Please do not cite or quote without permission

Draft prepared for the SciencesPo Graduate Conference

Adjudicating Fake News[1]

Filippo Maria Lancieri[2]
Caio Mario da Silva Pereira Neto[3]
Rodrigo Moura Karolczak[4]
Barbara Marchiori de Assis[5]

Abstract:

A pressing public interest problem is the moderation of speech in the digital world—or, more specifically, how to develop frameworks that help societies tackle problems around disinformation and online harassment while at the same time protecting the freedom of speech values that are essential to democratic governance. This article presents preliminary results of a large empirical project on the adjudication of fake news disputes by Brazilian Courts during the 2018 Brazilian presidential elections—one of the first evaluations of the on-the-ground impacts of policy changes in this area. It examines what led Brazilian judges to order the takedown of online content, what social networks and types of content were the most affected by judicial decisions and whether there is evidence that incumbent politicians abused the system, among others. This research just help improve the design of legal interventions aimed at limiting the spread of disinformation.

I. Introduction

The digital revolution has placed the internet at the center of the political debate. Twitter, Facebook, WhatsApp, YouTube, and other networks quickly became crucial to empower the voice of previously excluded citizens and activists. Concomitantly, social media has generated new

[1] The empirical research developed for this paper was carried own by the Research Group on Competiton Policy and Regulation of Digital Platforms at FGV Direito SP between 2019 and 2021, with collaboration of the following members: Antonio Bloch Belizario; Daniel Favoretto Rocha; Esther Simon Seroussi Souccar; Fernanda Mascarenhas Marques; Georges Vicentini El Hajj Moussa; Giulia de Paola; Helena Secaf dos Santos; Pedro Marques Neto; Raíssa Leite de Freitas Paixão; and Vitória Oliveira.
NOTE: Both Filippo Lancieri and Caio Pereira Neto have previously acted as external counsel to digital platforms (including Google, Facebook, Uber, among others). Lancieri's involvement ceased in 2015; Pereira Neto is still retained by these companies. This article, however, reflects solely our own independent judgment and it did not receive funding or was subject to any form of influence from outside sources. All errors are, of course, our own.

[2] Post-Doctoral Fellow, ETH Zurich Center for Law & Economics; Fellow, The George J. Stigler Center for the study of the Economy and the State, UChicago Booth; Contact: flancieri@ethz.ch
[3] Professor of Law: FGV Direito SP, São Paulo. LLM (2002) JSD (2005) Yale Law School. Contact: caio.pereira@fgv.br
[4] PhD candidate in Political Sciences, University of Illinois at Chicago. Masters in Political Sciences, New York University. Contact: rmoura2@uic.edu
[5] Research Fellow at the Competition, Public Policy, Innovation, and Technology (COMPPIT) Nucleus at FGV Direito SP. Masters in Public Administration (2014), Cornell University. Contact: bm524@cornell.edu

challenges to democracies across the globe. Major electoral upsets around the world quickly brought "fake news"[6] to the leading ranks of threats to democracy[7].

Digital platforms and countries worldwide are struggling to understand the dynamics and impacts of fake news on the electorate and to devise appropriate policy responses. They face a fundamental challenge: how to ensure freedom of expression and the spread of ideas—a core value in liberal democracies – while preventing disinformation from contaminating the public sphere? Responses vary immensely. Some are strict: Germany's hate speech and false news law attributes some balancing responsibility to online platforms and imposes fines up to EUR 50 million[8]; in the UK, the Ofcom will oversee a new duty of care imposed on digital platforms, one with the power to fine companies or imprison directors[9]; Singapore's law grants the Government authority to order the removal of statements of fact deemed false or misleading, allowing for fines or imprisonment in cases of violation[10]. The EU and the US, which have been largely relying on self-regulation, are considering coupling self-regulation mechanisms with mandatory obligations on digital platforms to increase transparency and curb the spread of false information; Czech Republic is experimenting with Governmental fact-checking[11]. Meanwhile, digital platforms are under a barrage of criticism

[6] We collapse fake news, false news, disinformation and misinformation campaigns as similar terms: fabricated stories presented as if from legitimate sources. This meaning includes both information that "*mimics traditional media content in form but not in organizational process and intent*" (fake news), "*information that is overtly false and misleading*" (misinformation) and "*information that is purposely spread to deceive people*" (disinformation). All types can and are used to influence the electoral process, from fake news websites to fake images, videos or memes. For the definitions, see David MJ Lazer and others, 'The Science of Fake News' (2018) 359 Science 1094. Pg. 1094. In doing so, we follow Soroush Vosoughi, Deb Roy and Sinan Aral, 'The Spread of True and False News Online' (2018) 359 Science 1146. Pg 1146, though they refrained from using the term "fake news" due to its political content.

[7] See Nathaniel Persily, 'The 2016 US Election: Can Democracy Survive the Internet?' (2017) 28 Journal of democracy 63.

[8] Anthony Faiola and Stephanie Kirchner, 'How Do You Stop Fake News? In Germany, with a Law.' *Washington Post* (5 April 2017) <https://www.washingtonpost.com/world/europe/how-do-you-stop-fake-news-in-germany-with-a-law/2017/04/05/e6834ad6-1a08-11e7-bcc2-7d1a0973e7b2_story.html> accessed 1 July 2018.

[9] See UK Home Office, 'Online Harms White Paper - Executive Summary' (*GOV.UK*, April 2019) <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2> accessed 26 April 2019.

[10] See Singapore's Singapore's Parliament, Protection from Online Falsehoods and Manipulation Bill 2019 [10/2019].

[11] Anthony Faiola, 'As Cold War Turns to Information War, a New Fake News Police Combats Disinformation' *Washington Post* (22 January 2017) <https://www.washingtonpost.com/world/europe/as-cold-war-turns-to-information-war-a-new-fake-news-police/2017/01/18/9bf49ff6-d80e-11e6-a0e6-d502d6751bc8_story.html> accessed 1 July 2018.

for either over or underenforcing their community guidelines and terms of service[12] – so much that some of them started advocating for governmental regulation delimitating what they should do[13].

As countries, companies, and societies scramble to find a solution to fight disinformation through new legal or extra-legal requirements, some questions remain mostly undiscussed or lack an empirical analysis: under what standards should governments intervene? Which authority would be responsible for enforcing these new regulatory standards? and how to design standards that are actually effective? The literature review on disinformation campaigns stressed the urgency for more research on "*the effects of new laws and regulations intended to limit the spread of disinformation*".[14]

This article hopes to shed light on this question by studying the performance of the Brazilian judiciary in this field, especially concerning disinformation during the electoral process. While still a work-in-progress, initial results indicate that the Brazilian model may be considered a step forward to democratic accountability (in comparison with pure moderation of the platforms). However, it faces important pitfalls and limitations that academics and policymakers must face. Understanding the current institutional framework, its operation, and its current limitations is relevant to inform future reforms and systems under implementation in other jurisdictions.

This piece starts by providing an overview of policy reactions to the disinformation phenomenon (Section II), then it explains Brazilian judiciary's role in tackling disinformation campaigns during elections (Section III), presents the theoretical virtues and flaws of the Brazilian system (Section IV), our research questions and database (Section V), presents an overview of initial results (Section VI) and discusses those results in a brief conclusion (Section VII). As you

[12] Will Oremus, 'Why Twitter Isn't Banning Alex Jones, According to a Top Executive' [2018] *Slate Magazine* <https://slate.com/technology/2018/07/twitters-vijaya-gadde-on-its-approach-to-free-speech-and-why-it-hasnt-banned-alex-jones.html> accessed 18 October 2018.

[13] See 'Opinion | Mark Zuckerberg: The Internet Needs New Rules. Let's Start in These Four Areas.' (*Washington Post*) <https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html> accessed 2 April 2019; Shannon Liao, 'Tim Cook Says Tech Needs to Be Regulated or It Could Cause "Great Damage to Society"' (*The Verge*, 23 April 2019) <https://www.theverge.com/2019/4/23/18512838/tim-cook-tech-regulation-society-damage-apple-ceo> accessed 24 April 2019. and Kent Walker, 'Smart Regulation for Combating Illegal Content' (*Google Public Policy Blog*, 14 February 2019) <https://www.blog.google/perspectives/kent-walker-perspectives/principles-evolving-technology-policy-2019/smart-regulation-combating-illegal-content/> accessed 6 June 2019.

[14] See Joshua Tucker and others, 'Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature' <https://www.hewlett.org/wp-content/uploads/2018/03/Social-Media-Political-Polarization-and-Political-Disinformation-Literature-Review.pdf> accessed 24 May 2019., p. 7.

will see, this is very much a work-in-progress, so questions and comments are particularly welcome.

II. Fighting disinformation: from theory to practice

a. In the books

Other than platform self-regulation, the threat posed by misinformation has led to several legal responses by Governments around the world. Policy reactions can be grouped into two main categories of intervention: (i) individual empowerment aiming to enable users to discern between real and false information (e.g., fact-checking or educational campaigns); and (ii) structural changes (i.e., changes in the legal framework) aimed at preventing the spread of the fake content before exposure (e.g., bans, restrictions, imposition of liability).[15]

Both face limitations. Individual empowerment initiatives are, at best, long-term solutions. These initiatives are softer and may take the form of digital literacy campaigns or more organized fact-checking environments. A growing body of research shows that citizens judge poorly whether news are real. Media literacy campaigns exist for years, and while evidence shows that they help, they are no panacea.[16] Moreover, research suggests that the general effects of fact-checking are ambiguous.[17] Studies indicate that "social fact-checking" (done by one's social circle) are ineffective in debunking myths.[18] Specific data on online news indicates that professional fact-checking has a small effect in making people distrust information labeled as false.[19] However, labeling may also lead to the "implied truth effect."[20] As fact-checking is restricted to a limited number of cases, readers may infer that false information is true simply because no one has told

[15] Lazer and others (n 6). Pg. 1095.
[16] Monica Bulger and Patrick Davison, 'The Promises, Challenges and Futures of Media Literacy'. pg. 18, reviewing a large literature on the pros and cons of media literacy campaign and affirming that "*from an evidence perspective, there remains uncertainty around whether media literacy can be successful in preparing citizens to resist 'fake news' and disinformation*".
[17] For a review see Gordon Pennycook and others, 'The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Stories Increases Perceived Accuracy of Stories without Warnings' <Available at SSRN: https://ssrn.com/abstract=3035384 or http://dx.doi.org/10.2139/ssrn.3035384>., pg 2-3.
[18] Tucker and others (n 14). pg. 19.
[19] ibid. pg. 53.
[20] Pennycook and others (n 17). pg. 9; 17. See also Adrien Friggeri and others, 'Rumor Cascades.', *ICWSM* (2014). Pg 8, about how having a rumor referred to Snopes.com (a fact-checking website) did not impact the relative distribution of real and false photos.

them otherwise.[21] If this last account is true, fact-checking may even make matters worse in some circumstances.

