Insper

Research Report

August 15th, 2022

Department: Center for Regulation and Democracy (CRD)

Contact: +55 11 4504-2400 | E-mail: crd@insper.edu.br

CONTENTS

1. Introduction

The fledgling legal research field of (private) online content moderation has spawned a large number of relevant and influential publications[1]. This is arguably a subfield of freedom of speech studies or a part of law and technology research and not all scholars devoted to this study subject identify themselves as focused on "content moderation", though they sometimes interact in conferences organized under this banner[2]. These works have made notable advances in documenting and evaluating the task performed by private platforms and websites – including, but not limited to, social media companies[3] - of determining what speech they allow in their spaces and applying such rules in each isolated case.

In a clear contrast to traditional constitutional law works devoted to assessing the compatibility of legal speech norms with constitutional guarantees and observing the role courts play in such assessment, studies of online content moderation have first described the characteristics of this rising, massive private review of speech and then moved to denouncing its many pitfalls. Works have provided big-picture frameworks of what moderation is on the multitude of networks that compose the internet[4], of the different scale profiles of platform moderation[5] and produced accounts of the challenges resulting from the fact that private companies wield incommensurable power over global speech[6].

---

[1] The reading list compiled by the Social Media Collective is perhaps the best curated source for academic works on content moderation: https://socialmediacollective.org/reading-lists/content-moderation-reading-list/.

[2] See, for instance, the Content Moderation at Scale, the fourth edition of which was held in 2019. Available at https://www.law.kuleuven.be/citip/blog/como-at-scale-brussels-the-4th-edition-of-the-content-moderation-conference/. The same line of studies and similar groups of authors also present and discuss their work on events under the header of "platform governance", such as the Workshop on Empirical Approaches in Platform Governance Research, hosted by the Alexander von Humboldt Institute for Internet and Society (HIIG) in 2020. Available at: https://www.hiig.de/events/workshop-on-empirical-approaches-in-platform-governance-research/.

[3] One early example is work by Laura DeNardis, claiming that the "focus on institutions, while important, sometimes misses core governance functions carried out via arrangements of technical architecture and through policy decisions of private industry" 'Hidden Levers of Internet Control' (2012), 15 Information, Communication & Society, p. 721.

[4] See James Grimmelmann, The virtues of moderation (2015), 17 Yale Journal of Law and Technology, Iss. 1, Art. 2, especially starting at page 55.

[5] This is one of the valuable contributions of Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches (2018), Data & Society Report. Available at: https://datasociety.net/library/content-or-context-moderation/.

[6] Nicolas Suzor, Lawless: The Secret Rules That Govern Our Digital Lives (Cambridge University Press, 2019).

Largely missing from both the predominantly descriptive and openly normative studies is an attempt to acknowledge the difficulties of a private speech review system working parallel to a judicial speech review system that in many countries does not even factor in all of the intricate moderation efforts employed by private platforms. Furthermore, proposals for the relationship between these two avenues for speech review remain hard to find. A first step in this direction, but one that still falls short of addressing these concerns is recognizing certain government-adjacent characteristics of digital platforms as they perform content moderation[7] and the need for careful and data-informed regulatory intervention[8].

Though some works describe in more detail what the Public Administration should do, a workable model requires consideration of how courts can and should fit into this new reality of how the limits of freedom of expression are determined. The level of court intervention and the legal liability standard they apply arguably influences platform rules and attitudes towards speech. If all content moderation decisions were simple and objective, different intermediary liability models would not have to consider risk-aversion and overcompensation. The reality, however, is that platforms make millions of subjective judgment calls about user-generated content using different mechanisms that vary in accuracy. The enforcement of internal content rules can be influenced by the fear of litigation and liability causing it to swing more or less conservative. As we consider new and improved models of intermediary liability, it is crucial to understand *i) why and how platform content review mechanisms can vary in accuracy and consistency and ii) what the effects might be of stricter liability standards on the behavior of social media vis-à-vis the speech they moderate.*

The content policy of most platforms treats court rulings about expression as isolated events that are relevant solely for the specific speech that the claimant litigated. There is no

[7] Many have made this argument. One the most recent and comprehensive contributions, focused specifically on what it means for American law, is Kate Klonick, 'The New Governors: The People, Rules, and Processes Governing Online Speech' (2018), 131 Harvard Law Review.

[8] "In sum, Google and Facebook have the power of ExxonMobil, the New York Times, JPMorgan Chase, the NRA, and Boeing combined. Furthermore, all this combined power rests in the hands of just three people." Luigi Zingales, Filippo Maria Lancieri, 'Committee on Digital Platforms: Policy Brief. Chiago Booth' (2019), Stigler Center for the Study of the Economics and the State.

institutionalized mechanism to allow or even stimulate jurisprudence on free speech produced by constitutional courts or international human rights courts to influence and enhance the platform's abstract content policy. Scholars have consistently criticized this, indicating that international human rights law would offer some support for content policy systems[9].

Platforms seem to assume they can treat expression removed because of court orders and expression removed because of terms of use violations as substantively separate and distinguishable. There are certain elements in this distinction that become increasingly harder to uphold. Most of the speech categories that companies decided to ban are illegal in the majority of countries they operate. They have constituted the bread and butter of judicial review of speech for decades.

As this separation continues to be forced, another unanswered question about the role of courts surfaces. If certain speech is litigated and upheld by a court with the argument that the Constitution guarantees it either because of substantive or procedural protections, can platforms then continue to censor that specific message or that type of speech without any kind of limitation on their autonomy?[10] The point here is not to defend or criticize court rulings that penalize social media companies for what they see as inaccurate content moderation. We worry about the overlooked effect of courts applying liability standards.

The threat of liability posed by a potential court ruling that says the platform should have removed a certain post usually leads the company to improve its community standards

[9] "[H]ow much will it matter ten or fifteen years from now that the First Amendment (and international human rights law) protect freedom of expression, if most communication happens online and is regulated by private platforms that do not—and are not required to—adhere to such long standing substantive norms on expression?" Evelyn Mary Asward, The future of freedom of expression online (2018), 17 Duke Law & Technology Review, 1, p. 31. "Viewing social media platforms through an Internet governance lens suggests several distinct areas of inquiry. One is the question of how national statutory mechanisms or international legal instruments attempt to, or should, regulate social media, whether for intellectual property rights enforcement, antitrust, privacy or other public interest concerns." Laura DeNardis, A.M. Hackl, Internet governance *by* social media platforms (2015), 39 Telecommunications Policy, p. 762.

[10] This phenomenon has been described as 'content reactivation'. In Brazil, there already is a large number of court rulings ordering social media platforms to reinstate content or accounts, with Superior Court of Justice precedent going so far as upholding a US$ 50 thousand penalty for a noncompliant social network. See "STJ mantém multa de R$ 254 mil ao Facebook por demora na reativação de página do Instagram". Migalhas. Sept 1st, 2020. Available at: https://www.migalhas.com.br/quentes/332732/stj-mantem-multa-de-r--254-mil-ao-facebook-por-demora-na-reativacao-de-pagina-do-instagram.

accordingly and enforce them so as to avoid a future conviction. The reverse is possibly also true: if courts, again based on a certain legal or constitutional standard, consistently reverse platform decisions to remove certain expression then the community standards would gradually evolve to reflect this more permissive stance. Again, the main question is: what are the effects of a legal liability standard and subsequent court rulings on the decisions a platform makes regarding its content moderation rules and their application? Would a stricter liability standard cause platforms to preemptively censor more expression than they otherwise would? How could we test such a hypothesis? What are ways to measure these effects in terms of platform behavior and social costs? The number of studies attempting to answer these questions is remarkably small.

2. How the Liability Standard Matters

As previously stated, why and how platform content review mechanisms vary in accuracy and consistency are important questions as they condition the possible effects of different intermediary liability regimes. A stricter or looser intermediary liability standard can cause significant variation in platform behavior towards expression before it is ever subject to legal requests for removal. That is because each decision about a specific post is highly subjective and context dependent. There will be room for different possible interpretations and because this is about expression, the law should ensure platforms – and, therefore, users - appropriate breathing room. Social media companies know in advance that no system for making this decision on billions of posts, videos and comments will ever hit the target squarely every single time. Legislators should understand that as well. As platforms make an effort to reduce the margin of error in one direction or another, they are susceptible to over or under censor as a reaction to any given legal liability standard. Furthermore, the more inconsistent and inaccurate a moderation mechanism is, the more variation a risk-averse behavior can cause.

This fact has been surprisingly absent from the debates and rationale used by legislators and public authorities when they make decisions about the role and responsibilities of platforms. Most national legal responses to the highly complex task of ascertaining what online speech should be restricted either give private platforms a blank check to remove whatever they see fit with no consequences or, more often, overestimate the capability of private platforms to review millions of individual stances of expression within short deadlines - subsequently fining them for failing to achieve the impossible. Intermediary liability is a trade-off and the costs must be measured or estimated whatever the chosen or proposed model.

What the solutions in e.g. the North-American CDA[11], the German Netzwerkdurchsetzungsgesetz[12] and the Brazilian Marco Civil[13] underestimate is the level of sophistication and diversity of content moderation tools afforded by technology and decentralized gatekeeping. To some extent, this is due to a difficulty to realize how unfit the constitutional rights balancing model has become for online speech today[14]. Legislators often disregard that hundreds of millions of posts are periodically already removed by digital platforms regardless of court orders in an effort by these companies to comply with the law and also fight harmful legal speech. At the same time, governments fail to even ask questions about either the accuracy of automated decisions on content or the context in which human moderators do their jobs.

