem não funcionaria. Portanto, modelos de checagem de conteúdo político devem ser treinados com conjuntos de dados de agências de checagem de fato locais. Nos modelos supervisionados, o desempenho depende do tamanho da base de dados de treinamento. O levantamento realizado mostra que os sites nacionais de checagem de notícias políticas possuem poucas amostras e usam diversos rótulos de classificação das notícias, o que dificulta a fusão de bases de notícia. Portanto, deve-se verificar a viabilidade de treinamento dos sistemas de identificação de conteúdo político com tão poucos dados.

A quarta diferença está relacionada à competência para a regulação de conteúdo. As atividades de regulação de conteúdo abusivo poderiam, sem grandes problemas, ser executadas no âmbito do Poder Executivo. No caso da regulação de conteúdo eleitoral, as atividades administrativas de moderação devem ser realizadas por uma entidade isenta, preferencialmente sem influência do Poder Executivo. Além da competência especial, destaca-se a necessidade de prazos e procedimentos especiais da matéria eleitoral.

O quinto desequilíbrio diz respeito ao efeito da aplicação indevida de uma medida de regulação. Na análise do grau de interferência das medidas, vemos uma assimetria clara. No caso da divulgação de pornografia infantil, e.g., a interferência está limitada a

um grupo pequeno de pessoas, como a família da criança. No caso de notícias eleitorais, a ameaça à liberdade de expressão afeta o processo democrático, e, portanto, toda a sociedade. Esta assimetria explica o grande debate em torno da regulamentação da autorregulação, enquanto não se observa muita oposição à aplicação de censura a notícias abusivas.

4. Conclusão

À medida que surgem tecnologias disruptivas com potencial uso na criação de *fake-news*, eleva-se o risco de interferência no processo eleitoral. Como resultado, percebe-se um crescente esforço de empresas, da sociedade civil e da academia para desenvolvimento de modelos de identificação de conteúdo, e de governos com propostas legais de regulação de conteúdo, inclusive político. Por outro lado, o receio do efeito silenciador, natural em qualquer democracia, desacelera este processo.

Como descrito, o PL 2630/2020 inova de forma perigosa, ao incluir dispositivos legais que, implicitamente, permitiriam a remoção de conteúdo eleitoral. Tal previsão não encontra paralelo no regulamento europeu de serviços digitais (DSA), nem na lei alemã de aplicação das redes (NetzDG). A previsão de multas elevadas, combinada com a dificuldade de classificação de conteúdo político e os baixos desempenhos observados dos sistemas de detecção de *fake news* e *deep fakes*, elevam o risco do efeito silenciador. Ademais, foram identificadas diversas lacunas, como a indefinição do órgão de regulação, e foram apresentados riscos, como possíveis ataques ao sistema de notificação de conteúdo. Com base nos riscos e desafios apontados, podemos fazer recomendações úteis ao debate de uma iniciativa legislativa.

Como mencionado, a classificação de notícias pode ter caráter subjetivo, e os sistemas de detecção de notícias falsas não são infalíveis. Portanto, não pode haver previsão legal de sanção por erro de classificação (Falso Positivo ou Falso Negativo), isolado ou reincidente. O emprego de medidas de desempenho mínimo também é difícil, pois dependeria da contabilização de todos acertos e erros, o que é impraticável diante do volume de informação. Portanto, a única exigência legal razoável seria a comprovação da implantação do sistema de detecção de fake-news textuais, com informações essenciais como a data do último treinamento, bases usadas no treinamento, tamanho das amostras e o desempenho medido no conjunto de teste. A norma

poderia prever uma periodicidade máxima do treinamento do sistema. Para a autorregulação decorrente da notificação, a norma deveria obrigar apenas a implantação do sistema de revisão com equipes treinadas, prevendo também procedimentos e prazos.

A lei também deve prevenir o viés dos sistemas da classificação de *fake-news* textual, cujo treinamento depende de conjunto de notícias rotuladas usado. A exemplo dos requisitos aos *Trusted Flaggers* previstos no *Article* 19 da DSA, seria recomendável que a norma de autorregulação de conteúdo político previsse requisitos para a seleção das bases de notícias usadas nos treinamentos dos sistemas, como a desvinculação, inclusive financeira, de partidos políticos ou governos, além da previsão de auditoria externa independente, checagem cruzada e a participação de jornalistas com compromisso ético-profissional.