Structural changes (i.e., changes in the legal framework), while preferred by some,[22] also have limitations. On the one hand, granting state agencies strong power to evaluate and ban fake news can pose severe restrictions on freedom of speech and may be considered unconstitutional in many liberal democracies. On the other, imposing additional liability on platforms through national laws and regulations risks either turning them into censors of acceptable content or be ineffective, depending on how these platforms weigh over-enforcement (type I) or under-enforcement (type II) errors. Finally, platforms' monetization strategies based on user engagement may diminish incentives to self-regulate, as misinformation can drive up engagement (at least in the short-term).[23]

Any changes in the legal framework should consider some aspects specific to disinformation. The definition of disinformation should be carefully crafted to avoid impacting lawful and non-harmful speech. Determining if the information is false or misleading requires some contextualization. It is also crucial to establish whether the author had an intent to spread misleading or false information. This is particularly important as disinformation has the potential to cause diffuse social damage. Given disinformation specificities, it may unwise to simply group it with other types of illegal or problematic online content such as online harassment and other forms of problematic online speech—disinformation deserves a framework of its own.[24]

Additionally, structural changes that attempt to impose or stimulate the adoption of controls based on Artificial Intelligence (AI) and automation strategies still face major challenges. The science behind AI monitoring of content leads to some difficulties: (i) the subtlety of language and context, as well as the increased use of images, creates challenges in the development of 100% automated AI monitoring tools, as the first wave of AI development was based on clear cut rules and correlations amongst pieces of data that had a hard time distinguishing misinformation from irony (for example); while (ii) the sheer volume of content makes purely human control

[21] Pennycook and others (n 17). pg. 11. This can be due to many reasons, either because fact-checkers did not have access to the news, or did not have resources to check it or because they could not ascertain the content's veracity.
[22] Xiaoyan Qiu and others, 'Limited Individual Attention and Online Virality of Low-Quality Information' (2017) 1 Nature Human Behaviour 0132. pg. 13.
[23] Persily (n 7). Pg 74. Tucker and others (n 14). pg. 37-38.
[24] Pielemeier, J. (2020). Disentangling disinformation: What makes regulating disinformation so difficult?. Utah Law Review, 2020(4), 917-940.

economically and physically difficult, even with thousands of content moderators.[25] Thus, structural solutions leveraging these tools fuel an (ongoing) arms race between platforms and content creators, whose winners and losers depend on the weighing of under or over-enforcement errors and how much societies are willing to accept in terms of platform censorship.[26]

The quest, therefore, has turned to finding a middle-ground that can promote structural changes while protecting online freedom of speech. Prof. Jack Balkin has a particularly useful framework in this area, affirming that online speech governance should be seen as a free-speech triangle between companies, users and the Government—one where decisions are necessarily intertwined. Balkin envisages different legal approaches when regulating digital platforms, such as the importance of balancing intermediary liability (which gives platforms incentives to police speech) with immunity (which protects some of their moderation decisions).[27]

The main risk of this interaction is that Governments successfully co-op companies as censor of users' legit manifestations, through the employment of some form of autocratic, non-transparent governance.[28] For Balkin, a possible solution is to impose on platforms an obligation to grant users due process rights and require that they operate as information fiduciaries who have duties of good faith and non-manipulation towards their users.[29] This would potentially enable platform governance, prevent governmental abuse, and protect users from platform abuse.

While proposals like this indicate a potential path forward, they are not tailored to addressing the specific problems of disinformation, limiting their direct applicability to the case at hand. It seems particularly difficult to use this due process framework inside platforms to deal with

[25] On the challenges of using AI to curated online content, see Tucker and others (n 14). pg. 38-39. A good example of the shortcoming of these methods is how, even with AI tools and thousands of new moderators, Facebook was incapable of limiting the live stream of the Christchurch shootings in New Zealand, leading the company to rethink how to monitor live videos. See Hamza Shaban, 'Facebook to Reexamine How Livestream Videos Are Flagged after Christchurch Shooting' (*Washington Post*, 21 March 2019) <https://www.washingtonpost.com/technology/2019/03/21/facebook-reexamine-how-recently-live-videos-are-flagged-after-christchurch-shooting/> accessed 25 April 2019.

[26] See Lazer and others (n 6). Pg 1096.; Issie Lapowsky, 'Fake Facebook Accounts Are Getting Harder to Trace' [2018] *Wired* <https://www.wired.com/story/facebook-uncovers-new-fake-accounts-ahead-of-midterm-elections/>. and Mark Scott, 'Why We're Losing the Battle against Fake News' (*POLITICO*, 7 October 2018) <https://www.politico.eu/article/fake-news-regulation-misinformation-europe-us-elections-midterms-bavaria/>.

[27] Jack M Balkin, 'How to Regulate (and Not Regulate) Social Media' [2020] Knight Institute Occasional Paper Series, at 25.

[28] Jack M Balkin, 'Free Speech Is a Triangle' (2018) 118 Columbia Law Review 2011. at 11; 15. See also Jack M Balkin, 'Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation' (2017) 51 UCDL Rev. 1149.

[29] Balkin, 'Free Speech Is a Triangle' (n 27). at 16.

disinformation in elections, considering the (i) political nature of the speech and the theoretical high-value associated with it, (ii) the short campaign periods, which also require fast interventions, and (iii) the intensity/amount of information flow in these periods. Platforms decisions usually end up benefiting one side of the political spectrum, putting them between a rock and hard place that risk displeasuring all those involved: something that could be seen in the aftermath of the 2020 US Presidential elections, where Democracts called for more content content moderation while Republicans accused platforms of censorship and suppression.[30] For any middle-ground solution to work, it is critical to develop trustworthy and responsive institutions that can oversee these obligations.

b. On the ground

The challenge in devising an optimal policy quickly led to diversity amongst jurisdictions. By comparing the different approaches to tackle the problem that are in place in the US, the EU, Czech Republic, France, the UK, Germany and Singapore, one can almost develop a continuum of policies ranging from self-regulation (US/EU), to direct governmental oversight (Singapore), as summarized below.

Historically, **the US** has mostly relied on a system of self-regulation, delegating to the digital platforms all responsibility in shaping their digital environments[31]—even if platforms themselves are increasingly asking for clearer Federal standards.[32] Actual policy responses focused on media-literacy platforms and campaigns, with some states across the US passing or at least discussing laws that encourage these programs.[33] However, it is noteworthy that states have started discussing legislation targeting online platforms to curb misinformation spread. For instance, California has introduced a bill that requires a person that operates a social media platform to disclose whether that platform has mechanisms or a policy in place to tackle the spread of

[30] 'New Poll: 75% Don't Trust Social Media to Make Fair Content Moderation Decisions, 60% Want More Control over Posts They See' (*Cato Institute*, 15 December 2021) <https://www.cato.org/blog/new-poll-75-dont-trust-social-media-make-fair-content-moderation-decisions-60-want-more>.
[31] See, for example, the White Paper proposed by Senator Mark Warner with initiatives to regulate social media and tech companies. Mark Warner, 'Potential Policy Proposals for Regulation of Social Media and Technology Firms' <https://slate.com/technology/2018/10/mark-warner-big-tech-if-then-transcript.html>.
[32] See 'Opinion | Mark Zuckerberg: The Internet Needs New Rules. Let's Start in These Four Areas.' (n 13).
[33] Bulger and Davison (n 16). pg 6. For an updated list see MLN, 'Your State Legislation' (*Media Literacy Now*) <https://medialiteracynow.org/your-state-legislation/> accessed 24 October 2018.

misinformation.[34] Texas and Florida have passed laws aimed at preventing "online censorship", though these are under litigation for potential violations of the First Amendment.[35]

The Federal Congress is yet to follow suit—American freedom of speech culture and legislation has historically been treated as the cornerstone of Internet freedom and relied on more information as a way to combat misinformation[36]—yet a shift appears to be underway. In 2018, Congress passed the FOSTA-SESTA Act to address concerns that online platforms were facilitating sex-trafficking (though its on the ground impact is far from clear);[37] and there are multiple bills under discussion that would reform the the immnuities granted by Section 230 of the Communications Decency Act.[38] While these trends are starting to reach disinformation,[39] concrete actions specifically targeting disinformation are limited.[40] Aside from COVID-19-related misinformation, the Federal Bureau of Investigation (FBI) and the Cybersecurity and Infrastructure Agency (CISA) issued announcements concerning potential threat posed by

[34] California Legislative Information, AB-35 Social Media Platforms: false information. <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB35> accessed on 27 April 2021.

[35] Benjamin Din, 'Federal Judge Blocks Florida's Social Media Law' (*POLITICO*) <https://www.politico.com/news/2021/06/30/judge-block-florida-social-media-law-497442>; James Pollard, 'Federal Judge Blocks Texas Law That Would Stop Social Media Firms from Banning Users for a "Viewpoint"' (*The Texas Tribune*, 2 December 2021) <https://www.texastribune.org/2021/12/01/texas-social-media-law-blocked/>.

[36] While the US first-amendment is complex, a common theme has been that chilling effects on speech are problematic and that government limitations on speech due to its content must normally survive strict scrutiny or prove a compelling governmental interest. See Geoffrey R Stone, 'Free Speech in the Twenty-First Century: Ten Lessons from the Twentieth Century' (2008) 36 Pepp. L. Rev. 273. Pg. 277; 282. For example, the Supreme Court affirmed that libel suits against news publishers require actual malice, or the statement has to be clearly false or the publisher has to act in reckless disregard of its truth or falsity. *New York Times Co v Sullivan* (1964) 376 254.

[37] 'Sex Trafficking - Online Platforms and Federal Prosecutions' (United States Government Accountability Office, June 2021) <https://www.gao.gov/assets/gao-21-385.pdf>.

[38] Daisuke Wakabayashi, 'Legal Shield for Social Media is Targeted by Lawmakers' *The New York Times* (15 December 2020), https://www.nytimes.com/2020/05/28/business/section-230-internet-speech.html; Section 230 Reform Legislative Tracker <https://slate.com/technology/2021/03/section-230-reform-legislative-tracker.html>

[39] Some Bills proposed in the 117th Congress to alter Section 230 include: Protecting Constitutional Rights From Online Platform Censorship Act; Curbing Abuse and Saving Expression in Technology (CASE-IT) Act; See Something, Say Something Online Act of 2021; Abandoning Online Censorship (AOC) Act; and Platform Accountability and Consumer Transparency (PACT) Act.

[40] One example, that yielded no results was the Trump Administration Executive Order 13925 from May 2020, which instructed federal departments and agencies to clarify the scope of Section 230 immunity for digital platforms and propose regulations to clarify this matter. https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-preventing-online-censorship/.

attempts from cybercriminals and foreign actors to spread disinformation about 2020 election results and cyberattacks on US voting systems.[41] But not much more has been done

The European Union initially relied on self-regulation. The European E-Commerce Directive: (i) establishes a safe-harbor shielding platforms from liability for third-party content whenever platforms are 'mere conduits' of information or whenever the platform is hosting information that it does not know it is illegal; (ii) prevents the imposition of general monitoring obligations on platforms;[42] but (iii) allows Governments to establish procedures for the removal of information, to impose obligations on platforms to remove illegal content and requires Courts to issue interim orders to cease illegal action by information society services.[43] Yet, European case law has left the door open for EU countries to determine their own approach, in some cases affirming the liability of online platforms regarding third-party content and in others affirming the safe harbor.

In 2018, an official policy report produced by a high-level group of 39 experts[44] for the European Commission defended an EU-wide multi-dimensional approach to fight misinformation based on five pillars: (i) transparency of online news;[45] (ii) media literacy programs;[46] (iii) empowering users and journalists to tackle misinformation;[47] (iv) safeguarding diversity of the

[41] FBI; CISA (September 2020). 'Foreign Actors and Cybercriminals Likely to Spread Disinformation Regarding 2020 Election Results' < https://www.ic3.gov/Media/Y2020/PSA200922>, and 'False Claims of Hacked Voter Information Likely Intended to Cast Doubt on Legitimacy of US Elections,' accessed 27 April 2021.
[42] Articles 12, 14 and 15 of Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market 2000 (OJ L 178).
[43] Articles 12, 14, 15 and 18 of ibid.
[44] See also two other official EU Documents, European Commission, 'Communication: Tackling Online Disinformation: A European Approach' <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236>. that largely defend similar policies though with some more details on online accountability, and European Parliament, 'European Parliament Resolution of 15 June 2017 on Online Platforms and the Digital Single Market (2016/2276(INI))' <https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2016/2276(INI)>., spurring the European Commission to take action to better regulate online platforms.
[45] Including understanding of who owns online websites, what type of content is sponsored, payments to human influencers and the use of robots, encouragement of credible journalism sources, dilution of disinformation with quality information (i.e. demoting), transparency on what drives platform algorithms, increased fact-checking and the promotion of a market for fact-checking and increased access to data controlled by platforms. High Level Group on fake news and online disinformation, 'A Multi-Dimensional Approach to Disinformation' <https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation>. pg. 22-25.
[46] Expanding into data literacy, or how data is used. ibid. pg. 26.
[47] Such as the development of source transparency indicators that ranks good providers, the promotion of client-based interfaces that empowers users on how to access information, the development of tools for online checking of

European news media ecosystem;[48] and (v) continuing research on the impacts of disinformation as a way to ensure that responses remain up-to-date.[49] This report resulted in the Code of Practice on Disinformation, [50] where platforms mostly listed what initiatives they were adopting to combat fake news, a conclusion deemed deeply unsatisfactory by the high-level expert panel.[51] In May 2020, an independent study concluded that, although the Code of Practice was an important step to tackle disinformation, its self-regulatory nature prevented platforms from being held accountable for not complying with the Code.[52]

The rise of "fake news" and the impact of COVID-19 on the spread of misinformation spurred Europeans into rethinking the EU-wide system by increasing the role of mandatory regulation. In addition to strengthening the Code of Practice on Disinformation, the European Commission, building on the E-Commerce Directive, has proposed the Digital Services Act and the Digital Markets Act, encompassing a new and single regulatory framework applicable across EU countries.[53]

Notably, the Digital Services Act establishes rules according to online intermediaries' functions and sizes, seeking to improve the removal of illegal content, increasing the oversight of online platforms that reach more than 10% of EU's population. Given the risk of disseminating false information, large platforms will need to adopt risk management measures that include independent audits. To facilitate supervision and research into potential disinformation risks, large

content such as audiovisual and text-based reports, training of journalists on how to navigate the online world, and media innovation projects. ibid. pg. 27-29.