As the legislator debates or implements stricter intermediary liability standards, impact analysis assessments are thoroughly absent. Social costs are sometimes considered, but never properly weighed. There is questionable quality of analysis and collateral damage involved in the two alternatives for large-scale assessment of illegal or harmful content and the mechanism for their external review. Governments seem to have placed much faith on private (automated or blue-collar-work) and public (judicial) decisions. Stricter standards have aligned the

[11] Section 230 of the Communication Decency Act (CDA, 1996) famously states that "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider" (47 U.S.C. § 230). This immunity has been discussed extensively and, when criticized, it is normally based on the view that it provides an arrangement that errs on the side of too much speech. One possible countervailing interest that is not sufficiently addressed is the protection of youth, as they "tend to focus more on the potential benefits of information disclosure than they do on potential harms." Urs Gasser et al., Response to FCC Notice of Inquiry 09-94. Empowering Parents and Protecting Children in an Evolving Media Landscape (2010), p. 12. Available at: http://ssrn.com/abstract=1559208. More recently, some level of jurisprudential erosion of such immunity would be required in order to protect against online terrorism. Danielle Keats Citron, Benjamin Wittes, The Internet will not break: Denying Bad Samaritans Sec. 230 Immunity (2017), 86 Fordham Law Review; Anka Elisabeth Jayne Goodman, 'When You Give a Terrorist a Twitter: Holding Social Media Companies Liable for Their Support of Terrorism' (2018) 46 Pepp L Rev 147, especially starting at p. 182.

[12] Gerald Spindler, 'Internet Intermediary Liability Reloaded The New German Act on Responsibility of Social Networks and its (In-) Compatibility with European Law' (2017), 8 JIPITEC. The author points to elements of the German law that result in disproportional suppression of speech.

[13] One of the biggest substantive innovations of the Brazilian Civil Rights Framework for the Internet was the adoption of judicial notice and takedown, a liability standard much more protective of online speech than notice and takedown. For an explanation of this new model, see Nicolo Zingales, The Brazilian approach to internet intermediary liability: blueprint for a global regime? (2015), 4 Internet Policy Review 4.

[14] I attempt to explain this in the first half of Ivar A. Hartmann, A new framework for online content moderation (2020), 36 Computer Law & Security Review.

incentives for an increase in automated decisions that are supposed to efficiently solve the problem. The result so far is that they push platforms further against the wall while civil society notices that the spread of hate speech and fake news has only increased.

How have these systems fared? How susceptible are they to risk-aversion tendencies, overcompensation and thus excessive removals? The alternatives for speech review that strict liability, result-focused standards rely on surely do not have to be perfect. But understanding their shortcomings is key to estimating the effects of diverse liability regimes.

2.1 The Alternatives

It might seem at first glance that judicial review would be a safe bet for the course-correction in cases where platforms fail to comply with the legal standard. However, court review is both unsustainable and subject to as much, if not more, bias than the human decisions produced by private content moderators. As platforms publish more speech in the last few years, millions of posts are reviewed every day either because they are flagged, notified or detected by artificial intelligence, or as part of punishment administered to users[15]. Notice and takedown liability schemes fail in that they keep this review away from courts, inflating the power of platforms over speech all the while creating significant incentives for companies to remove everything that is subject to notice[16].

Judicial notice and take down at least reduces the perceived risk of platforms, such that they are not as inclined to remove first and ask questions later. Very few studies have raised a red flag about legal content that platforms unintentionally remove because of misaligned incentives. Danielle Citron was one of the first to do so[17]. Social media companies obviously

[15] YouTube alone removed whopping 11.401.696 videos in only three months between April and June, 2020. Automated detection might push numbers up, but even excluding removals after automated flagging the number is still 552.062. Available at: https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:Y2020Q2;exclude_automated:human_only&lu=total_removed_videos.

[16] Wendy Seltzer, 'Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of The DMCA on The First Amendment' (2010), 24 Harvard Journal of Law and Technology.

[17] Danielle Keats Citron, 'Technological Due Process' (2008), 85 WASH. U. L. REV.

cannot be required to sustain perfect filtering, so the more governments and courts push for a decrease in false negatives, with stricter liability regimes, the more companies will increase the false positives – legal content that is unduly removed.

Few content moderation studies acknowledge that judges are also biased, and this affects their rulings on speech[18], with the aggravation that courts are rarely racially or gender-inclusive, especially at the top[19]. This is especially troubling given that a large share of abusive speech that society is trying to curb is precisely hate speech targeted at minorities. While civil society rightfully put the pernicious effects of private moderation under the spotlight, it is worth keeping in mind that more Judicial review of speech is not inherently a good thing – especially when it is purposely viewpoint-oriented. The balancing of fundamental rights, a solution offered by constitutional law scholarship in many countries for judges to tackle hard speech cases[20], is very sophisticated in theory, but its application by lower courts has seldom been tested quantitatively. Empirical results from Brazil indicate a chasm between doctrinal intent and judicial practice[21] that results in unpredictable rulings that are much less protective of speech

[18] Lee Epstein, Christopher M Parker, Jeffrey A. Segal, 'Do Justices Defend the Speech They Hate? In-Group Bias, Opportunism, and the First Amendment' (2013), American Political Science Association Annual Meeting.
[19] This is a worldwide phenomenon. See Alice J. Kang et al., Breaking the Judicial Glass Ceiling: The Appointment of Women to High Courts Worldwide (2020), The Journal of Politics (preprint), https://doi.org/10.1086/710017.
[20] There is a monumental number of works in different countries on constitutional law and rights balancing, especially in the case of solving conflicts between free speech and opposing rights such as privacy and honor. Alec Stone Sweet, Jud Mathews, 'Proportionality balancing and Global Constitutionalism' (2008), 47 Columbia Journal of Transnational Law. Balancing has been adopted by international courts (see, for instance, Eduardo Andrés Bertoni, 'The Inter-American Court of Human Rights and the European Court of Human Rights: a dialogue on freedom of expression standards' (2009), 03 European Human Rights Law Review; Jean-François Flauss, 'The European Court of Human Rights and the Freedom of Expression' (2009), 84 (03) Indiana Law Journal) and national, constitutional courts such as the German (see Klaus Stern, *Das Staatsrecht Der Bundesrepublik Deutschland. Band IV/1. Die einzelnen Grundrechte* (C.H. Beck, 2006), p. 62; Christian Starck, *Kommentar zum Grundgesetz. Band I. Band 1, Präambel, Artikel 1 bis 19* (Franz Vahlen GmbH, 2005), p. 591; Josef Isensee, Paul Kirchhof, *Handbuch des Staatsrechts. Band IV – Freiheitsrechte* (C.F. Müller Juristischer Verlag, 1989), p. 662) and Brazilian courts. One of the current Justices of the Brazilian Supreme Court had already defended this model in a widely cited work 10 years prior to his joining the Court: Luís Roberto Barroso, 'Colisão entre liberdade de expressão e direitos da personalidade. Critérios de ponderação. Interpretação constitucionalmente adequada do Código Civil e da Lei de Imprensa' (2004), 235 Revista de Direito Administrativo.
[21] Studies produced with data science and random sampling show that state trial courts in Rio de Janeiro, the second largest court in the country, when deciding about freedom of expression, apply balancing or cite any jurisprudence at all in only half of the rulings. Ivar A. Hartmann, A Liberdade de Expressão na Primeira Instância do TJ-RJ (2020), 18 Revista Opinião Jurídica issue 27. When the Brazilian Supreme Court decide this type of case in the last two decades, 70% of its citations to precedent referred to rulings by the Court itself where not workable

than expected. Unpredictability increases risks for platforms, likewise increasing the incentive to remove speech they are not entirely certain to be illegal.

A large part of the risk in judicial review of abusive content such as hate speech, defamation and fake news is arbitrary decisions where a judge believes their common sense will enable them to reach a fair verdict. Ironically, the manual work[22] performed by private contractors of companies paid by digital platforms is much less concerning in that respect. Moderators are trained to apply extremely detailed sets of rules for speech without overthinking the rules themselves and prioritizing the value of consistency and coherency. Some level of predictability, however, is the only positive aspect of this second alternative. Even though they perform essentially the same job as judges ascertaining whether certain expression is permissible according to certain rules, private moderators receive a small fraction of the formal training and compensation[23]. But this is not the most urgent problem.

Unlike judges, their entire heavy workload consists of arbitrating the merits of posts, pictures, and videos. Because only questionable content is directed to their attention, private moderators are consistently exposed to the most vile and terrifying expression known to mankind. The profound and probably long-lasting negative effects on their mental health have been documented by qualitative studies[24]. At the very least, in order for research to begin

substantive standard on freedom of expression had been presented. Ivar A. Hartmann, A Crise dos Precedentes no Supremo: O Caso dos Precedentes sobre Liberdade de Expressão (2020), 6 Revista Estudos Institucionais issue 1.

[22] The first widely circulated report on the work of content moderators shocked both because of its information and because of the utter lack of transparency that had shrouded this moderation strategy until then. Adrian Chen, Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads' (2012), Gawker. Available at: https://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads.

[23] "Facebook says in a company statement that moderators receive 'extensive training' that includes 'on-boarding, hands-on practice, and ongoing support and training.' Gray describes his training as only eight days of 'pretty cursory' PowerPoint displays presented in rote fashion by a CPL staff member." Paul M. Barrett, Who Moderates the Social Media Giants? A Call to End Outsourcing (2020), NYU Stern Report, p. 13. Available at https://www.stern.nyu.edu/experience-stern/faculty-research/who-moderates-social-media-giants-call-end-outsourcing.

[24] Sarah Roberts conducted a large number of interviews with private moderators, resulting in the most comprehensive empirical study on the topic so far. Sarah T. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media (Yale University Press, 2019).

assessing the sustainability of this alternative, more data is required about the numbers, profiles, work conditions and decision reversal rate of private moderators[25].

If one could look past this, there is the issue of local and regional culture. Whatever the work conditions of private moderators, the system will remain highly questionable if platforms assign workers to make decisions on content produced in a culture they do not know or understand[26]. Most if not all types of abusive speech suppressed by companies is directly tied to a particular language and set of mores. Addressing hate speech produced in a cultural and geographic setting with groups of moderators that are disconnected from it results in a high number of both false positives and false negatives[27].