A detecção de *deep fakes* é neutra, pois não depende do tipo de conteúdo, portanto, permite o emprego de sistemas globais. Entretanto, a falta de experimentos em larga escala dificulta o dimensionamento de uma estrutura de TI necessária. Uma opção interessante seria a implantação dos sistemas por fases. Em um primeiro momento, seriam verificados apenas os conteúdos notificados, o que é viável, dado que o percentual de mídias criadas por DF é ínfimo. Com um volume menor de análises, esta fase serviria de teste para a implantação da autorregulação interna.

Como foi discutido, a autorregulação de notícias políticas demandaria atividades complexas e demoradas para a implantação, como coleta das bases de notícias, treinamento de modelos supervisionados, testes de desempenho, além de seleção, contratação e treinamento de equipes de revisão. Portanto, uma abordagem pragmática para a regulação estatal, visando a evolução gradual dos serviços, seria a previsão de prazos de implantação escalonados, com requisitos crescentes ao longo do tempo.

A norma deve também mitigar o risco do efeito silenciador decorrente da ameaça de sanções. A autorregulação regulada, conforme o modelo de regulação responsiva, permitiria a cooperação e negociação, combinando estratégias de persuasão e sanção escalonada e gradual, dependente da reincidência. A norma deve definir uma dosimetria para a aplicação de sanções administrativas. No âmbito da Justiça Eleitoral, a atuação ativa e célere, com direito à ampla defesa, poderia ser viabilizada pela definição de um procedimento administrativo específico.

A norma também deve mitigar o risco de ataques sistêmicos à notificação de notícias abusivas, seja pela obrigatoriedade da autenticação prévia do autor, seja pela previsão de sanções por notificações maliciosas.

Seria recomendável igualmente que a norma estabelecesse um critério objetivo para a escolha da medida de moderação. A suspensão ou remoção de conteúdo ou contas deve ser aplicada de forma proporcional, observando atributos de conteúdo, como a veracidade e nocividade, além de atributos de rede ou de usuário, como alcance e propagação da notícia, reincidência e transparência da intenção do autor.

Por fim, a norma de regulação deve prever, com o máximo detalhe possível, a competência das atividades de fiscalização e regulação. Qualquer proposta de órgão regulador minimamente vinculada ao Poder Executivo significaria um retrocesso de décadas, à época anterior à criação da Justiça Eleitoral, que assumiu todas as funções eleitorais em 1932. A regulação de conteúdo eleitoral deve preferencialmente ser executada pela Justiça Eleitoral. Diante da dificuldade de estimar o volume de trabalho das atividades administrativas de fiscalização, a composição e o efetivo do órgão devem ser flexíveis. Uma alternativa natural seria o emprego de uma estrutura administrativa no âmbito da Justiça Federal, composta pelo quadro de servidores, atuando em conjunto com as CFPE. As atividades meramente administrativas seriam realizadas pelos servidores, e a atividade jurisdicional e a aplicação de sanções seriam atribuídas aos Juízes Eleitorais.

Como foi constatado, não existe uma solução milagrosa, e esta passa pelo debate aprofundado e pela construção gradativa de conhecimento sobre o tema. A autorregulação de conteúdo político pelo Estado é um problema complexo e deve ser bastante debatida e detalhada. Como visto, as diferenças entre a regulação de conteúdo eleitoral e de conteúdo abusivo justificam a elaboração de uma norma específica para a regulação de conteúdo eleitoral, o que permitiria uma análise mais aprofundada e detalhada, evitando que lacunas sejam preenchidas a posteriori de forma assistemática. A elaboração inicial de uma norma sobre conteúdo abusivo, além dos benefícios que ela própria almeja, pavimentaria o caminho para a parceria entre Estado e plataformas, servindo de laboratório para a implantação e operação da regulação de conteúdo eleitoral.

Referências

ALTUNCU, E.; FRANQUEIRA, V. N. L.; LI, S. 2022. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. *Arxiv*, p. 1-31.