[48] Such as the protection of press freedom and no intervention in editorial independence, providing funding to support quality journalism (such as VAT exemptions and direct state aid) and funding of media innovation. ibid. pgs. 29-30.

[49] ibid. pg. 31.

[50] European Commission, 'EU Code of Practice on Misinformation' <https://ec.europa.eu/digital-single-market/en/news/code-practice-disinformation>.

[51] See an opinion issued by an EU institution, the Sounding Board, on the promises made by the digital platforms: "As outlined by the Sounding Board's previous written feedback and comments, the "Code of practice" as presented by the working group contains no common approach, no clear and meaningful commitments, no measurable objectives or KPIs, hence no possibility to monitor process, and no compliance or enforcement tool: it is by no means self-regulation, and therefore the Platforms, despite their efforts, have not delivered a Code of Practice. "European Commission, 'EU Code of Practice on Misinformation - Annex III - Opinion of the Sounding Board' <https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=54456> accessed 18 October 2018.

[52] European Commission, 'Study for the assessment of the implementation of the Code of Practice on Disinformation'<https://digital-strategy.ec.europa.eu/en/library/study-assessment-implementation-code-practice-disinformation>

[53] The Digital Services Act focuses on regulating online intermediaries and the Digital Market Act aims to regulate digital platforms that operate as gatekeepers between businesses and consumers for digital services.

platforms will be required to ensure access to platforms' key data repositories for authorities and researchers.

The **Czech Republic** created a new Center Against Terrorism and Hybrid Threats within the Ministry of the Interior. One of the Center's main goals was to prevent the spread of misinformation campaigns, leading some to call it a "Truth Ministry."[54] The country does not have specific legislation addressing disinformation, mostly relying on Czech criminal law.[55] The Center had a more active role during the COVID-19 pandemic by identifying the most common disinformation narratives in the country and partnering with other organizations.[56]

In **France**, a law against information manipulation was signed in December 2018.[57] Particularly concerned over potential influence on election outcomes, digital platforms must publish the auhtor's name and the amount paid of sponsored content during campaign periods. Similar to Brazil, there is a legal injunction system that enables candidates, political parties, public prosecutors, or any person to file a motion to remove false information online during the three months preceding the election's first day and until the election's end. Interim judges may order to remove online content if it is inaccurate or misleading, artificially or automated disseminated on a massive scale, and could potentially disturb the peace or compromise election results. The court must rule within 48 hours. Between elections, the French Broadcasting Authority verifies platforms' efforts on combating the spread of false information. Such efforts include the implementation of mechanisms to easily flag false information and combat accounts that massively distribute fake information, as well as improving algorithm transparency and raising awareness among users.[58] The law will effectively come into force in the 2022 elections.

The UK Government is proposing a new legal and regulatory framework establishing digital plataforms' duty of care towards their users. After the launch of the Online Harms White

[54] See Faiola (n 11).
[55] Center Against Terrorism and Hybrid Threats, 'Criminal Law Regulation,' <https://www.mvcr.cz/cthh/clanek/dezinformacni-kampane-trestnepravni-uprava-trestnepravni-uprava.aspx> accessed on 27 April 2021.
[56] Center Against Terrorism and Hybrid Threats, 'Coronavirus: An overview of the Main Disinformation Narratives in the Czech Republic,' <https://www.mvcr.cz/cthh/clanek/coronavirus-an-overview-of-the-main-disinformation-narratives-in-the-czech-republic.aspx> accessed on 27 April 2021.
[57] Loi n° 2018-1202 du 22 décembre 2018 relative à la lutte contre la manipulation de l'information (Dec. 22, 2018), <https://perma.cc/QH5N-25MC> accessed on 27 April 2021.
[58] Nicolas Boring, 'Government Response to Disinformation on Social Media Platforms: France' (*Library of Congress*, September 2019) < https://www.loc.gov/law/help/social-media-disinformation/france.php> accessed on 27 April 2021.

Paper consultation in April 2019,[59] the UK Government published its full government response in December 2020[60] and introduced a strict Online Safety Bill in 2021. This new framework will apply to those platforms that host user-generated content with access in the UK and/or facilitate public and private online interaction between service users, one of whom is in the UK. In addition to tackling illegal online content and activity (e.g., hate crime, drug-related offenses, fraud and financial crime, terrorism, child sexual abuse), this Bill would also address "legal but harmful content and activity."

However, in December 2021, the House of the Lords and House of the Commons Joint Committee on the Draft Online Safefty Bill issued a 192-page report recommending that the controversial "legal but harmful content" is removed. Instead, the report recommended taking into account existing criminal legislation about offline harmful behavior that could also be applicable online, avoiding the definition of "legal but harmful content" that could be inadequately broad and negatively impact freedom of expression.[61]

Notably, platforms would need to proactively remove illegal content and activity –that is, even without users' reporting– and ensure that users and children are not exposed to such content. To this end, platforms' algorithms and other functionalities would have to automatically detect illegal content and remove them to minimize users' exposure to such material.[62] False information i*ntentionally* sent to cause harm would be illegal (e.g., hoax bomb threats), in contrast to misinformation cases, in which lacks users' intentionality to spread false information. In this context, the report provided some examples of disinformation activities that could be "recognized as letigimate grounds for interference in freedom of expression," such as: knowingly false information that could substantially harm –physically or psychologically– a reasonable person; disinformation that could put public health at risk; and disinformaton that could negatively impact electoral systems' integrity.

---

[59] See HM Government, 'Online Harms White Paper' <https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper-executive-summary--2>.
[60] See HM Government, 'Online Harms White Paper: Full Government Response to the consultation'< https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/944310/Online_Harms_White_Paper_Full_Government_Response_to_the_consultation_CP_354_CCS001_CCS1220695430-001__V2.pdf>
[61] "House of Lords; House of Commons; 'Draft Online Safety Bill – Report of Sessions 2021-22' (December 2021). < https://publications.parliament.uk/pa/jt5802/jtselect/jtonlinesafety/129/129.pdf>.
[62] Gov.UK, 'Online safety law to be strengthened to stamp out illegal content'< https://www.gov.uk/government/news/online-safety-law-to-be-strengthened-to-stamp-out-illegal-content>.

Named as the regulator the Office of Communications (Ofcom) would follow a proportionate and risk-based approach, including non-financial sanctions, fines, and business disruption measures. Specifically, enforcement will range from notices of non-compliance to fines of up to 10% of a company's annual global turnover. Ofcom will also be able to block access to services in the UK as a last resort measure.

Germany passed a strict hate speech law (NetzDG) that mandates the removal of manifestly illegal content in less than 24 hours after notification. Digital platforms that fail to comply may be fined up to 50 million Euros per violation. The definition of what may be manifestly illegal content is broad and includes almost 24 definitions of the German criminal code, including some associated with insult, defamation, privacy protections, forgery and others.[63] This system mostly places the burden on users and online platforms to police hate speech and, collaterally, misinformation. Users or governmental entities are required to issue takedown orders and platforms need to respond swiftly – mostly within 24 hours.

Many commentators have argued that this process could lead to online censorship as platforms become more risk-averse and remove most content under the risk of being fined.[64] Reports by Facebook indicate that the company received 4,211 NetzDG reports between July/December 2020, and 1,117 of them led to a total of 1,276 posts blocked or deleted.[65] On other platforms, the number of purported illegal content under the NetzDG has been substantially higher. YouTube received 323,792 reports and removed 73,477 during the same period.[66] Similarly, Twitter received 811,469 reports and removed 118,797.[67] One possible explanation for the significant gap between Facebook's data and those published by YouTube and Twitter is the

[63] For a summary of the law see Center for Democracy & Technology, 'Overview of the NetzDG Network Enforcement Law' (*CDT Insights*, 17 July 2017) <https://cdt.org/insight/overview-of-the-netzdg-network-enforcement-law/> accessed 30 October 2018. The categories of criminal defamation, privacy violations, forgery and others are broad in many EU nations, including Germany. See Article 19, 'Responding to "Hate Speech": Comparative Overview of Six EU Countries' 19 <https://www.article19.org/resources/united-kingdom-responding-to-hate-speech/>. Pg 21. Indeed, reports by Facebook indicate that the bulk of removal requests involve insults and defamation requests. See Facebook, 'NetzDG Transparency Report' <https://fbnewsroomus.files.wordpress.com/2019/01/facebook_netzdg_january_2019_english71.pdf>.
[64] See Balkin, 'Free Speech Is a Triangle' (n 27). pg. 21.
[65] Facebook (n 60).
[66] Google Transparency Report, 'Removals under the Network Enforcement Law,' < https://transparencyreport.google.com/netzdg/youtube?hl=en&items_by_submitter=period:Y2020H2&lu=items_by_submitter> accessed on 27 April 2021.
[67] Twitter, Transparency – Germany, <https://transparency.twitter.com/en/reports/countries/de.html> accessed on 27 April 2021.

different complaint mechanisms implemented by these platforms.[68] Despite these takedown measures, it is unclear whether this legislation significantly contributed to preventing hate speech and amendments are under discussion. A very interesting study comparing the prevalence of hate speech by extreme-right Twitter in Germany before and after the passage of the law found that the regulation reduced the intensity of hateful speech in Germany by about 2%, and also had spill-over effects in Austria.[69]

Finally, in October 2019, **Singapore**'s Protection from Online Falsehoods and Manipulation Act (POFMA) took effect. POFMA granted the Government authority to order the removal of statements of fact deemed false or misleading, allowing for fines or imprisonment in cases of violation.[70] This complex and all-encompassing law authorizes the Singaporean government to order the removal of content it deems false or misleading, block the access to websites and require platforms and ISPs to prevent future access and dissemination of information by that source and authorize the publication of a "correction notice" that reflects the real facts, among others. Penalties vary depending on the types of violation, but normally include hefty fines (up to USD 750,000) and imprisonment terms (up to 10 years) for non-compliance.[71]

The law provides ample power for a Government authority to require previous monitoring from platforms and direct the limits on acceptable speech—powers condemned by a Special Rapporteur of the United Nations High Commissioner for Human Rights as incompatible with international Human Rights Law.[72] Noteworthy, the number of "correction notices" issued by Ministers under the framework of POFMA increased during the election campaign in July 2020. Many of those notices were issued against opposition politicians and political parties.[73]

---

[68] Heldt, A. (2019). Reading between the lines and the numbers: an analysis of the first NetzDG reports. Internet Policy Review, 8(2). DOI: 10.14763/2019.2.1398, p. 11.
[69] Raphaela Andres and Olga Slivko, 'Combating Online Hate Speech: The Impact of Legislation on Twitter' [2021] ZEW-Centre for European Economic Research Discussion Paper. at 2.
[70] See Singapore's Protection from Online Falsehoods and Manipulation Act (POFMA)
[71] Singapore's Parliament Protection from Online Falsehoods and Manipulation Bill (n 10). See items 27, 7(3)(c).
[72] David Kaye, 'Statement by the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression on the Protection from Online Falsehoods and Manipulation Bill' <https://www.ohchr.org/Documents/Issues/Opinion/Legislation/OL_SGP_3_2019.pdf>. pg. 5.
[73] Human Rights Watch (2021), 'World Report 2021 – Events of 2020,' <https://www.hrw.org/sites/default/files/media_2021/01/2021_hrw_world_report.pdf> p. 589, and Aqil Haziq Mahmud (October 2020). 'In Focus: Has POFMA been effective? A look at the fake news law, 1 year since it kicked in.' In *Channel New Asia* < https://www.channelnewsasia.com/news/singapore/singapore-pofma-fake-news-law-1-year-kicked-in-13163404> accessed on 27 April 2021.