Propping up artificial intelligence against abusive content is the alternative with the lowest human and financial cost in implementation and the most consistent decisions – but only at the surface[28]. The worst part of automated content moderation is that policy makers usually misunderstand its functioning and overestimate its accuracy. It is much easier to explain to lawmakers the downsides of the previous two alternatives than to make them fully understand the difference between supervised and unsupervised machine learning or to have them grasp what the relationship is between regressions and natural language processing. The opacity of artificial intelligence[29] has allowed an overconfidence in its accuracy for arbitrating difficult speech review cases. As a result, policy makers tend to see a strict intermediary liability standard coupled with fully

[25] "Outsourcing saves social media companies significant amounts of money on moderation, just as it lowers costs for janitorial, food, and security services. Contract moderators don't enjoy the generous pay scales and benefits characteristic of Silicon Valley. Outsourcing also has given the tech companies greater flexibility to hire moderators in a hurry without having to worry about laying them off if demand for their services wanes." Barret, supra note, p. 4.

[26] Barret, for example, suggests social media companies should "Further expand moderation in at-risk countries in Asia, Africa, and elsewhere" supra note, p. 25.

[27] Most observers agree that the lack of sensibility for regional cultural realities contributed to grave human rights violations in Myanmar. See Timothy McLaughlin, How Facebook's Rise Fueled Chaos and Confusion in Myanmar (2018), Wired. Available at: https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar/.

[28] Robert Gorwa, Reuben Binns and Christian Katzenbach, Algorithmic content moderation: Technical and political challenges in the automation of platform governance (2020), Big Data & Society, January–June, 1–15.

[29] Matthew U. Scherer, Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies (2016), 29 Harv. J. L. & Tech. 353.

automated content moderation as a solution with a small or nonexistent margin of error and no visible, substantial fallout. None of these assumptions has been empirically tested.

A research project on multistakeholder roles in content moderation performed over fifty interviews with members of the Brazilian Congress, the Judiciary – including sitting judges of the Brazilian Superior Court of Justice – and the Federal Administration, as well as journalists from national circulation newspapers. In the semi-structured interviews where interviewers prodded them about the challenges of fighting abusive content on platforms, interviewees never even *mentioned* concerns with automated moderation. Nobody was remotely preoccupied with accuracy, transparency or explainability[30] of the models used by social media companies to cast hundreds of millions of decisions about freedom of expression[31].

Under pressure to come up with solutions, lawmakers in many countries have turned to explicit or indirect obligations for platforms to swiftly and silently remove as much problematic content as possible using software[32]. Meanwhile, studies indicate that artificial intelligence is still years away from being a decent substitute to human moderators in ascertaining the limits of free speech[33]. Prior automated filtering to detect copyright violations already used by YouTube for several years remains a blunt and opaque instrument impervious to any public accountability[34].

---

[30] "Software also faces limits of explainability, which is a problem for legal decision- making. Software can often explain how it reached a decision, but not why. That may be fine for a thermostat, but is a limitation for a system that is supposed to both satisfy those subjected to it and prompt acceptance of an adverse ruling." Tim Wu, Will Artificial Intelligence Eat The Law? The Rise of Hybrid Social-Ordering Systems (2019), 119 Columbia Law Review, issue 1, p. 21.

[31] Ivar Hartmann, Yasmin Curzi, Nicolo Zingales and Clara Almeida (Orgs.), Moderação de conteúdo online: contexto, cenário brasileiro e suas perspectivas regulatórias (Alameda, 2022, forthcoming).

[32] Section 3, (2), 2 of the German Network Enforcement Act of 2017 forces platforms to remove "manifestly unlawful" content in 24 hours. The European Directive on copyright in the Digital Single Market that came into force in 2019 almost featured an obligation for platforms to use upload filters. The EU Parliament later watered down that clause – to the dissatisfaction of copyright attorneys. See Axel Nordemann, Upload Filters and the EU Copyright Reform (2019), 50 International Review of Intellectual Property and Competition Law.

[33] State of the art automated moderation technologies today allow at best for a mixed system of human and automated moderation. OFCOM, Use of AI in Online Content Moderation (2019). Report, p. 35. Available at: https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

[34] "Going beyond the statutory framework, voluntary mechanisms of algorithmic copyright enforcement do not afford alleged infringers with even the minimum due process protections set by the DMCA: they do not grant alleged infringers the right to contest content restrictions through a counter notice procedure, and they do very

The use of artificial intelligence in content moderation runs much deeper than merely removing abusive content, a process that only subjects a subset of online expression. AI is also the foundation of recommender systems. It does not influence solely what content platforms keep us from seeing, rather it also decides with what content we actually end up interacting[35]. This parallels the fact that purchase decisions on Amazon are increasingly the result of successful algorithm recommendations[36]. Algorithms that decide who should be exposed to what content in order to drive engagement and raise ad revenue are therefore a much more influential cog in the social media machine than content moderation schemes. Regulating recommending[37] would have far more impact than imposing stricter liability standards with the hopes of regulating content itself. We should expect, but have not yet tried to model or measure, the effects of stricter intermediary liability regimes on recommending system design and practice. Would content of ambiguous legality be recommended less often to less people?

Policy makers tend to overestimate the performance of platform mechanisms that issue decisions on specific instances of speech – be they in disagreement with community standards or with the law – and therefore underestimate their inaccuracy in the regulation of expression. In a notice-and-takedown jurisdiction, platforms are forced to anticipate the outcome of unpredictable court rulings about speech and then design and implement fallible content review systems by contracted moderators or artificial intelligence that attempt to hit a moving target. Under stricter

---

little in terms of validating copyright ownership rights." Maayan Perel and Niva Elkin-Koren, Accountability in Algorithmic Copyright Enforcement (2016), 19 Stanford Technology Law Review, p. 508.

[35] "In fact, it's the algorithm that chooses what to show a user that is credited with TikTok's popularity, and it's the ultimate ownership of that algorithm that is the sticking point in the sale of the company." Gregg Leslie, TikTok and the First Amendment. TikTok users have free speech rights—and courts should pay attention (2020), Slate. Available at: https://slate.com/technology/2020/09/tiktok-wechat-first-amendment-free-speech.html.

[36] Back in 2013, "35 percent of what consumers purchase on Amazon and 75 percent of what they watch on Netflix come from product recommendations based on such algorithms." Ian MacKenzie et al., How retailers can keep up with consumers (2013), McKinsey & Company. Available at: https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers.

[37] For a detailed account of the pervasiveness of algorithmic influence in the behavior of social media users, see Jennifer Cobbe and Jatinder Singh, Regulating recommending. Motivations, Considerations, and Principles (2019), 10 (3) Europen Journal of Law and Tecnology. The authors point that algorithmic editorializing is not covered by safe harbor intermediary protection. Instead of suggesting strict liability to regulate this practice, however, the authors defend a more narrowly-tailored approach that involves procedural guidelines.

intermediary liability standards, the impact on platform behavior could be significant, generating a substantial excess of removals and thus significantly hurting freedom of expression. These effects must be described and estimated.

3. Economic Literature on Social Media

The economic importance of laws is a traditional research topic of economic literature. Economic agents do not interact with each other removed from their legal environment, making legal institutions essential to economic efficiency.[38] There is a strand of scientific works assessing long-term impact of legal origins to economic development.[39]

On a micro level, there are several works that evaluate the economic impact of regulations raging from minimum wage[40], healthcare [41] and crime[42]. Recently, there are a seriesof articles that analyzed how the introduction of the European General Data Protection Regulation (GDPR) in 2018 affected multiple markets, from advertising to healthcare. For example, Peukert et al.[43] showed how internet traffic between the European Union and other regions was affected by the legislation, while Goldberg, Johnson e Shriver[44] presented the effects of the data protection legislation on ecommerce.

The economic literature has methods to check and evaluate ex-post effects of regulation even if knowledge of how a specific market operates is not completely entirely developed. This is the case with online content moderation: while there is an increasing number of studies on

[38] In his seminal article Coase argued that well defined property could lead to efficient bargaining between agents. On the other hand, under positive transaction costs, such as those incurring from legal uncertainty, the economic welfare could be hampered. Coase, R. H. "The Problem of Social Cost." *The Journal of Law & Economics*, vol. 3, 1960, pp 1–44.

[39] For example, see the works of Mahoney, Paul G. "The Common Law and Economic Growth: Hayek Might Be Right." The Journal of Legal Studies, vol. 30, no. 2, 2001, pp. 503–25. and La Porta, Rafael, et al. "The Economic Consequences of Legal Origins." Journal of Economic Literature, vol. 46, no. 2, 2008, pp. 285–332.These are articles focusing on the economic consequence of legal origins.

[40] Card, D. and Krueger, A., 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. The American Economic Review, 84(4), pp.772-793.

[41] Paul J Eliason, Benjamin Heebsh, Ryan C McDevitt, James W Roberts, How Acquisitions Affect Firm Behavior and Performance: Evidence from the Dialysis Industry, The Quarterly Journal of Economics, Volume 135, Issue 1, (2020), Pages 221–267, https://doi.org/10.1093/qje/qjz034

[42] The Use of Violence in Illegal Markets: Evidence from Mahogany Trade in the Brazilian Amazon (with Ariaster B. Chimeli). American Economic Journal: Applied Economics, 9(4), October 2017, 30-57.

[43] Peukert, Christian & Bechtold, Stefan & Batikas, Michail & Kretschmer, Tobias. (2020). European Privacy Law and Global Markets for Data. SSRN Electronic Journal. 10.2139/ssrn.3560392.

[44] Goldberg, Samuel, Garrett Johnson, and Scott Shriver. 2021. "Regulating Privacy Online: An Economic Evaluation of the GDPR." Available at SSRN 3421731.

the factors conditioning social platforms to moderate content, there is a lack of empirical works describing and modeling platform behavior.

Part of the theoretical literature is concerned about identifying what each platform will adopt on their own as the appropriate level of content moderation. Platforms have incentives to self-moderate their content, since content such as harassment and hate speech might drive users away. Platform revenue models are often based on advertising and ad buyers normally do not want their brand associated with inappropriate content. Madio and Quinn[45] proposed a model that incorporates brand safety concerns and the role of platforms in moderating user-generated content. Liu et al.[46] analyzed how moderation technology and revenue system (subscription or advertising) affect the incentives platforms already must moderate.