ALVIM, Frederico Franco et al. 2023. *Guerras Cognitivas na Arena Eleitoral: O Controle Judicial da Desinformação*. Rio de Janeiro-RJ. Lumen Juris.

BREWSTER, J.; ARVANITIS, L.; SADEGHI, M. 2023 The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation At Unprecedented Scale. *NewsGuard*.

CAMURÇA, Eulália Emília Pinho. 2021. *Judicialização das fake news na desordem do ecossistema informacional digital: devires do campo eleitoral*. Fortaleza. UFC.

CASTELLS, M., 2018. *O poder da identidade- A era da informação: economia, sociedade e cultura*, vol. 2. São Paulo-SP. Paz e Terra, 9ª Edição.

CHAVES, Mauro César Santiago. 2023. *Regulação Responsiva e Agências Reguladoras Federais: recorte jurídico-institucional sob a perspectiva da Advocacia-Geral da União e do Poder Judiciário Federal*. Brasília-DF. TCU.

CHESNEY, R.; CITRON, D. K. 2018. Deep Fakes: A Looming Challenge for A looming crisis for national security, democracy and privacy. *The Lawfare Blog*.

DEMENTIEVA, D.; KUIMOV, M.; PANCHENKO. 2023. A. Multiverse: Multilingual Evidence for Fake News Detection. *J. Imaging*, vol. 9, n.4, p. 77.

GRAÇA, Guilherme Mello. 2019. *Fake news e processo eleitoral: a cruzada quixotesca do tribunal superior eleitoral de combate às notícias falsas*. Niteroi-RJ. UFF.

HAMED, S.; AZIZ, M. J.; YAAKUB, M. R. 2023. A review of fake news detection approaches: a critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon*, v. 9, n. 1, p. 1-23.

HOES, Emma et al. 2023. Leveraging ChatGPT for Efficient Fact-Checking. *Psyarchiv Preprints*, p. 1-16.

JASPER, S. How we detect, remove and report. child sexual abuse material. 2022. *Google*.

LIM, C. 2018. Checking how fact-checkers check. *Research & Politics*, v. 5, n. 3, p. 205.

MARANHÃO, J.; CAMPOS, R.; GUEDES, J.; OLIVEIRA, S. R.; GRINGS, Maria Gabriela. 2021. Regulação de "Fake News" no Brasil. *Instituto Legal Grounds*.

MULLIN, J. 2022. Google's Scans of Private Photos Led to False Accusations of Child Abuse. *Electronic Frontier Foundation*.

OSÓRIO, Aline. 2017. *Direito eleitoral e liberdade de expressão*. Belo Horizonte-MG. Editora Fórum.

PENNYCOOK, G., BEAR; A.; COLLINS, E. T.; RAND, D. G. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, vol.66, n.11, p. 4944–4957.

PL 2.630-Câmara de Deputados (comp.). *PL Nº 2.630, 2023*. Disponível em: https://infoleg-autenticidade-assinatura.camara.leg.br/CD237493373700. 25 set. 2023.

POMPEU, L. 2023. Relator retira órgão regulador do PL das Fake News. *O Globo*.

RAIS, D.; FALCÃO, D.; GIACHETTA, A.Z. 2018. *Direito eleitoral digital*. São Paulo-SP. Revista dos Tribunais.

RAIS, D.; BATTISTI, R. 2022. O mito do PL de "fake news". *IV Simpósio de Direito Eleitoral do Nordeste*, vol. 1, n. 1, p. 50-63.

SANTOS, I. M. R. 2022. As formas de autorregulação na LGPD a partir da regulação responsiva. *Revista de Direito Setorial e Regulatório*, vol. 8, n. 1, p. 149-162.

TSE. *Fake News: TSE lança página para esclarecer eleitores*. Disponível em: https://www.tse.jus.br/comunicacao/noticias/2018/Outubro/fake-news-tse-lanca-pagina-para-esclarecer-eleitores-sobre-a-verdade. 25 maio 2023.

TWOREK, H.; LEERSSEN, P. 2019. An Analysis of Germany's NetzDG Law. *Amsterdam: Transatlantic Working Group*.

YI, J. 2022. ADD 2022: the first audio deep synthesis detection challenge. *ICASSP*, p. 9216-9220.