In short, after failed attempts to push self-regulation into digital platforms to avoid the proliferation of misinformation and disinformation, many countries have been shifting to mandatory regulatory frameworks while delegating the enforcement to government authorities. Many proposed frameworks put digital platforms under an obligation to be more transparent about their algorithms, facilitate user reporting of false information, and enable access to their data. Still, many governments' frameworks encourage some degree of self-regulation (e.g., codes of good practice).

III. Elections and the role of the Brazilian judiciary in tackling disinformation

Brazil hosts a complex and mature election system. Elections take place every even year, always in October/November. Every four years (e.g. 2014, 2018, 2022), the country elects a new President, a new federal House of Representatives and a portion of the Federal Senate,[74] new state Governors, new state Congresses. Two years after Federal and State elections (e.g. 2020, 2024), the country votes on mayors and municipal legislatures. Elections for President, Governors and Mayors have run-offs with the two most voted candidates if no candidate reaches at least 50% of the valid votes in the first run.

With a population of approximately 210 million people, Brazil is the world's fourth-largest democracy and a highly connected nation. According to the most recent data, 130 million Brazilians are active social network users, and the country is among the largest markets for Facebook (127 million monthly active users), WhatsApp (120 million), Instagram, YouTube and Twitter. Moreover, Brazilians trail citizens from developed countries in educational attainments, potentially making them more susceptible to disinformation. As a result, Brazilian authorities identified "fake news" as a potential threat to the high-stakes 2018 presidential elections—the most polarized election to date—and adjusted the election monitoring structures to deal with it.

Brazil hosts a rare and complex electoral court system tasked with the organization of fair and safe elections. During election periods, normally three-months before and after an election, an army of around 3,000 State and Federal Judges and 3,000 State and Federal prosecutors is temporarily transferred to electoral courts to oversee the fairness of the poll. Electoral courts are

[74] Each state is represented by 3 Senators, which serve an 8-year term. Senatorial elections are organized every four years, together with the presidential election, renewing 1/3 or 2/3 of the Senate every other election.

divided into three branches: around 2,500 electoral district courts; 27 Electoral Courts of Appeals (*Tribunais Regionais Eleitorais,* or TREs), one for each of the 26 Brazilian States and another for the Federal District around the Brazilian Capital; and the Superior Electoral Tribunal (*Tribunal Superior Eleitoral*, or TSE), Brazil's highest electoral court. Career judges and prosecutors are appointed to each of the electoral districts[75]. The TSE is composed of seven justices[76]. The TREs follow a similar composition, reflecting lower court appointments[77]. Both the TSE and TREs also appoint auxiliary justices who are responsible for specific matters, whose decisions are appealable to the full-court.

Brazilian Courts have been trying to address problems associated with online disinformation and irregular electoral advertisement for years. Since 2009, the Brazilian electoral law has specific provisions: (i) affirming that online platforms must abide byelectoral courts' decisions; and (ii) establishing a safe harbor that exempts platforms from direct liability associated with the illegal content as long as companies were unaware of the content's illegal nature. As amended in 2017, the law grants the TSE powers to establish "good practices" regarding political advertisement on online platforms. More controversial, in 2017, the TSE expanded on such powers to establish, through a formal rulemaking process, restrictions on disseminating "*facts known to be untrue*", the standard for taking down false online content. The same regulation also affirms that disinformation is not part of voters' general freedom of speech; therefore, can be abridged through judicial rulings.

Brazilian law has also been expanded over the years to restrict undue attribution of content to third parties and the use of fake profiles during electoral campaigns. Current law criminalizes or imposes fines against anyone who: (i) unduly attributes the authorship of online electoral advertisement to third parties; (ii) directly or indirectly contracts out individuals or groups to offend candidates or a party through online messages or comments, punishing both the contracting party as well as the contracted parties; or (iii) disseminates content through fake online profiles.

[75] Brazil has a career judiciary similar to France. Judges and prosecutors of lowers courts are selected based on nationwide exams where positions are allocated according to test scores. Judges and prosecutors then slowly move up the ranks until they reach appeal courts or leadership positions. Supreme Court Justices are appointed by the President and confirmed by the Senate, as in the US.
[76] Three of these are Brazilian Supreme Court Justices, two are Justices at the Superior Court of Justice (STJ) and two civilians appointed by the Brazilian President from a short list provided by the Supreme Court (usually lawyers or law professors). Seven substitute-judges can replace the justices in case of absences.
[77] Instead of Supreme Court Justices, TRE members are Circuit Court judges.

Any candidate, political party, electoral coalition, or public prosecutor may file a complaint alleging the violation of Brazilian Electoral Law. Citizens and companies may not file complaints but may be defendants. Complaints are usually filed before district courts. District-level decisions can then be appealed to TREs and, in restricted circumstances, to the TSE.

These decisions are taken under expedite procedural rules—candidates have between 24 to 72 hours to complain against offensive content. Judges must rule on the matter within 24 hours, and normally issue preliminary rulings before hearing defendants. Defendants have 48 hours to either challenge the initial complaint or the interim order. Judges then issue a final ruling. Appeals follow a similar expedited time-frame: 24 hours to file the appeal, 48 hours to file counter-arguments and 48 hours for the appellate court to rule. In some cases, complaints may be filed directly before TREs/TSE, depending on the claimant and the nature of the claim. This was the case for disinformation proceedings relating to the 2018 Brazilian elections: as the election involved only Executive and Legislative positions in the State and Federal Governments, only the TREs and the TSE were involved in adjudication. Relevant to this discussion, the 27 TREs and the TSE had each previously appointed three auxiliary judges (84 in total) that would be responsible for handling disinformation claims.

Judges have basically four alternatives when deciding on a complaint alleging that someone is spreading disinformation in violation of Brazilian laws:

(i) decline jurisdiction (if the matter does not impact the electoral process/results);
(ii) grant candidates a "right of reply," through which they may express their views on the matter on the same venue as the offender;
(iii) issue a takedown order against the infringing content, including orders prohibiting the content from being reposted online; and/or
(iv) impose civil fines—including fines against platforms to ensure compliance.

These can be applied concomitantly, meaning that judges may grant a right of reply, issue a takedown order and impose a civil fine as the result of a single complaint. in their decisions, Judges are expressly required to balance out the potential restriction of freedom of expression rights vis-à-vis the potential harm to isonomy between candidates and the fairness of the electoral process. The law also requires that all complaints against online content identify the specific URL linking to that content, meaning that judges can decide between ordering the takedown of a specific

URL or issuing a general order for removal of that content. Judges may also order the takedown of the entire hosting platform if the platform repeatedly fails to comply with Court rulings. All electoral takedown orders are interim in nature and are automatically revoked once elections are over. Parties may then file civil lawsuits to require the permanent removal of the content and request compensation for any damages.

Over the years, the TSE and some lower courts issued controversial decisions that directly restricted freedom of expression rights in ways that would probably be difficult to reconcile with constitutional protections in countries like the US. These include a prohibition on candidates from openly criticizing their competitors (a prohibition on "negative political advertisement"), one qualifying a satirical YouTube movie where a citizen depicted a candidate as having horns and wearing a clown nose as negative political advertisement (therefore illegal); and another prohibiting candidates from using telemarketing services.

IV. The theoretical virtues and flaws of the Brazilian system

In this context, a key question then becomes under what standards one can judge the institutional design of public policies to combat disinformation like the ones put in place in Brazil and elsewhere. In a seminal article, Prof. Jack Balkin lists four main challenges that any governance system for online speech much overcome to be considered fair and effective:[78] (i) developing effective standards that are protective of freedom of speech; (ii) avoiding a global jurisdiction (or some form of Brussels' effect where one country effectively imposes standards on all others);[79] (iii) preventing an obscure privatized bureaucracy from serving as prosecutor, judge and executioner without any form of accountability; and (iv) preventing nation-states from co-opting private infrastructures for surveillance, data collection and analysis to increase their control over civil society. One could also independently add that a system that intends to monitor and tackle disinformation during elections must be responsive and acknowledge the rapid dynamic of these processes, where new developments may quickly but fundamentally impact public opinion and voting outcomes.

From the perspective of a democratic theory of freedom of speech, one that sees the open and robust public debate as essential to a democracy, Courts may be the right locus (or the most

[78] Balkin, 'Free Speech Is a Triangle' (n 27). pg 22.
[79] On the Brussel's effect, see Anu Bradford, 'The Brussels Effect' (2012) 107 Nw. UL Rev. 1.

acceptable one) to restrict false statements of facts and preserve the public sphere as a space for deliberation. If the intention of the Law protecting freedom of speech is *"to broaden the terms of public discussion as a way of enabling common citizens to become aware of issues before them and of the arguments on all sides and thus to pursue their ends fully and freely"*,[80] judicial intervention against disinformation may help to achieve this end by preserving the overall fairness of the process, avoiding distortions in the information environment. Based on this framework, one can argue that the Brazilian model has several theoretical strengths that (at least partially) address all abovementioned concerns:

i. It is an open (multiple parties may file complaints) and appealable system, where independent judges and prosecutors decide what content should be taken down. This allows for the development of clear standards on what is acceptable online speech through the evolution of case law;

ii. The inherent publicity of court decisions allows some form of democratic accountability of the judicial standards—both civil society and Congress may criticize the results and even pass laws to correct interpretations. The possibility of appeals provides an institutional avenue to correct mistakes;

iii. As access is reasonably open to all parties involved in election disputes, and Courts remain responsible for balancing freedom of speech and election integrity, the system prevents social media platforms from becoming the unique arbiters of online speech;[81]

iv. The reliance on the Judiciary (instead of a regulatory agency, as proposed in the UK or enacted in Singapore), also diminishes the risk that the party in power abuse its authority and illegally prosecutes its own citizens or political oponents;

v. Streamlined procedural rules (most decisions are taken within 24 hours of the complaint being filed) empower the Judiciary to act quicky, twarthing disinformation campaigns at their incipiency;

[80] Owen Fiss, The Irony of Free Speech (1996), Introduction [Kindle Edition].
[81] In theory, the judiciary could even develop standards that are later enforceable by platforms.

vi. Finally, takedown orders are restricted to Brazil and no longer valid after the election ends, diminishing the risk that decisions over-restrict online speech or that Brazil imposes global standards;

Nevertheless, relying on courts to decide what is true or false in a very dynamic electoral process can also lead to distortions tha hinder public discourse. The distinction between facts and opinion can be particularly blurred in a polarized election context.[82] The ideological aspects of elections may impact Courts' ability to act neutrally, in particular in open-ended standards like the one existent in Brazil (i.e. remove *"facts known to be untrue"*). Even evaluating clear factual claims in this context may be complicated, as *"whenever the state attempts definitively to determine the truth or falsity of a specific factual statement, it truncates a potentially infinite process of investigation and therefore runs a significant risk of inaccuracy."*[83] Also, the expedited time-frame can lead to superficial decisions, and the multiple Courts involved may easily spiral into contradictory standards. Another potential shortcoming of the system is that it still relies on candidates, parties or prosecutors to flag topics and file complaints, which could create distortions based on parties' willingness or capacity to litigate.