A few works observe how distinctions in liability models might affect platform moderation strategies. De Chiara et al[47] analyze the incentives of platforms to remove content after a notification from a copyright holder. Hua and Spier[48] describe the optimal levels of platform liability for platforms that host harmful companies in their pool of clients, and Jeon atal.[49] observe how content screening is affected by the different level of liabilities that subject such companies.

These theoretical models assume that some level of content moderation will already happen regardless of intermediary liability, and that if somehow the law changes, then the new equilibrium can harm not only those companies but also consumers if the new regime stimulates companies to over-moderate. As previously explained, the alternative for content moderation isalso an open question, as artificial intelligence might not be the most efficient method to curb some forms of illegal content such as hate speech, especially if moderation is

[45] Madio, Leonardo and Quinn, "Martin, Content moderation and advertising in social media platforms" (2021). Available at SSRN: https://ssrn.com/abstract=3551103 or http://dx.doi.org/10.2139/ssrn.3551103

[46] Liu, Yi and Yildirim, Pinar and Zhang, Z. John, Implications of Revenue Models and Technology for Content Moderation Strategies (November 23, 2021). Available at SSRN: https://ssrn.com/abstract=3969938 or http://dx.doi.org/10.2139/ssrn.3969938

[47] De, C., Alessandro, E. M., Antoni, R.-P., & Adrian, S.-M. (2021). "Efficient copyright filters for online hosting platforms". NET Institute Working Paper.

[48] Hua, Xinyu & Spier, Kathryn. (2021). Holding Platforms Liable. SSRN Electronic Journal. 10.2139/ssrn.3985066.

[49] Doh-Shin Jeon & Yassine Lefouili & Leonardo Madio, (2021). "Platform Liability and Innovation," Working Papers 21-05, NET Institute.

focused on content rather than the user sub-networks and their intricate dynamics[50]. Jimenes-Duaran[51] showed through an experiment that users posting hate speech were not deterred by content removal.

A risk associated with of over regulation, either by the government or by the terms of use established by each platform, is that the user might be subject to a chilling effect. That is, they will interact less (e.g. they will be inclined to reduce their posts and comments) with the platform if they perceive a potential risk of retaliation when their conduct is deemed inadequate. This is obviously the goal of moderation for users who disseminate harmful and illegal content, but the chilling effect extends well beyond abusive users and hurts legitimate expression. Observing users that received automated U.S. Digital Millennium Copyright Right (DMCA) notices, Matias et al.[52] showed that they interacted less with the platform after receiving automated notifications. This result is also observed by Penney[53] after surveying users to identify the profile of people more susceptible to chilling effects.

In addition to chilling effects, there is also a more recent literature discussing the value to consumers of access to social networks. Some platforms have incentives to charge users zero price because they operate in a two-sided or multisided market[54], in which advertising "subsidizes" services to the user side. There is a risk of ignoring the fact that these consumers can also be affected by regulation even if they consume a service at zero price.

Using the concept of willingness to pay - how much a consumer is willing to pay for a product - and willingness to accept - how much the consumer must be paid to provide a product - recent papers have shown that users obtain benefits by consuming social media since they

[50] Bharath Ganesh, The Ungovernability of Digital Hate Culture (2018), 71(2), Journal of International Affairs.
[51] Jiménez Durán, Rafael, The Economics of Content Moderation: Theory and Experimental Evidence from Hate Speech on Twitter (2022). Available at SSRN: https://ssrn.com/abstract=4044098 or http://dx.doi.org/10.2139/ssrn.4044098
[52] Matias, J. N., Mou, M. E., Penney, J., & Klein, M. (2020). Do Automated Legal Threats Reduce Freedom of Expression Online? Preliminary Results from a Natural Experiment. https://osf.io/nc7e2/
[53] Penney, J. W. (2019). Privacy and Legal Automation: The DMCA as a Case Study. Stanford Tech. L. R., 22, 412.
[54] Geoffrey Parker, Marshall W. Van Alstyne, Sangeet Paul Choudary, Platform Revolution: How Networked Markets Are Transforming the Economy and How to Make Them Work for You (W. W. Norton & Company, 2016).

have positive values of WTP and WTA. The works of Corrigan et al.[55], Brynjolfsson et al.[56], Mosquera et al. (2019)[57] and Alcott (2020)[58], discussed in more detail in another section of this study, all show that consumers attribute a positive price to the use of social media. As such, a change in regulation that reduces consumer access to social media (e.g. by reducing the content available to consumers in social media) must account for this impact.

The recent economic literature on content moderation is still developing. Nevertheless, some characteristics are clear. On the company side, platforms already have incentives to moderate their content regardless of regulation. Furthermore, a change in the liability standard will alter the equilibrium potentially affecting the welfare of users susceptible to over moderation.

[55] Corrigan JR, Alhabash S, Rousu M, Cash SB (2018) How much is social media worth? Estimating the value of Facebook by paying users to stop using it. PLoS ONE 13(12): e0207101. https://doi.org/10.1371/journal.pone.0207101

[56] Brynjolfsson, Erik and Collis, Avinash and Eggers, Felix, Using Massive Online Choice Experiments to Measure Changes in Well-Being (March 26, 2019). Proceedings of the National Academy of Sciences, 116 (15) 7250-7255, April 2019., Available at SSRN: https://ssrn.com/abstract=3163559 or http://dx.doi.org/10.2139/ssrn.3163559

[57] Mosquera, R., Odunowo, M., McNamara, T. et al. The economic effects of Facebook. Exp Econ 23, 575–602 (2020). https://doi.org/10.1007/s10683-019-09625-y

[58] Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." American Economic Review, 110 (3): 629-76.

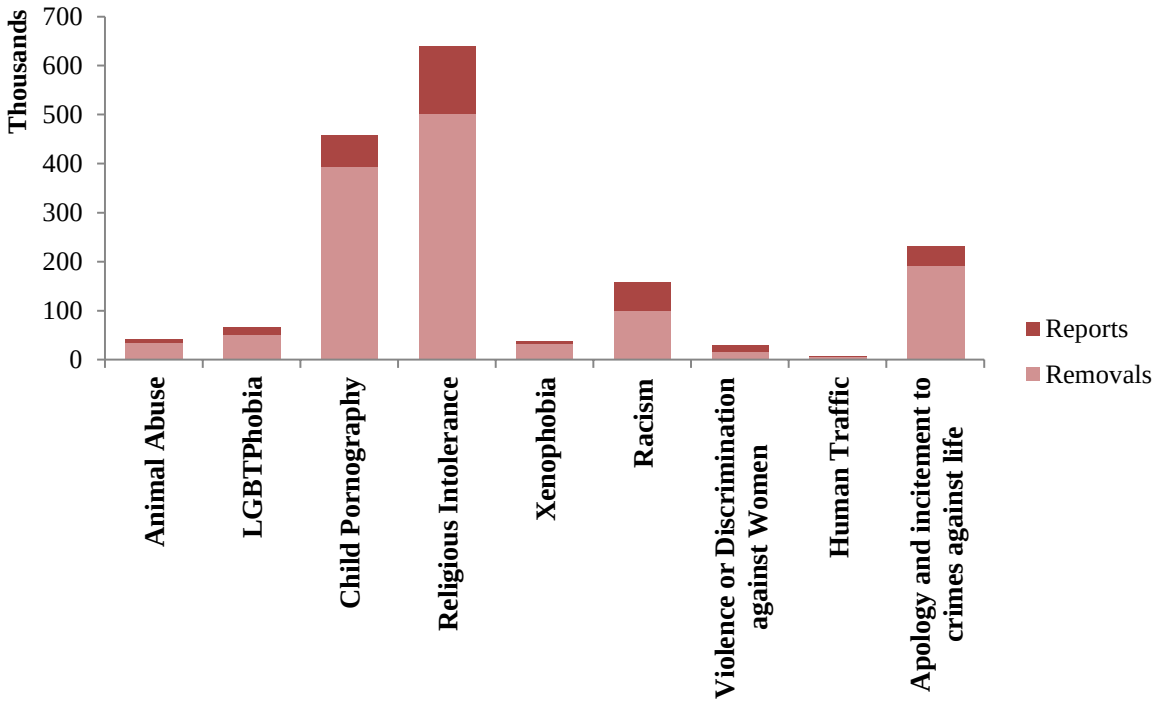
4. Methodology and Results

We apply three different methodologies to estimate the economic effects of digital platforms liability due to user-published content,including company financial impact of social media liability through a change in standards. We present each of the methodologies and their results in the following subsections.

4.1 Reports and removals

Data obtained from Safernet, a non-governmental organization "focused on the promotion and defense of Human Rights on the Internet in Brazil" (Safernet.org), shows that among the social networks present in the ranking of 10 sites with the most complaints and removals of profiles from 2006 to 2021, more than 80% of the complaints received resulted in the exclusion of profiles by the platform. The ranking includes social networks with the highest number of users in each period, such as Orkut (until 2014), Facebook, Twitter, Instagram, Youtube and Tiktok. Safernet data also contains the distribution of reports and exclusion of profiles by report motives, as summarized by **Figure 1** below.

Figure 1 – Distribution of Reports and Removals by Motive



Source: Safernet

Based on the annual variation (2006 – 2021) of complaints and profile removals on social networks carried out by Safernet, we performed a linear regression using the annual variation of profile removal as a dependent variable, and the annual variation of complaints received as an independent variable.

A linear regression is a mathematical model that measures the correlation between two variables[59]. In this case, we are measuring the correlation between the variation in the removal of posts by social networks and the variation in the number of complaints received.

[59] For more information, see Wooldridge, J. M., Econometric Analysis of Cross Section and Panel Data: The MIT Press, 2012

The regression shows the variation in the number of profiles that each social network excludes per year (dependent variable Y_{it}) being explained by the variation in the number of reports that each social network receives (independent variable X_{it}). In this case, the linear regression takes the functional form $\Delta Y_{it} = \alpha + \beta \Delta X_{it} + \epsilon_{it}$. **Figure 2** below shows the results we have obtained from ordinary least squares estimation.