In summary, the Brazilian framework is an institutional alternative that deserves to be closely evaluated. It represents one of the most structured global attempts to implement an open and fair system to help oversee online speech and, as such, can provide valuable information to countries weighing the risks and benefits of increasing public interventions in this complex world of public speech in digital markets.

V. Research questions and database

This larger project (though not necessarily this paper) aims to evaluate whether Brazilian Courts succesfully balance freedom of expression and the integrity of elections. By evaluating all cases involving takedown requests associated with disinformation/fake news handled by Brazilian Courts during the 2018 Brazilian elections, this project aims to answer five research questions:

[82] According to Robert Post, factual statements claim general validity regardless of community standards, assuming it is possible to achieve convergence on the evaluation of truth or falsity, while statements of opinion claim validity grounded on certain community standards. Robert C. Post, The Constitutional Concept of Public Discourse: Outrageous Opinion, Democratic Deliberation, and Hustler Magazine v. Falwell, 103 HARVARD LAW REVIEW 601 (1990) at 660.
[83] *Id.*, at 659.

(i) Did Brazilian courts develop a clear and consistent standard to measure whether an online publication qualifies as disinformation?

(ii) What are the determinants of judicial decision-making in this area?

(iii) Are courts generally successful in seprating disinformation from legitimate expression when looking at these cases?

(iv) What institutional improvements could better equip Brazilian Courts to handle these cases; and, finally

(v) What lessons can the Brazilian experience teach the world as countries design their systems to tackle electoral disinformation?

This requires first-hand access to Court cases, something that has proven a significant challenge. Brazil has no private or public, comprehensive, searchable case law database. Historically, each Brazilian Tribunal has been responsible for its own IT systems, leading to significant discrepancies. TSE itself hosts a searchable public database[84]. However, this database is not reliable[85] and unfit for robust quantitative analysis.

As an alternative, since 2016, Brazilian electoral courts started using a new, nationwide electronic system known as "Processo Judicial eletrônico," or "PJe". The PJe is slowly digitalizing the entire Brazilian judiciary, meaning that all procedural steps are now electronic: parties file electronic complaints (including evidence), judges then issue electronic decisions, parties file electronic appeals and so on. The use of the PJe was mandatory for all classes of complaints that could encompass disinformation cases during the 2018 elections. However, while the PJe significantly increases the efficiency of the judiciary and is a reliable source of information for researchers, it only allows for individualized access (a researcher can only access one proceeding at a time), and there is no centralized general search functionality.

To overcome these challenges, FGV-CEPI, a research center on innovation and technology policy within FGV Law School in São Paulo, developed a comprehensive electoral case database

[84] Available at http://www.tse.jus.br/jurisprudencia/decisoes/jurisprudencia

[85] First, it is optimized to return fast queries and a maximum of 1000 results; this sacrifices accuracy, meaning that similar searches yield different results. Second, and most worrisome, the database requires judges to manually upload cases to TSE's system, which becomes particularly challenging when judges are overloaded with 24-hour deadlines. The differences in resources between jurisdictions could also indicate that results are biased as judges in more resourceful courts are more likely to upload cases than those in less resourceful courts. TSE judges themselves told us they do not trust their database.

for the 2018 Brazilian Elections.[86] This is a machine-readable database that contains all online disinformation cases and their respective metadata to enable scientific research. The creation of the database was made possible by an unconditional grant from Facebook, meaning that Facebook funded the works but retained no overseeing powers over the construction of the database nor had any influence on overall results. This paper is the first comprehensive research project based on the FGV-CEPI database. While this research project relies on the FGV-CEPI database, it does not rely on any support from Facebook, and Facebook had no early access or other forms of influence on the results of this project.

The construction of the database consisted of three distinct steps. First, FGV-CEPI developed a web scraper that allowed the download of judge-issued documents available at PJe. This effort resulted in a dataset of over 95,000 cases between December 2017 and March 2019, covering more than the entire 2018 electoral period. Next, it developed a word search utilizing regular expressions of terms in three categories: (i) fake news, disinformation, and "known to be untrue" information; (ii) online content, digital content, and social media-related terms; (iii) date and time terms to restrict the focus to the 2018 elections. This resulted in over 850 useful expressions and a preliminary sample of 2,928 cases that match one wording at least in categories (i) and (ii). Finally, CEPI manually curated the cases to filter out any disputes that did not involve online "fake news" claims; that is, online content challenged for untruthfulness, fake news, disinformation, libel or similar. Because of the lack of strict definitions for what constitutes online disinformation, it incorporated a broad scope of language that minimally signifies that a content may be untrue. The final sample consists of **1,492 cases**, which should be close to the entire population of online misinformation cases decided by Brazilian courts during the 2018 Federal and State Elections.[87]

[86] Rodrigo Karolczak was the main researcher in charge of tdeveloping he database.

[87] The web scraping process makes 100 attempts at accessing estimated identification numbers after the last known collected case in a given court. There is regulation on case identification numbering and Courts tend to follow the labeling procedure. However, due to idiosyncrasies or error, a case may be misnumbered. Cases that are numbered in such a way that they are over 100 units of distance since the last known case were not collected, as searching for them would incur significant processing costs. This was a conservative approach to ensure a complete sample. Therefore, while some cases may be outside of our numbering sample, these cases are random and non-systematic and, as such, should not materially impact our results.

For this paper, we then trained a group of 10 undergraduate and graduate law students to manually code these cases, which resulted in **2,850 coded decisions**.[88] We are just finalizing the review of the coding to correct any inconsistencies. Our codebook includes information such as:

(i) the involved parties (e.g. if plaintiffs are incumbent politicians and whether defendants are politicians, platforms, citizens or journalists);

(ii) whether the decision resulted in content takedown, fine imposition, rights of response, and/or bans on reposting similar content;

(iii) The URLs challenged, the type of content (post, image, video, etc.), including whether the specific URL has been taken down or not;

(iv) How plaintiffs characterized their claim (facts "known to be untrue," defamation, libel, slander, offense against honor) and how judges ruled on these claims;

(v) Under what legal basis judges ordered the takedown of the content;

(vi) Objective information on the judges (including gender and appointment process).

VI. Results

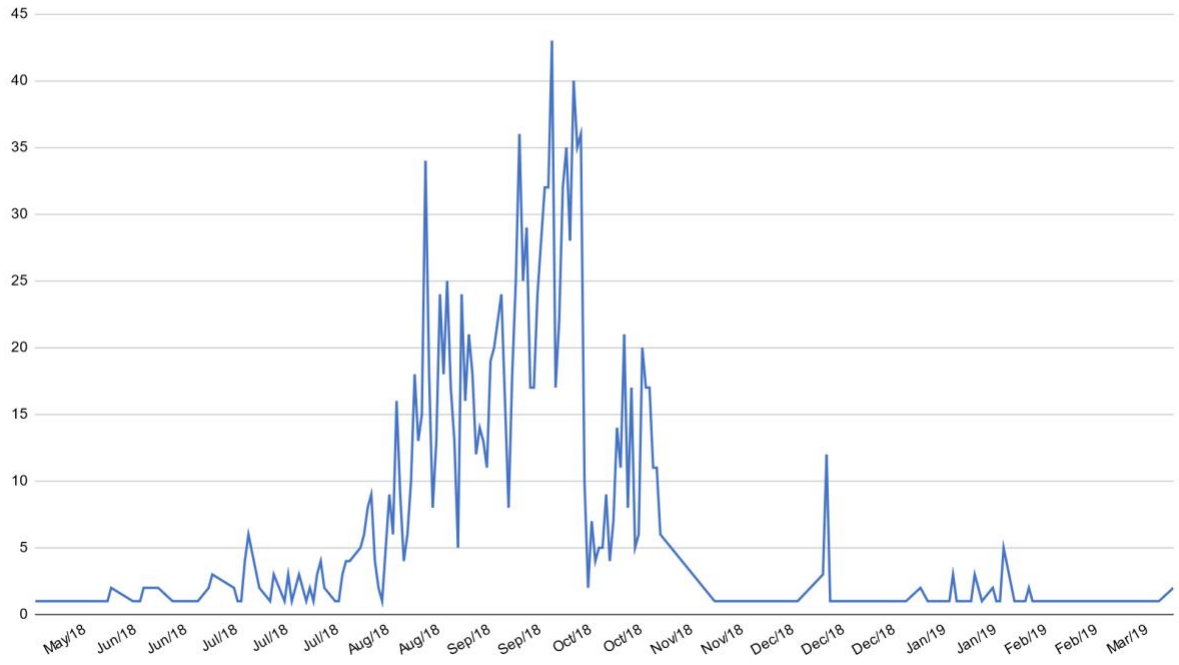
Below you can find preliminary results based on the dataset, prepared for the conference. These so far mostly consist of descriptive statistics that provide an overview of our findings so far.

a. General trends

Our database covers the entire Brazilian election period for 2018. As expected, the number of cases grows as the first and second voting rounds get closer (October 7th and 28th, respectively). It is also worth noting a spike of cases in early August, though we cannot clearly identify why.

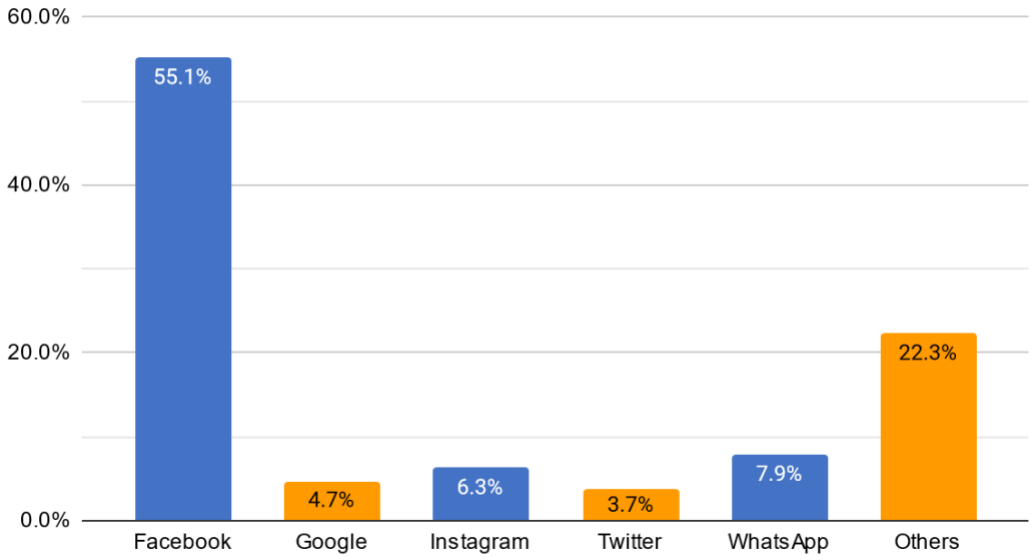
[88] We developed a template for coding the cases. Then, we initially provided a sample of coded cases to teach students how to correctly label the cases. We then split students into groups to promote exchanges and add consistency within the groups. Finally, we held bi-weekly meetings with all students to discuss the coding and settle any potential discrepancies.

Figure 1: timeline of initial rulings



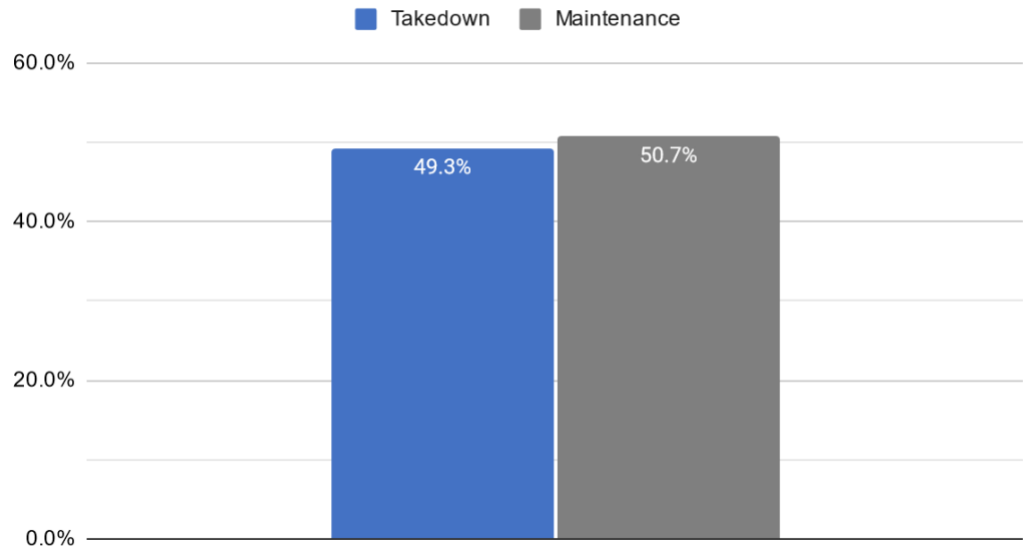
Facebook accounts for the bulk of requests for removals (almost 70% of the URLs present in our database, considering Facebook/WhatsApp and Instagram). Google only received a small fraction of requests (Google encompasses YouTube). Others impacted agents include blogs and other websites such as news platforms.

Figure 2: URLs by platform



Despite the several legal reminders that Courts must ensure freedom of expression, Brazilian judges ordered the takedown of content in almost 50% of cases brought before them, [89] a significant percentage by any account.

Figure 3: Takedown ratio for the full sample



[89] This figure considers "takedown" as the judge ordering the removal of any type of content in a lawsuit. It is not broken down per requested URLs as in Figure 2.

More worrisome, they ordered the takedown of content in 54.5% of cases involving lawsuits against ordinary citizens. When faced with specific requests to takedown entire profiles, judges ordered their entire removal in 27.6% of cases. When candidates were defendants, judges ruled on taking down content in 43.5% of the cases. Moreover, there were 62% more decisions in which citizens are defendants (876) than candidates (540).

Figure 4: Takedown ratio of claims against citizens

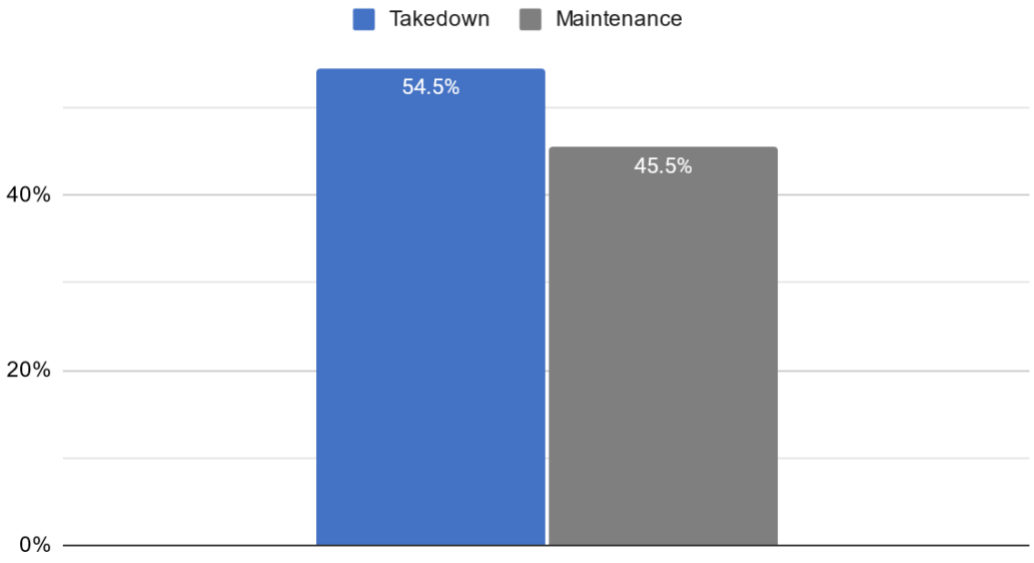
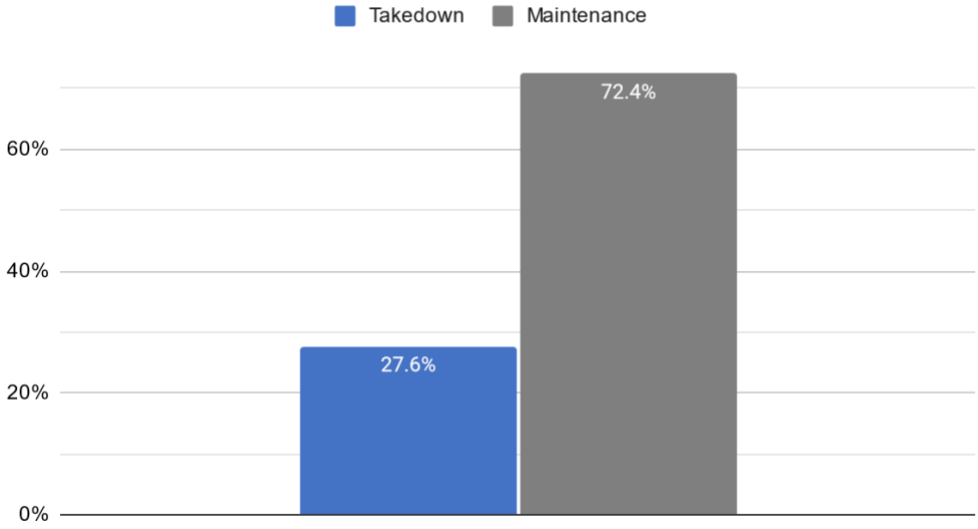
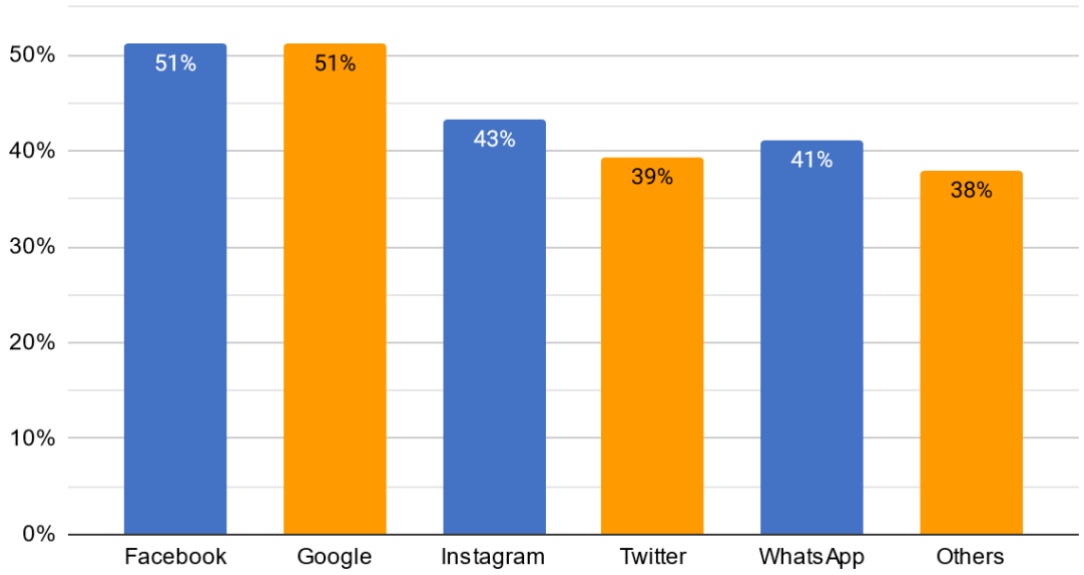


Figure 5: Takedown ratio of claims against social media profiles



Many expressed concerns that fighting disinformation will become harder the more users migrate towards encrypted peer-to-peer communication systems such as WhatsApp. While it represented only 8% of specific requests, judges issued takedown orders in roughly 41% of those cases,[90] a ratio close to other platforms.

Figure 6: Takedown ratio by platform

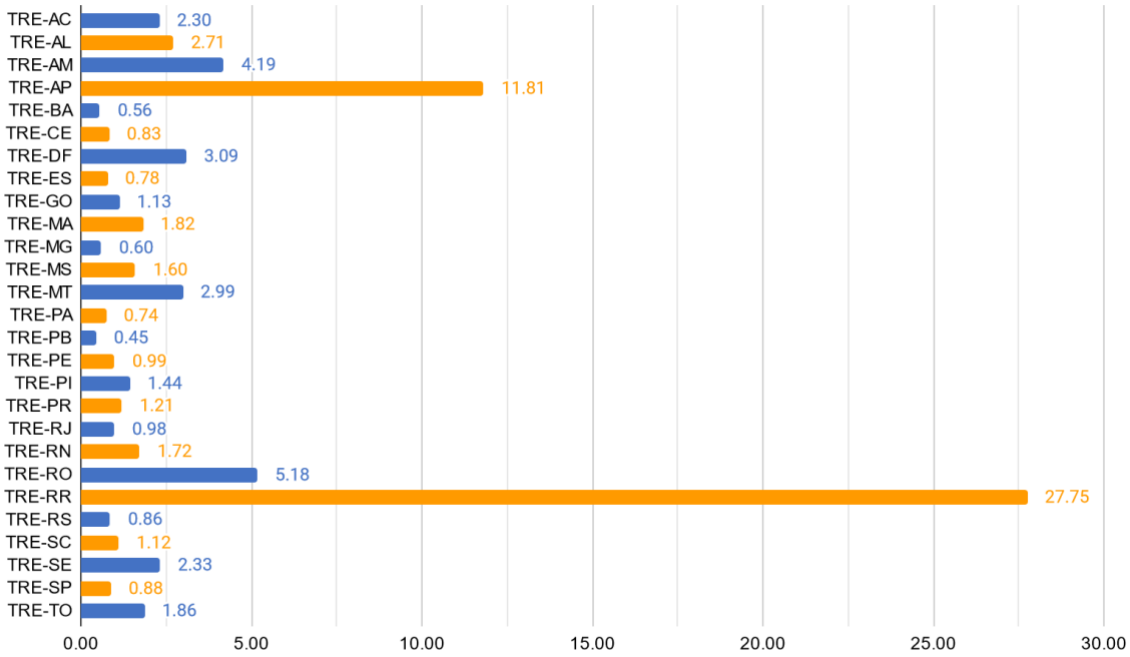


b. Capture

In general, broad freedom of expression rights are justified on a general fear of abuse by those in power: if a door is open for courts to control speech, elites and the government will quickly ensure that their points of view prevail. There is some initial evidence corroborating this fear: TREs in smaller, less economically developed States (e.g. Amapá, Rondônia, Roraima and Amazonas) received a disproportionally larger number of claims when compared to larger, more economically developed Brazilian States (e.g. São Paulo, Rio de Janeiro and Minas Gerais).