Figure 2 – Panel regression output

post removal variation	0.850**
	(0.040)
Constant	-341.845
	(316.53)
Observations	48
R^2	0.909

Standard errors statistics in parentheses
* $p < 0.05$, ** $p < 0.01$

Source: Authors' calculations

The results show statistically significant coefficients, which means that we can say with 95% confidence that the correlation between the variation in removals and variation in complaints is different from zero. In addition, we obtain a coefficient of determination of almost 90%. The coefficient of determination is a statistical tool that measures the proportion of the variation in the dependent variable that is predictable based on the variation in the independent variable. In the case of removal of posts on social networks, these results indicate that the variation in the number of complaints is significant for the behavior of removal of posts on social networks.

This base model is used with cross-section data. When considering data at different points in time, we can use a regression with panel data. In the case of our model, we performed a panel regression that explains the evolution of the removal/reporting ratio before and after the Civil

Rights Framework for the Internet (Marco Civil da Internet), controlling for the fixed effects of the different social networks.

As previously discussed, Marco Civil instituted in 2014 a judicial notice and takedown liability regime where social networks are afforded greater legal certainty regarding the responsibility for the removal of posts. Thus, we are performing a regression considering the period before and after the statute.

Our regression, from panel data[60], is specified as: $\left(\frac{Removal}{Reporting}\right)_{it} = \alpha + \beta_0(\text{Marco Civil}) + \beta_1(\text{Instagram}) + \beta_2(\text{Orkut}) + \beta_3(\text{Reddit}) + \beta_4(\text{Tiktok}) + \beta_5(\text{Twitter}) + \beta_6(\text{Youtube}) + \varepsilon_{it}$, where "mc" is the binary variable for Marco Civil da Internet being in force. Based on this model, we obtain the following result.

Figure 3 – Panel regression output

Marco Civil	-0.175**
	(0.036)
Instagram	0.623**
	(0.053)
Orkut	0.464**
	(0.055)
Reddit	0.606**
	(0.109)
Tiktok	0.541**
	(0.109)
Twitter	-0.046
	(0.047)
Youtube	0.458**
	(0.044)
Constant	0.536**
	(0.042)
Observations	51
R^2	0.896
FE of Social Networks	YES

Standard errors statistics in parentheses
* $p < 0.05$, ** $p < 0.01$

[60] For more information, see Wooldridge, J. M., Econometric Analysis of Cross Section and Panel Data: The MIT Press, 2012

The outcomes not only show statistically significant results at 95% confidence level but also obtain a high coefficient of determination (R^2), which is almost 90%.

Kohli and Jaworski[61] define market orientation as the dissemination of information throughout the organization and the appropriate responses related to the needs and preferences of customers and competition. Kumar[62] adds to this framework that an organization's market orientation position leads the organization to better performance because management and employees have information about implicit and expressed needs of customers and competitors, as well as strengths and a strong motivation to achieve superior customers satisfaction. Narver and Slater[63] separate Market Orientation into three elements, namely Customer Orientation, Competitor Orientation and Interfunctional Coordination. The results obtained in the regression for the variation in the number of removals of posts on social networks suggest that social networks guide their content removal policy based on the behavior of the volume of complaints, indicating that this may be a component of the Customer Orientation policy of these companies.

These results corroborate the results obtained by Kumar and Narver & Slater[64] regarding the Theory of Market Orientation. Kumar shows that companies which maintain an orientation centered on knowledge, that is, developed on their customer ecosystem (i.e., market orientation), obtain sustainable competitive advantage. Narver and Slater also show that the

[61] Kohli, Ajay K. and Bernard J. Jaworski (1990), "Market Orientation: The Construct, Research Propositions, and Managerial Implications," Journal of Marketing, 54 (April), 1–18.
[62] Kumar, V., Venkatesan R. and Robert P. Leone (2011), "Is Market Orientation a Source of Sustainable Competitive Advantageor Simply the Cost of Competing?," Journal of Marketing Research, 75 (Jan), 16-30.
[63]Narver and Stanley F. Slater (1990), "The Effect of a Market Orientation on Business Profitability," Journal of Marketing, 54 (October), 20–35

[64]Narver and Stanley F. Slater (1990), "The Effect of a Market Orientation on Business Profitability," Journal of Marketing, 54 (October), 20–35

adoption of market orientation yields greater economic profits for companies[65]. Under the hypothesis that social networks adopt, at some level, a policy of market orientation, which implies greater profitability (Narver and Slater) and competitive advantages (Kumar), we would expect to see an efficient policy to remove flagged profiles on social networks. Efficiency in this scenario goes far beyond mere accuracy in decisions regarding the application of community guidelines and substantive legal standards for expression, which includes predictions on possible litigation, their outcomes, and costs.

The adoption of an efficient profile removal policy is also in line with results obtained by Stahl, Heitmann, Lehmann, and Neslin[66]. This study shows that brand positioning has statistically relevant results on customer acquisition and retention, and on companies' profit margins. A policy that does not remove profiles that disseminate harmful content would have a high potential to hurt the brand positioning of that social network and, consequently, deteriorate platform's financials.

The empirical and theoretical results above suggest not only that social networks already carry out a moderation policy that includes restrictions on content and users, but also that they have strong enough private motivations to carry out the exclusion of abusive profiles in a system that they find effective.

Safernet data shows that although social networks exclude profiles that violate their rules based on private notice and user flags, almost 20% of reports do not result in profile deletion. A stricter intermediary liability regime could jeopardize the balance in private policy and force companies to increase the share of profile deletion.

[65] Narver and Slater (1990) define market orientation as "the organization culture that most effectively and efficiently creates the necessary behaviors for the creation of superior value for buyers and, thus, continuous superior performance for the business".
[66]Florian Stahl, Mark Heitmann, Donald R. Lehmann, & Scott A. Neslin (2012), "T The Impact of Brand Equity on Customer Acquisition, Retention, and Profit Margin," Journal of Marketing, 76 (July), 44–63

We take a closer look at the case of Facebook, the platform with the largest number of users in the world. It had a market value of over 450 billion dollars in May 2022, and the Brazilian market represents close to 5% of its users worldwide. A different liability standard could stimulate the platform to ban even more users to avoid expensive litigation, jeopardizing the company's operation with a market value equivalent to more than one-fourth of Brazil's GDP, without necessarily bringing any economic improvement, as discussed above.

We take Facebook's annual revenue per user as a proxy for the entire social media market and assume that the share of the annual revenue per user relative to Brazil is proportional to the quotient of the GDP per capita of Brazil in relation to the GDP per capita worldwide. From Facebook transparency reports, we obtain the total content removed by Facebook in the world. We estimate the total content removed by Facebook in Brazil by multiplying the total content removed by Facebook in the world by the percentage of Facebook users in Brazil (in relation to the total worldwide). This assumes uniform moderation across countries and regions worldwide, which is unlikely. However, platform usually do not disclose country-specific data on moderation practices. Therefore, we move further assuming moderation in Brazil is better proxied by the world average. We divide the estimate of content removed by Facebook in Brazil by the average annual number of posts per user in Brazil, and obtain an estimate for the number of profiles for posts deleted in Brazil. We divide this last estimate by the proportion of excluded profiles from Safernet database and get the estimated reported profiles on Facebook Brazil. Safernet data only represents a fraction of profile reports, but is crucial to estimate the effects of action taken as a result of profile reports.

To obtain the estimate of profiles not removed by Facebook Brazil, we multiply the estimate for the number of profiles for posts deleted in Brazil by 0.2, based on Safernet's deletion quotient, and obtain the estimate of profiles not removed (even after reporting) for Facebook in Brazil. We multiply this last result by the average value per Facebook user in Brazil and divide this result by Facebook's Brazilian market share, finally arriving at an estimated figure for annual lost revenues. Similarly, we multiply the proportion of non-excluded users (in relation to total users) by the sum of the social media valuation to obtain the valuation loss estimate.

This would mean that the social media market revenue in Brazil is approximately 4.5 billion reais per year. If regulation were to encourage a higher share of profile removals in response to private notice and user flags, then this means the gap between 80 and 100% of bans described above could close. The current 20% of unfounded complaints that do not result in profile removal could turn into 20% of unfounded complaints that, because of the risk created by liability, end up causing profile removal. This would imply, given the estimated revenue per user in Brazil, in an annual revenue loss of 56.3 million reais per year for the social media market in Brazil. The detail of this calculation is presented in **Figure 4**.