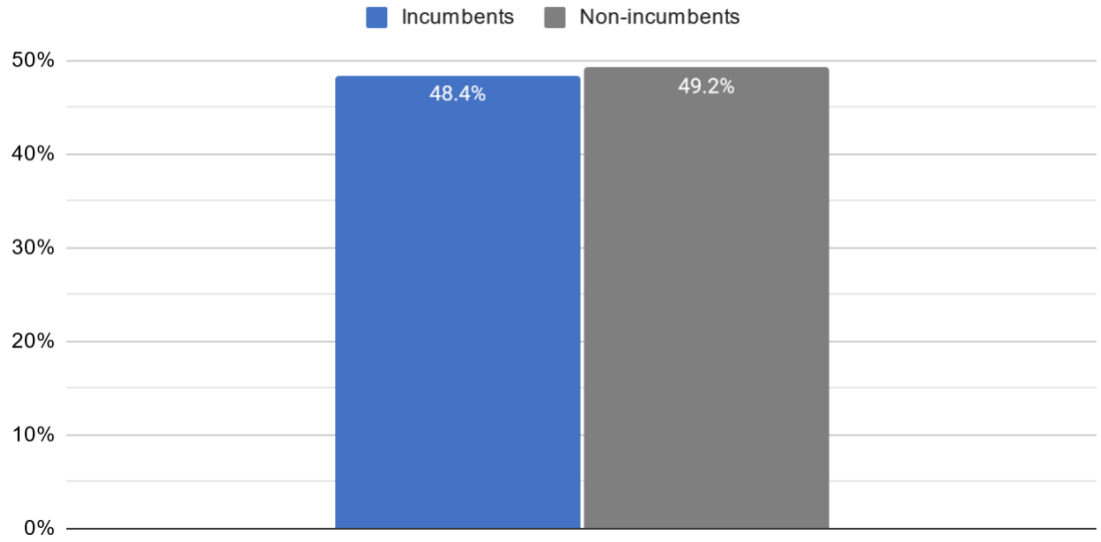
[90] It is difficult, if not impossible, to enforce these orders against individuals or the Whatsapp, considering its peer-to-peer encrypted system. So, it is surprising to see these decisions in the sample.

Figure 7: Cases by population (in 100.000) per State Court



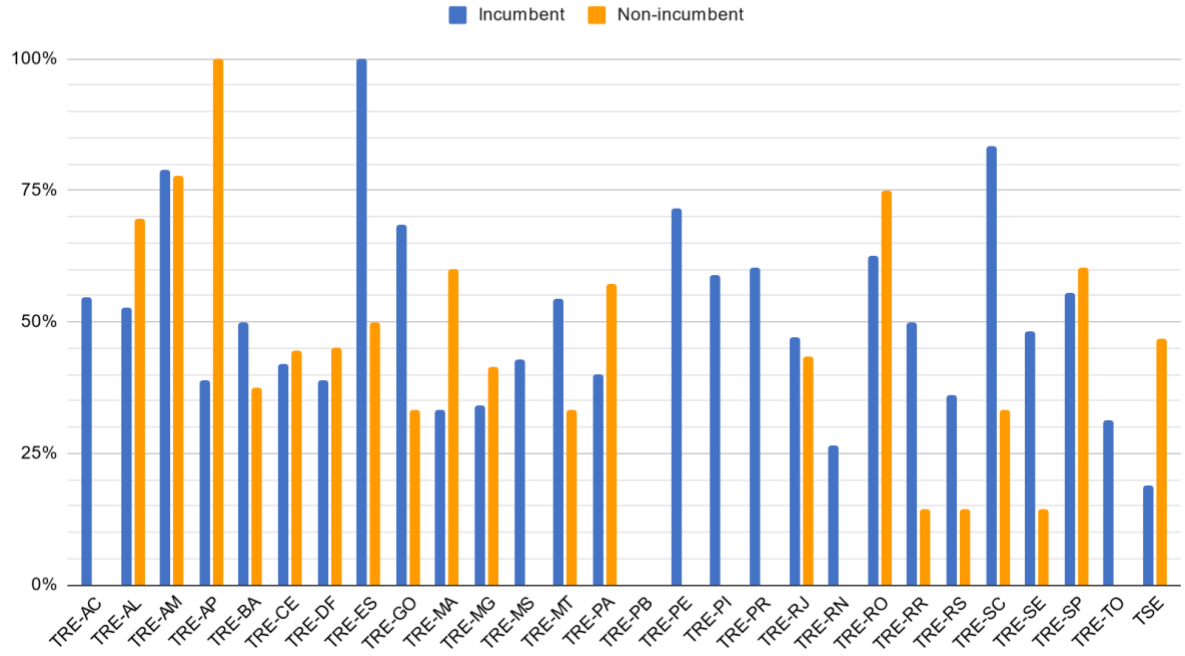
Nationwide, incumbents (considered those who occupied any public office in the year before the election) and non-incumbent politicians were equally likely to win takedown requests.

Figure 8: Initial success rate of incumbents and non-incumbents



Yet, it is interesting to notice how these averages mask significant heterogeneity within States. As shown below, incumbents are apparently significantly more likely to win cases than non-incumbents in many Brazilian States. Indeed, the average and the median win rate for incumbents within States is 49%, versus 35% for non-incumbents—showcasing how there is at least preliminary evidence of some form of capture by politicians. It is worth noting, that the incumbents also used the system much more, being responsible for 1,181 decisions versus only 396 for non-incumbents.

Figure 9: Incumbent and non-incumbent initial takedown success rate by Court



c. Standards

Another concern relates to how judges behave and whether the system may develop coherent standards. Our initial results point to a somewhat worrisome picture. Judges only reverted 15% of interim rulings, which are normally decided without any form of defendant involvement. In addition, in 53% of rulings judges do not quote a single piece of case law, a potential proxy for how argued are the decisions.

Figure 10: Reversal rate of interim ruling

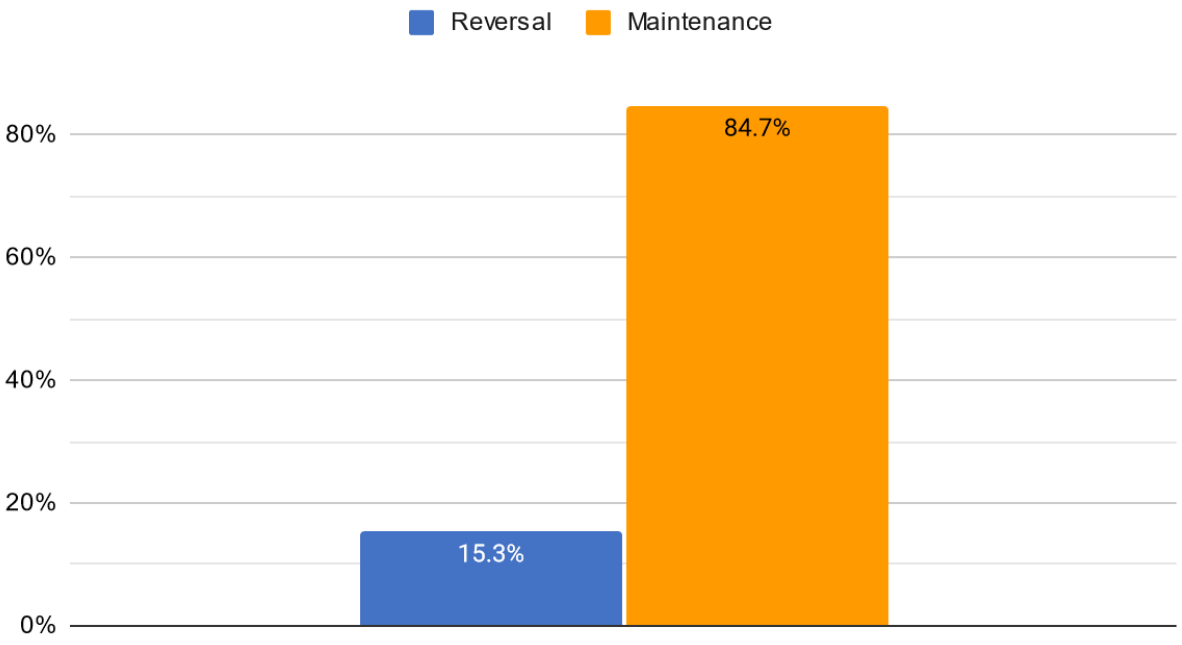
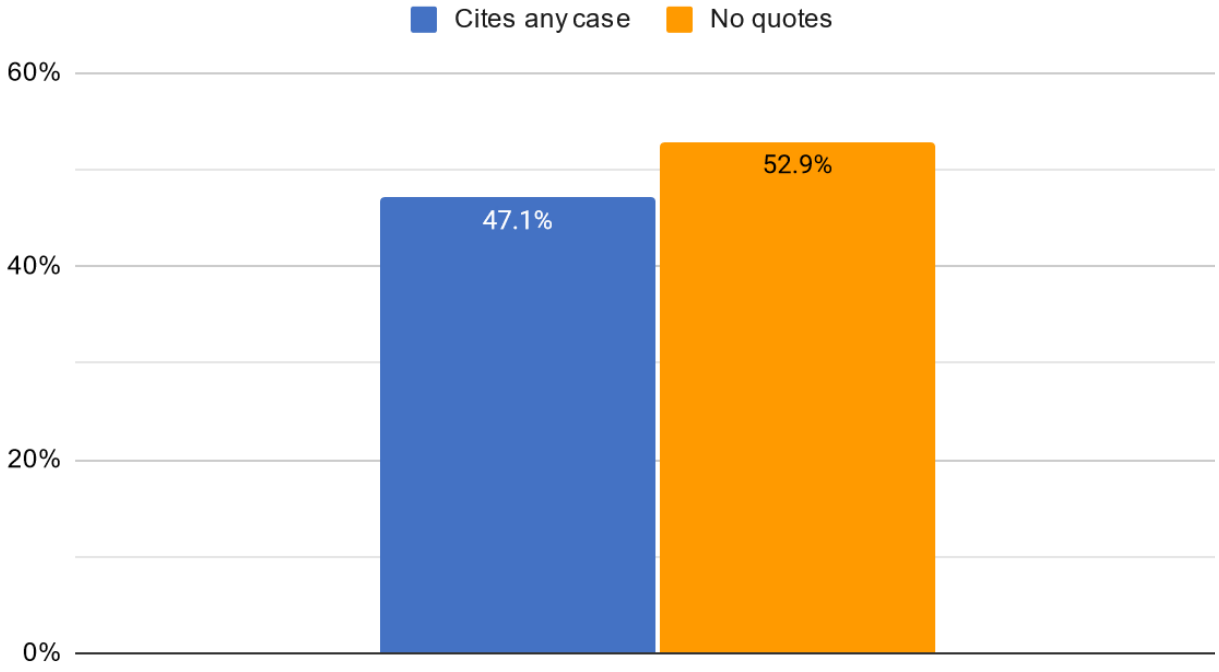
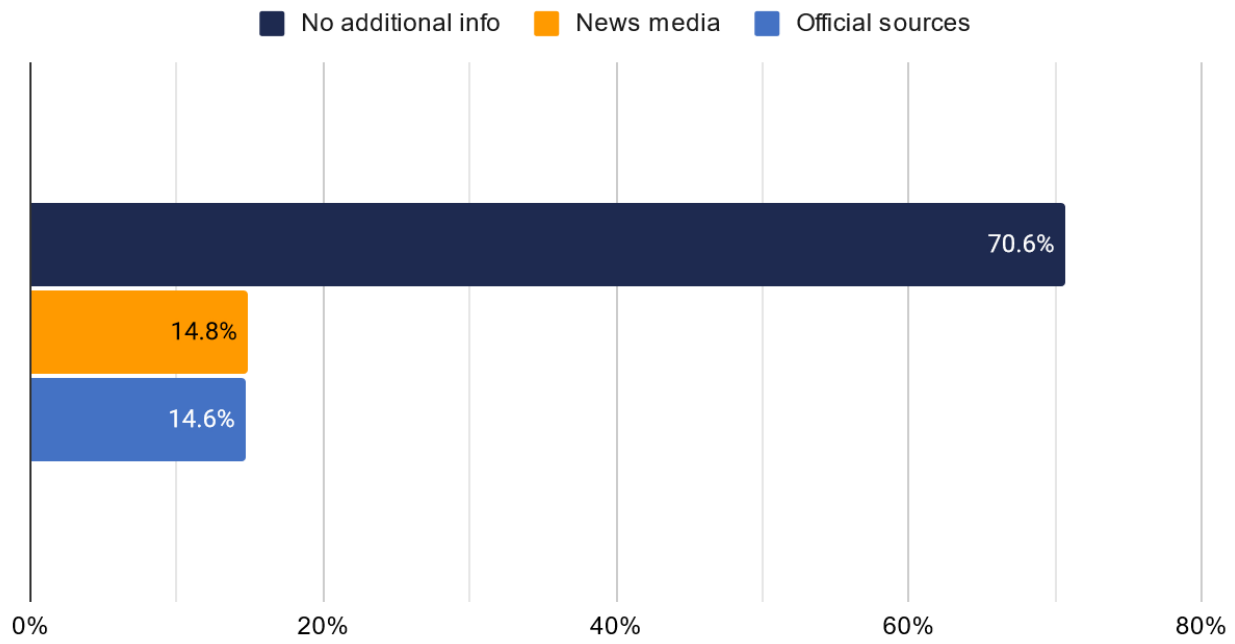


Figure 11: Case law quote rate



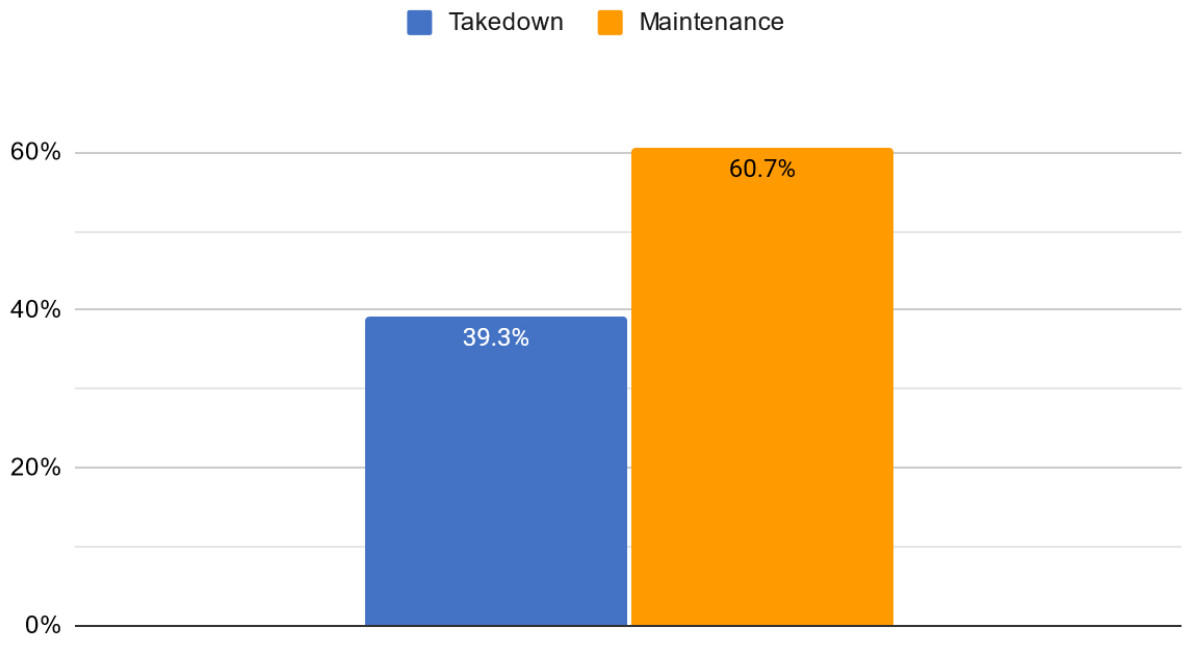
This might indicate that judges mostly decide cases based on first impressions. This concern is corroborated by the fact that in almost 70% of cases, judges did not cite or add any additional information or evidence to their rulings. They appear to simply judge by looking at the content of the material being challenged (complaints must include a copy of the challenged material).

Figure 12: Standards for "facts known to be untrue"



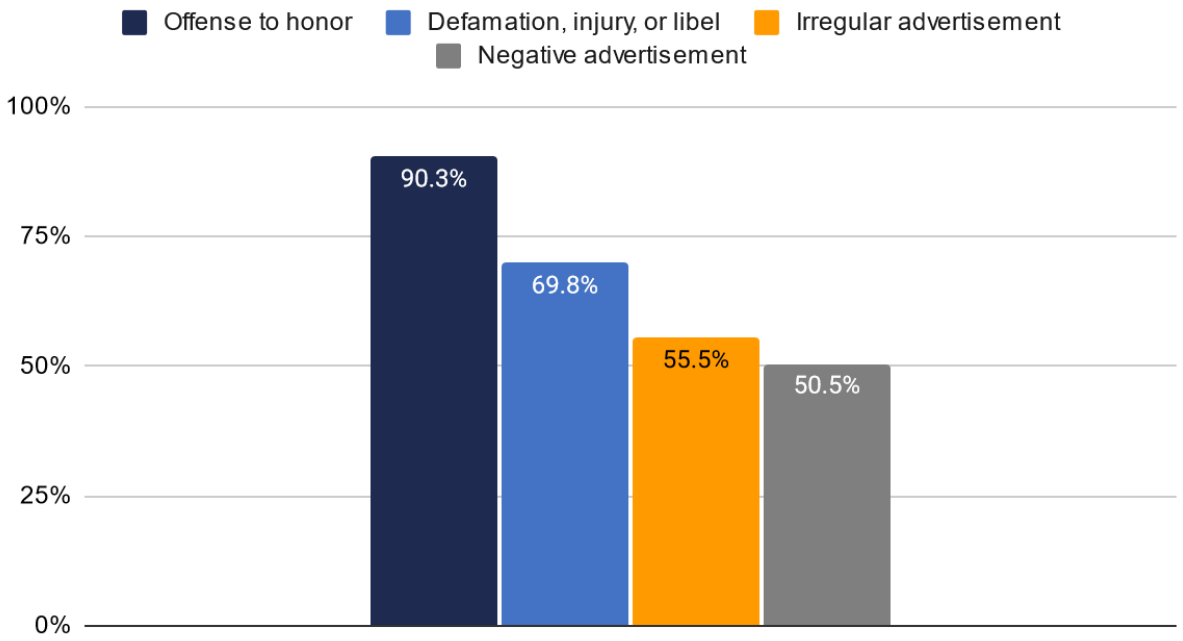
Perhaps surprisingly, takedown ratios are significantly smaller when judges research and quote case law. This may signal either that these are more complex cases that require additional research, that the simple fact of researching the legal standards makes judges more reluctant to take-down content or that judges believe they need bolder justifications to keep content online. We are trying to further investigate this.

Figure 13: Takedown ratio if judge cites any case law



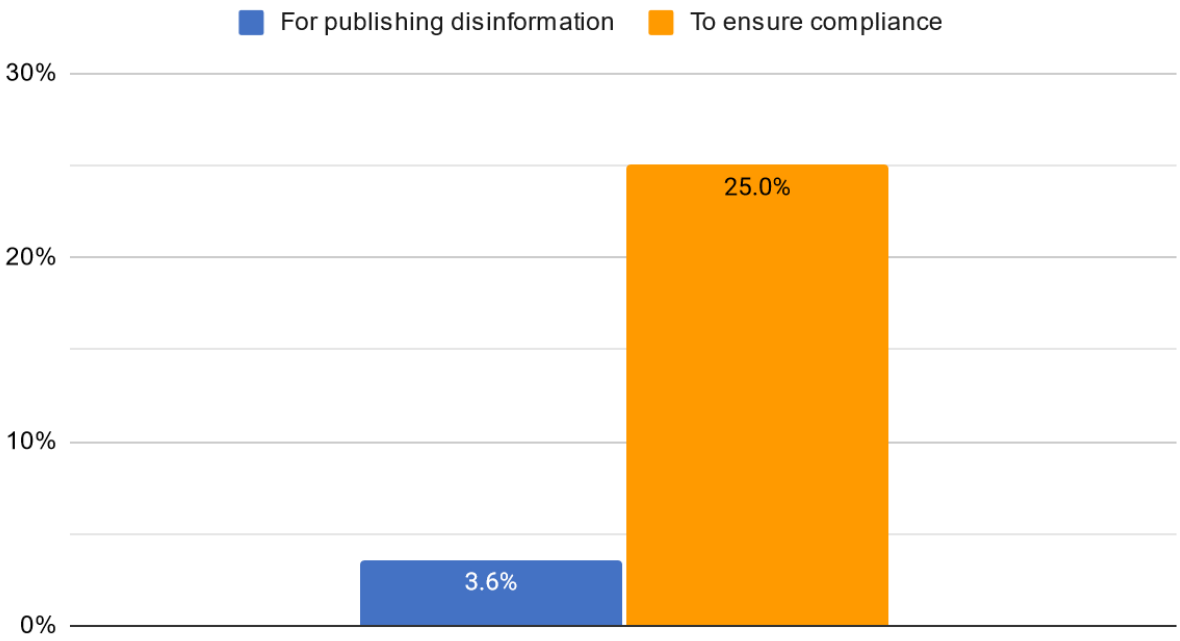
Finally, 90% of cases also alleged offense to honor, and 70% of claims combined accusations of disinformation with a punishable crime, such as defamation or libel. This paints a grim picture of attempts to only remove "illegal content", as done by jurisdictions like Gremany (in its NetzDG law). Speech is a fluid field, and it is easy to claim that sharing disinformation about a given candidate also leads to some form of criminal behavior.

Figure 14: Distribution of other legal arguments by case



Finally, it is worth noting that despite their clear legal powers to impose fines for sharing disinformation and the overall 50% general takedown ratio, judges were very reluctant to fine parties in legal proceedings. In only 3% of cases, judges imposed fines for the sharing of disinformation; in 25% of cases, judges imposed daily fines in case parties refused to comply with takedown orders.

Figure 15: Fine ratio by ruling



VII. Initial Findings

Given the ongoing stage of the research, at this moment we can only share initial findings.

The Brazilian system has many theoretical merits—particularly the introduction of some form of democratic accountability to online speech. Yet, the data on its performance indicates many significant shortcomings, many previously identified in other notice and takedown systems: the burden is unduly placed solely on politicians and public prosecutors to monitor the avalanche of online speech. This monitoring was already costly, and will become almost untenable as interactions move to encrypted, peer-to-peer communications (such as WhatsApp).

Judges also seem to be generally inclined to take down content, potentially overenforcing the provision and harming freedom of speech. Even after accounting for a strong bias in the cases that make it to Court—a candidate has to spend resources to file a legal complaint against the material—a takedown ratio of around 50% in interim decisions (and 54% in cases against citizens) seems high. More worrisome, however, is the low number of reversals, in particular when combined with the fact that most rulings do cite a single case law or any other data source other than the initial complaint. This indicates that decisions are mostly taken based on first-impressions

that are hardly reassessed. This also implies that the legal standard "*facts known to be untrue"* seems overlybroad, depriving judges of proper guidance and allowing them to rely primarly on their instincts or prior knowledge when deciding cases.

Finally, one must note that the system is costly to maintain, and is not designed to be scaled up to handle more cases. The total number of cases (i.e. 1,492) and decisions (i.e. around 2,500) is only a scratch on the surface of the disinformation phenomenon. Yet, electoral courts must adjudicate many other types of cases during elections, making it harder for them to handle a much larger number of disinformation lawsuits. On the other hand, the limited number of appeals and reversals means that the system is also not designed to generate the type of clear cut standards on what is acceptable speech that platforms and other could easily incorporate in their own moderation attempts—depriving it of another potential source of scale.

To us, this appears to be the focus of future reforms: strengthening the development of a coherent appeals process in order to allow the development of clear standards that platforms may incorporate in their reviewing algorithms and/or that judges may use to review cases. Without clearer standards, the impression of random decisions abridging the freedom of expression will undermine the legitimacy of court interventions on alleged disinformation.

The challenge ahead is enormous and will require thoughtful institutional design to devise a system capable of contributing to a healthier information environment and a robust public sphere. Should courts have any relevant role in such a system, the Brazilian experience seems a critical laboratory to watch.