Figure 4 – Detailed Annual Revenue Loss Estimate

Total Content Removed Facebook (World)	4,275,899,904
Estimated Content Removed (Facebook Brazil)	203,614,281
Estimate Removed Profiles by Post (Facebook Brazil)	1,696,786
Reported Profiles Estimate (Facebook Brazil)	2,056,976
Estimate of unremoved profiles (Facebook Brazil)	429,443
Estimate profiles not removed (All platforms Brazil)	1,713,697
Estimated annual revenue loss	**R$ 56,278,081**
Estimated market value loss	**R$ 27,562,871,323**

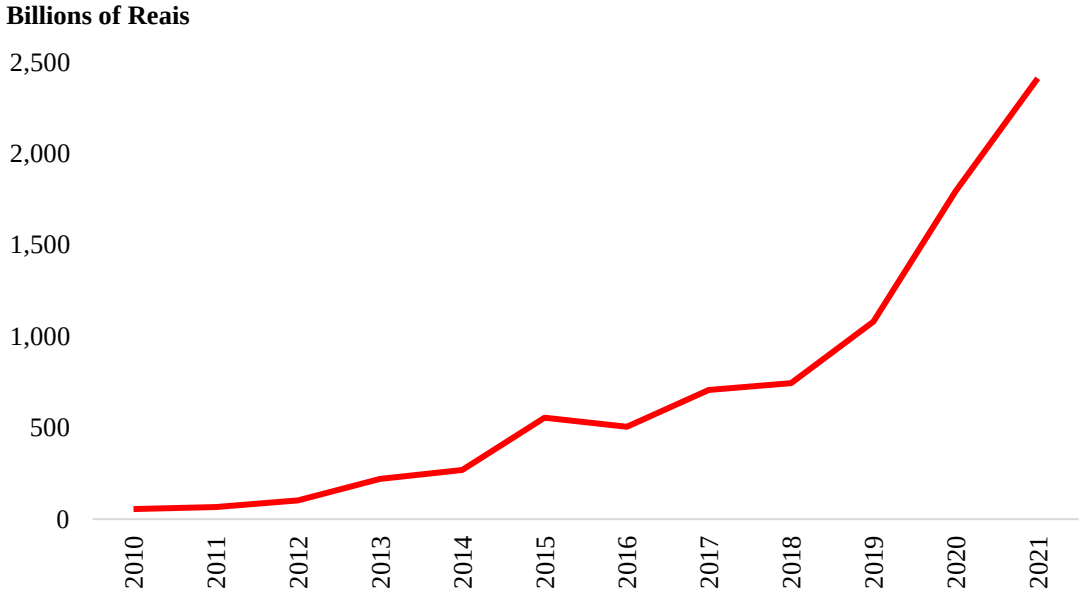
Source: Authors' calculations

These calculations can still be considered as a lower bound, as they do not take into account the possibility that as social networks become more reactive to content and profile reports users and non-users might have a greater incentive to flag more content and profiles – both accurately and inaccurately.

We move further in order to evaluate the market capitalization loss for the social platforms industry. To do so, we change our measure of revenues by the total market capitalization from our benchmark social media, Facebook. We find an estimated market value reduction of R$ 27.6 billion reais due to excess removals in fear of litigation.

To compare the magnitude of the impact, we plot intertemporal platforms' industry market value in Brazil based on publicly traded social networks. This estimate was made considering the world market value of each listed company and, to estimate the share of the market value corresponding to Brazil, we used as a proxy the quotient of Brazil's GDP in relation to the world GDP. **Figures 5** summarizes this data.

Figure 5 - Social Network Market Value in Brazil



Sources: Bloomberg, IPEAdata, Authors' calculations

Now, instead of the hypothesis that companies would "close the gap" and remove all reported profiles, we can estimate the outcome in the case that that companies' removal/reporting behavior returns to that prior to the enactment of Marco Civil. Based on the regression and results show in **Figure 3**, the revenue loss and valuation estimates would be somewhat smaller in magnitude.

Figure 6 – Detailed Annual Revenue Loss Estimate

Total Content Removed Facebook (World)	4,275,899,904
Estimated Content Removed (Facebook Brazil)	203,614,281
Estimate Removed Profiles by Post (Facebook Brazil)	1,696,786
Reported Profiles Estimate (Facebook Brazil)	2,056,976
Estimate of unremoved profiles (Facebook Brazil)	360,191
Estimate profiles not removed (All platforms Brazil)	1,437,344
Estimated annual revenue loss	**R$ 47,202,344**
Estimated market value loss	**R$ 23,118,055,309**

This result implies that, in the event of a regulatory change that encourages companies to change their content moderation policy reverting to pre-Marco Civil risk-aversion behavior, based on Facebook data and assuming this change to be proportional for the entire market (considering market share), we would come to an annual revenue loss of approximately R$ 47.2 million and a loss of valuation of R$ 23.1 billion in the entire social media market.

The estimated social network revenue in Brazil implies an estimated annual revenue of R$ 16.5 billion for the entire market. A PwC study estimates that the internet advertising market in 2021 was R$ 18.7 billion, which would indicate that the social media market in Brazil represents approximately 88% of the internet advertising sector's revenue. It should be noted that our estimate differs from the revenue estimates for the sector made by Statista Company DB, which estimates that the social media market was approximately R$ 7.8 billion, implying that the social media represents only 42% of the online advertising market. If, instead of the revenue estimate presented above, we use the revenue estimates by the Statista Company DB sector, the estimated lost revenue would be R$ 22.2 million per year.

4.2 Impact to Consumers

The interaction between consumers and firms in a market is not a static process, as firms compete with others for the consumer. Each market can be subject to specific shocks such as a merger or acquisition that affects the market or a change in regulation. As such, there are factors that are endogenous to the characteristics of the market, such as how firms operate and compete in it and other factors that might be considered disruptive to the business. Regardless of the source of change in market characteristics, the traditional method to evaluate impact to consumers, apart from quality, is to observe how the change in prices affects consumers.

The immediate effect of a price change is a variation in the quantity of the product that is consumed and if we know how much users value the product then we can estimate the variation in consumer welfare due to price changes. The difference between how much each consumer values a product and how much they pay for it is the consumer surplus. It is possible

to obtain consumer surplus estimating the demand function for a specific product. Therefore, given a price change, with demand curves we can obtain consumer surplus.

The price analysis is not easily implemented when the product is access to a social network because users usually pay zero price for access. Usual methods of demand estimation cannot be applied. The fact that the price is zero does not imply that the consumer benefit is zero. For instance, if consumers were willing to pay for the access, then that value is the economic benefitof the social network to them.

Social network usually operates in at least two markets. In one, they allow users who consume content (one side) to view such content produced by other users (other side). Usually, an additional side in a multisided social media market is comprised of the advertising segment. The cost to add a new user to the existing pool of users is very small and effectively zero. Once a social network already spent on their operational infrastructure (e.g. servers), the maintenance costs (e.g. content moderation system, electricity, legal team) associated with a new user are marginal.

On the other hand, each additional user increases the benefit for other users in either side since there will be more people to communicate and to create content for, there is a network benefit of adding users to platforms[67]. Also, each additional user is another consumer that will be targeted by ads. Social networks have a positive benefit in adding users. Thus, the optimal pricing for social networks is to offer users zero price access and try to include as many users as possible in the platform. Data collection and processing are part of the network effects that propel social media[68].

Platform revenue comes from the (algorithmic) advertising side of the platform[69], but users can still have a welfare benefit even if for accounting purposes like the GDP this benefit does not exist. For the same reason, a change in regulation should consider these non-

[67] Julie E. Cohen, Law for the Platform Economy (2017), 51 U.C.D. L. Rev. 133:144.
[68] Shoshana Zuboff, Big other: surveillance capitalism and the prospects of an information civilization (2015), 30 Journal of Information Technology.
[69] Richard Graham, Google and advertising: digital capitalism in the context of Post-Fordism, the reification of language, and the rise of fake News (2017), 3(45) Palgrave Communications 1.

accountable welfare effects to consumers. To quantify welfare, it is necessary to obtain the value of social network to consumers. A liability regime that causes an excess of content removal and profile exclusions negatively impacts the welfare of social media users. The question then is how to estimate a figure for such impact.

How much a consumer is willing to pay for a product can be associated with different concepts: Willingness to Pay (WTP) is the maximum amount that they are willing to pay to obtain a product. Willingness to Accept (WTA) is the minimum amount that a consumer is willing to be paid to sell or to relinquish access to a service. Under traditional economic theory, those values should be the same, however there is robust evidence documenting that WTA > WTP[70] for the same product. A possible explanation for this is the endowment effect, the fact that individuals tend to place more value in an item they already own.

Recent papers tackling the social media market value adopt experimental methods to obtain WTA from their participants and then obtain consumer surplus by multiplying that by the number of users. Since Facebook is the largest social media in the world, it is not surprisingthat such studies calculate the social media welfare for that platform.

As previously explained, we estimated a loss of 1.7 million Facebook users in Brazil if there is a change to a stricter intermediary liability standard. To estimate the consumer welfare impact, we take the estimated value of Facebook from each related paper in the literature, and calculate how much of the mean personal yearly income[71] consumers are willing to spend on Facebook. This yearly income is from the United States, so we restrict the analysis to articles that the experiment is performed in the United States. Herzong[72] performed an experiment with participants from multiple countries, but we do not include it in this analysis. These studies report median and mean value of Facebook access to the consumer, so we also adopt each of the available data.

[70] Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler. 1991. "Anomalies: The Endowment Effect, Loss Aversion, and Status Quo Bias." Journal of Economic Perspectives, 5 (1): 193-206.
[71] U.S. Census Bureau, Mean Personal Income in the United States [MAPAINUSA646N], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/MAPAINUSA646N, June 29, 2022.
[72] Herzog, Bodo. (2018). Valuation of Digital Platforms: Experimental Evidence for Google and Facebook. Review of Financial Studies. 6. 87. 10.3390/ijfs6040087.

We take both figures in each paper an apply that proportion to the latest yearly income obtained from "PNAD Contínua"[73] for Brazil. This calculation yields how much each Facebookuser in Brazil values access, which we in turn multiply by the 1.7 million users that are estimatedto be lost due to a change in regulation.

Brynjolfsson et al.[74] selected a sample representative of Facebook users in the United States for experiments performed in 2016 and 2017. They report results for a single-binary discrete choice, in which each participant is asked once to forgo access to Facebook for a specific price. These prices are systematically changed across the participants. With these responses they then estimate a demand curve of payment to forgo Facebook for a month and the percentage of participants that decide to forgo. From the demand curve, they obtain a medianWTA of $48.49 per month in 2016 and $37.76 in 2017.

Corrigan et al.[75] conducted a series of second price auctions in which participants bid on the minimum price they were willing to accept to give up Facebook, and the winner of the auction had to receive the amount of the second lowest bid. The participants were an online sample recruited using Amazon's Mechanical Turk platform.

Mosquera et al.[76] conduct an experiment using university students from Texax AM in 2017. They adopted Becker–DeGroot–Marschak (BDM) mechanics in which the participant declares for how much they are willing to accept to give up Facebook. The researcher randomly selects a price and if the price selected is higher than the chosen price then the participant receives the money to deactivate. This experiment had multiple phases and some participants that already received compensation and had given up Facebook for a week were exposed to the

[73] Available at: https://www.ibge.gov.br/estatisticas/sociais/rendimento-despesa-e-consumo/9171-pesquisa-nacional-por-amostra-de-domicilios-continua-mensal.html?=&t=destaques

[74] Brynjolfsson, Erik & Collis, Avinash & Eggers, Felix. (2019). Using massive online choice experiments to measure changes in well-being. Proceedings of the National Academy of Sciences. 116. 201815663. 10.1073/pnas.1815663116.

[75] Corrigan JR, Alhabash S, Rousu M, Cash SB (2018) How much is social media worth? Estimating the value of Facebook by paying users to stop using it. PLoS ONE 13(12): e0207101. https://doi.org/10.1371/journal.pone.0207101

[76] Roberto Mosquera & Mofioluwasademi Odunowo & Trent McNamara & Xiongfei Guo & Ragan Petrie, 2020. "The economic effects of Facebook," Experimental Economics, Springer;Economic Science Association, vol. 23(2), pages 575-602, June.

BDM mechanism a second time to accurately estimate the affects after they already had given up a first time. Participants had a median value of 40 dollars per week and an average value of 67 dollars per week.

Allcott et al.[77] constructed a sample of online users that were paid to participate in the exam and at the end of the experiment the results were made to match Facebook on observables. They used BDM in tree periods during this mechanism for the participant sample and eventually obtained a median WTP of 100 and average of 180. Another interesting result from the paper is that they elicit WTA three times.

The results clearly diverge between these studies. One possible explanation is the difference in samples. Some studies tried to replicate the population that used Facebook in the United States, while others obtained samples in multi-phase experiments including periods where the individuals were already not accessing their Facebook account. Either way, we can observe a large impact in lost consumer welfare that will not be accounted for in traditional calculations.

Figure 7 was produced using the values from each paper. Data on yearly income in the United States comes from series MAPAINUSA646N in the federal reserve database and Brazilian data is from "Pnad Contínua" (IBGE). The number of users removed from the platform, was set at 1,696,786. According to this estimate, the lower bound for annual loss in consumer welfare due to a stricter intermediary liability standard is R$ 532 million and the upper bound is R$ 4.1 billion.

Figure 7 – Facebook Consumer Impact

Article	Corrigan et al. (2018)	Brynjolfsson et al. (2018)	Brynjolfsson et al. (2018)	Mosquera et al. (2020)	Mosquera et al. (2020)	Allcott et al. (2020)	Allcott et al. (2020)
Year of the Sample in the Study	2015 (Auction 3)	2016	2017	2017	2017	2018	2018
Value	US$ 1,921 per year	US$ 48.49 per month	US$ 37.76 per month	US$ 40 per week	US$ 67 per week	100 per month	180 per month
Measure	Mean	Median	Median	Median	Mean	Median	Mean
Percentage of yearly income	4.32%	1.31%	1.02%	4.67%	7.83%	2.38%	4.28%
Impact due to removal of users	R$ 2,257,574,600	R$ 683,830,040	R$ 532,510,256	R$ 2,444,432,674	R$ 4,094,424,730	R$ 1,245,119,524	R$ 2,241,215,144

[77] Corrigan JR, Alhabash S, Rousu M, Cash SB (2018) How much is social media worth? Estimating the value of Facebook by paying users to stop using it. PLoS ONE 13(12): e0207101. https://doi.org/10.1371/journal.pone.0207101

5. Concluding Remarks

Social networks make up a market that has skyrocketed in economic relevance for Brazil, with growth of more than 2,000% in the last 10 years. In this context, the main question we address in this policy paper is: what are the economic effects of a stricter intermediary liability standard on the decisions a platform makes regarding its content moderation rules and their application?

Empirical findings support a model for how sensitive social networks are to complaints and how they react with user removals. Market orientation and branding literature corroborate results that show that social networks already have incentives to maintain an efficient policy for removing reported content even in the absence of legal obligations.

We apply three different methodologies to estimate the economic effects of digital platforms liability due to user-published content. When using linear and panel regression models to analyze possible scenarios of a regulatory change that would encourage companies to adapt their removal policy, our estimates show large potential losses of, at least, R$ 47 million in revenues per year and R$ 23 billion in market value. Moreover, from the users' perspective, we estimate annual losses in consumer welfare due to a stricter intermediary liability standard between R$ 532 million and R$ 4.1 billion.

In light of a possible change to a stricter intermediary liability standard, where social networks are responsible for user generated content before a court ruling or court order, we conclude there is no evidence that such a change would improve content removal policies or economic welfare. In fact, it could lead to sizable losses for platforms, merchants, and consumers.

Appendix

Synthetic Control: the case of Germany and France

Germany and France have recently adopted laws aimed at combating hate speech on social media. The German law Netzwerkdurchsetzungsgesetz (NetzDG) was passed in June 2017 and came into force in January 2018. It fines social networks with more than 2 million users who do not remove illegal content reported by other users within 24 hours. Similarly, France passed Loi Avia in May 2020, which also requested the 24-hour removal of content deemed illegal. The French Constitutional Council found that the 24-hour content removal disposition violated freedom of expression and was therefore unconstitutional. What remains from the law is the possibility of higher fines. These statutes represent stricter rules for social networks in these countries as they perform content moderation.

The law may change platforms' incentives on the appropriate level of moderation, as the possibility of high fines will likely cause companies to excessively moderate their users. Under these scenarios, the need to trigger the state to remove content is reduced so that the amount of court notice requiring content removal is reduced. On the other hand, stricter laws that facilitate judicialization such as the NetzDG can elevate judicialization if platforms do not effectively follow new takedown guidelines, implying an increase in judicialization. Instead of notice and takedown, a switch to judicial notice and takedown in Brazil has been found to gradually decrease litigation – faster rulings and less appeals – precisely because the boundaries for platform moderation become clear and predictable[78].

To analyze the impacts of the two laws, the synthetic control method (CS) was adopted, which is appropriate for evaluating policy impacts in situations where the number of control groups is reduced. CS can be applied when there is only one treatment group available.

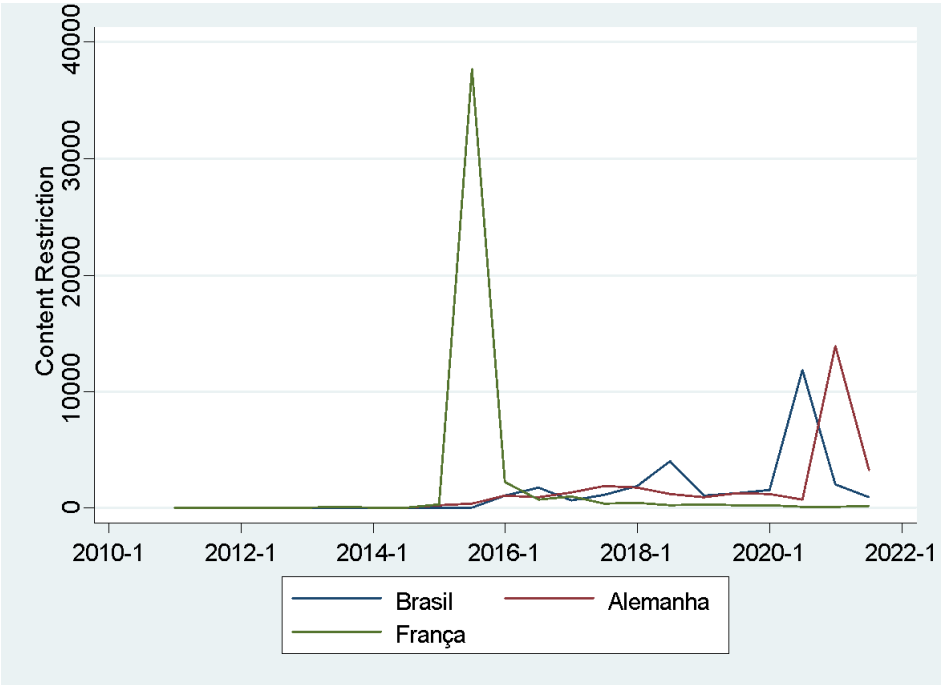
The dependent variables of the analysis are the number of removal requests made by the government obtained from transparency reports by Google, Meta and Twitter. As Google separates which service was affected by the removal order, Google data is separated into YouTube, Google Search and Google Total. Market metric is based on Facebook market share according to StatsCounter data.

For the construction of the model, the following predictor variables were collected from the World Bank: Life Expectancy, GDP per Capita, Proportion of Fixed Broadband Points and Population. All these variables are annual starting from 2011.

Dependent variables are also used as a predictor before the implementation of the respective laws. Facebook's Market Shares have a monthly frequency from 2011 onwards, since the number of removals is published every six months, with the datasets going back to the beginning of the last decade. A limitation of the order base is the presence of high volatility from many outliers.

Thus, for the construction of the model, removal data from the last year before the law was used for the treated countries. For the possible pool of donors, countries that one year and two years later showed variation between semesters of more than 500% were removed. To illustrate the process, we present **Figure A1** below.

Figure A1 - Content Meta restrictions for Germany, Brazil, and France.

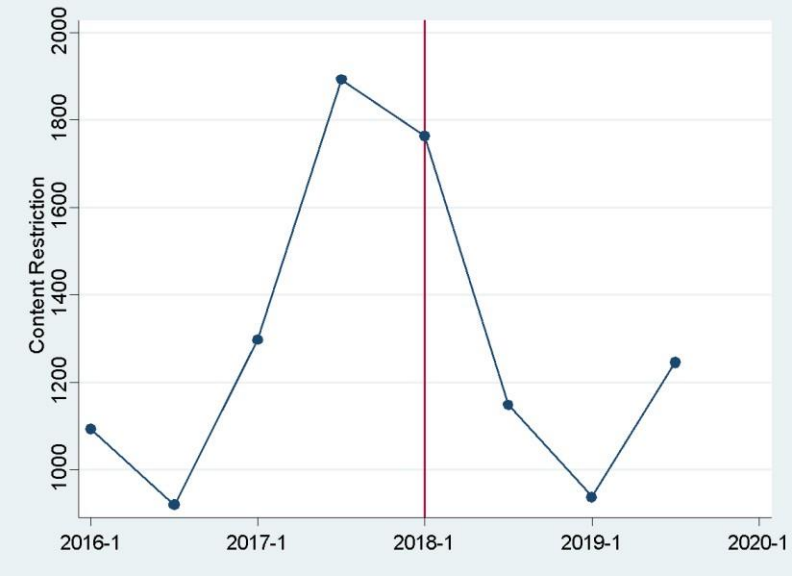


Source: Meta transparency reports.

[78] Marcelo Guedes Nunes, Julio Trecenti. Parecer Acerca Do Impacto Do Marco Civil Da Internet Nas Ações De Remoção De Conteúdo (2021).

It is possible to notice the presence of several outlier data points in the series for the selected countries. Specifically, for France and Germany in a period around the adoption of the laws.

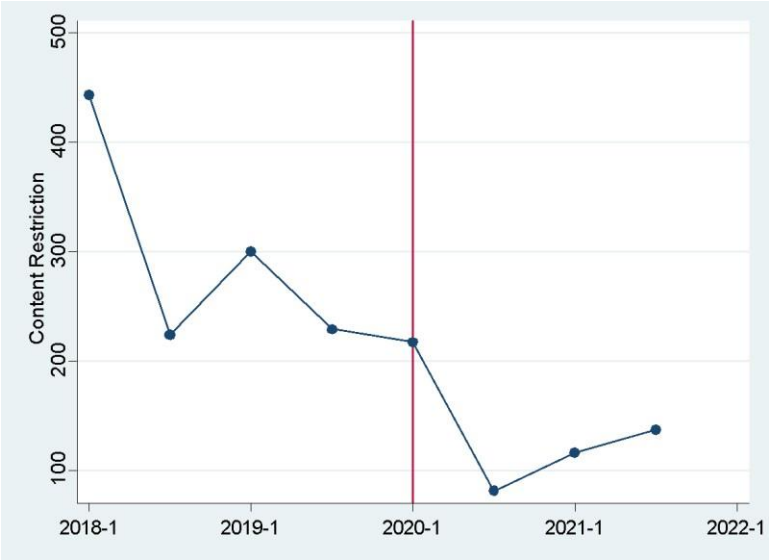
Figure A2 - Content Goal Restrictions for Germany



Source: Meta transparency reports.

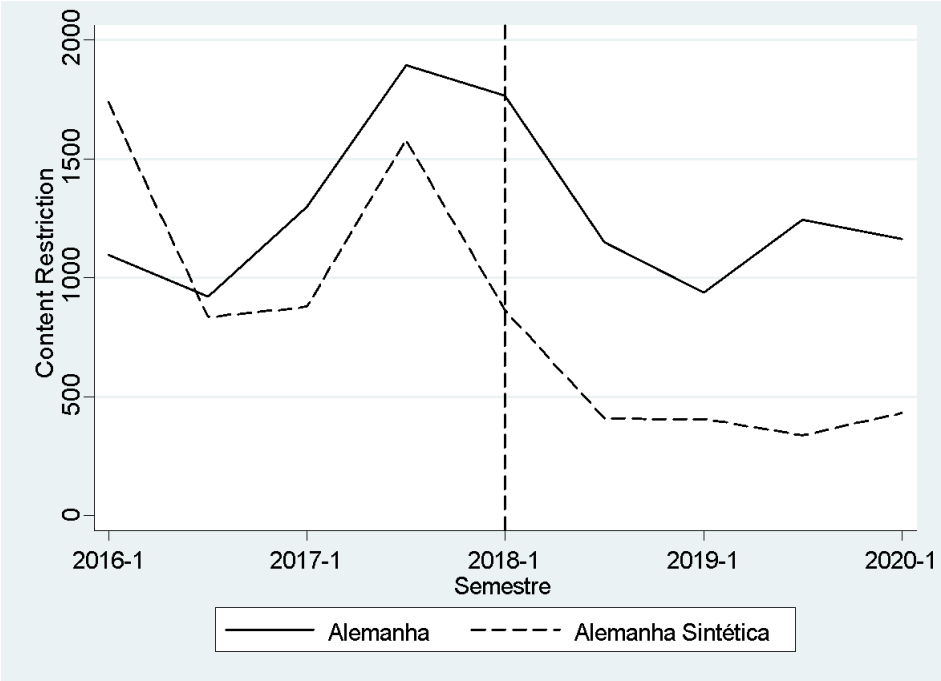
In **Figures A2 and A3**, we note the high dispersion of our constructed dataset. In the range of four years, removals in Germany almost doubled just the semester before the law and for France the restrictions constantly dropped. After running the synthetic control method, we obtain the following series shown in **Figure A4**. Synthetic Germany is made up of 64.8% of France and 35.2% of Turkey.

Figure A3 – Content Goal Restrictions for France



Source: Meta transparency reports.

Figure A4 – Synthetic Control for Target Removal in Germany.



Source: Transparency Report. Law was implemented in January 2018

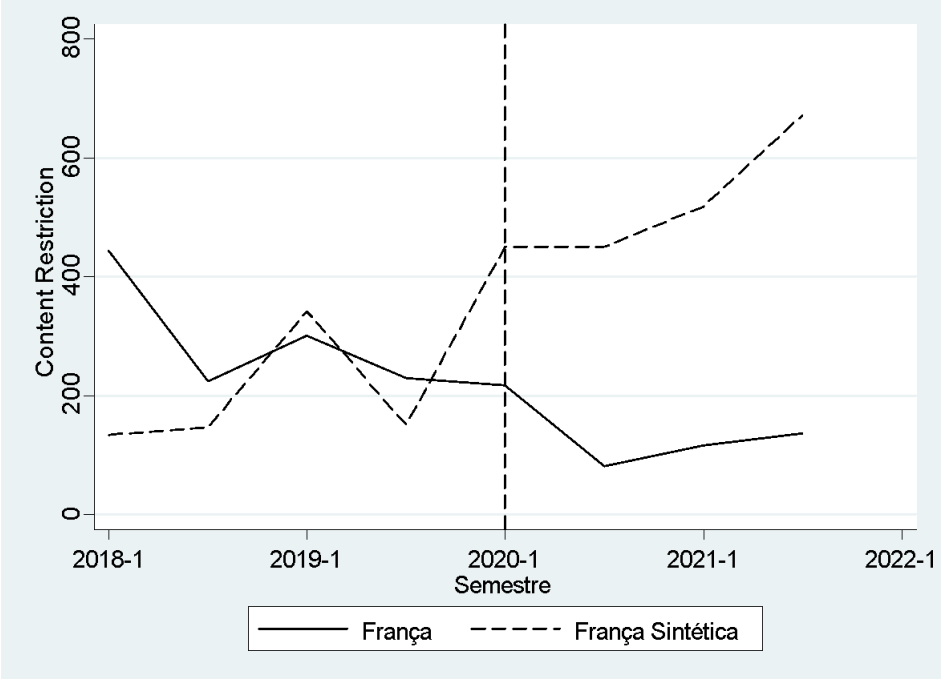
We note in **Figure A5** that we have a reasonably balanced counterfactual before the intervention,except on removal requests.

Figure A5 – Averages Balancing Before the Law

	Original Germany	**Synthetic** Germany
Removal Requests 1 year before the law	1,595	1,227
Fixed Broadband Points per 100 inhabitants	36.62	30.29
GDP per capita log	10.70	10.16
Population Log	18.21	18.06
Life Expectancy	80.74	80.16

Source: Authors' calculation.

Figure A6 – Synthetic Control for Target Removals in France.



Source: Transparency Report. Law was implemented in May 2020.

The CS charts for Germany and France are, respectively, **Figure A4** and **Figure A6**. The effect of the law then derives from the difference between the performed curves and synthetic curves. **Figure A7** summarizes German and French differential point estimates for Meta, Google, and Twitter.

Figure A7 - Averages for dependents two years after the laws

Treated	Synthetic	Treated/Synthetic	Event
1,123	395	284%	Content Removal - Meta Germany
111	546	20%	Content Removal - Meta France
263	280	94%	Content Removal - Google Germany
300	523	57%	Content Removal - Google France
65	66	97%	Market Share - Facebook
69	66	105%	Market Share - Facebook
42	101	41%	Content Removal - Twitter Germany
287	594	48%	Content Removal - Twitter France
145	175	83%	Content Removal - Google Search Germany
253	315	80%	Content Removal - Google Search France
15	16	95%	Content Removal - Youtube Germany

Figure A8 shows that these results are maintained even when we implement placebo tests, which consist of applying synthetic control removing a country that contributed in the weight of the synthetic country. For example, Synthetic Germany for Meta removal is a combination of Turkey and France. So, the placebo consists of applying the method twice, once excluding Turkey from the base and the other excluding France. The table below reports the mean for each event from all placebo trials[80].

[80] Note that by excluding Meta in Germany the laws appear to reduce takedown demands for content removal. It is worth mentioning that the removal results were calculated for periods of two years after the laws since the data presented high volatility and many outliers.

Figure A8 – Placebo Test: Averages Values

Treated	Synthetic	Treated/Synthetic	Event
263	215	123%	Content Removal - Google Germany
300	558	54%	Content Removal - Google France
1,123	615	183%	Content Removal - Meta Germany
111	484	23%	Content Removal - Meta France
65	67	97%	Market Share - Facebook Germany
69	66	105%	Market Share - Facebook France
42	129	32%	Content Removal - Twitter Germany
287	498	57%	Content Removal - Twitter France
145	176	82%	Content Removal - Google Web Search Germany
253	342	74%	Content Removal - Google Web Search France
15	17	89%	Content Removal - YouTube Germany

In this subsection, we conclude that both French and German interventions seem to reduce, instead of increase, removals. In the best scenario, we observed mixed results, where we cannot infer directly the economic benefit of the intervention on either users nor platforms.