RANDOMIZED EXPERIMENTS FOR ONLINE CONTENT MODERATION POLICY

BY
IMANOL RAMÍREZ

Senior Associate at Dentons López-Velarde (Mexico City). Lawyer (Universidad Autónoma del Estado de Hidalgo), LL.M. (Harvard Law School). The article represents only the author's views. Contact: iramirez@llm20.law.harvard.edu and https://www.linkedin.com/in/imanol-ramirez/.

TechREG CHRONICLE
JUNE 2022

THE TAMING OF INTERNET PLATFORMS – A LOOK AT THE EUROPEAN DIGITAL SERVICES ACT
By Natascha Just

WHY TECH REGULATORS NEED TO THINK LIKE GOOGLE
By Terry Flew

WHAT WE TALK ABOUT WHEN WE TALK ABOUT DEPLATFORMING TRUMP
By Scott Babwah Brennen & Matt Perault

AESTHETICS, TECHNOLOGY, AND REGULATIONS
By Gregory Day

RANDOMIZED EXPERIMENTS FOR ONLINE CONTENT MODERATION POLICY
By Imanol Ramírez

PLATFORM CONTENT STANDARDIZATION
By Marta Cantero Gamito

Visit **www.competitionpolicyinternational.com** for access to these articles and more!

RANDOMIZED EXPERIMENTS FOR ONLINE CONTENT MODERATION POLICY

By Imanol Ramírez

The difficulty of achieving consensuses over the regulation of online content moderation has created a stringent and divergent regulatory framework around the world. This fragmentation increases the cost of operating in digital global markets due to greater entry and expansion barriers. Given that countries' legal standards over the regulation of content moderation remain too far apart from each other, international law does not seem to offer a solution in this respect. Nonetheless, policymakers and researchers could start taking advantage of the divergent legal environment and the data richness that characterizes the digital economy to test and prove the effects of the different regulations in place. Today, opinions and proposals are based to a great extent on intuitive assumptions or theoretical ideas. Thus, it is necessary to start questioning and testing those ideas through empirical methods. Applying the rationale used in randomized experiments to validate the intuitive assumptions that fill the debate, such as the alleged effects of intermediary liability, including the chilling-effect over speech, would allow policymakers to better understand how the different regimes shape the conduct of intermediaries and make policy decisions accordingly.

01 INTRODUCTION

Much ink has been spilled on how to balance the different policy objectives of online content moderation regulation. These policy goals include preventing online harms, promoting free speech, encouraging technical innovation, and guaranteeing competition, among others.[2] Nonetheless, optimal solutions or consensuses have proven difficult to achieve and today, as a result, there is a fragmented legal and regulatory landscape over online content moderation all over the world.

Although this increasingly stringent and divergent legal environment raises various concerns, especially for innovation and competition policy due to increased entry and expansion barriers to markets and data, the way forward in terms of harmonization and cohesion does not look promising. International law, which could be the right avenue to address issues with global dimensions, does not seem at hand since countries' legal standards remain too far from each other.

For instance, in the debate of intermediary liability/immunity, the United States and now its trading partners under the United States-Mexico-Canada Agreement ("USMCA") maintain broad protections for intermediaries who, with certain exceptions, are not liable for the effects of user-generated content posted in their platforms and of the moderation decisions taken with respect to it. On the other hand, the European Union and its member countries aim to establish liability-based regimes for intermediaries, including significant fines for failure to comply with regulations, as demonstrated by the recent approval of the Digital Services Act ("DSA").

Moreover, regulation proposals tend to be based on intuitive assumptions or even anecdotal evidence, most likely reflecting commentators' values or implicit guesses about the possible effects of regulation. At a theoretical or logical level, there are plenty of compelling arguments supporting many of the views out there, such as the pros and cons of intermediary liability/immunity, which proponents from both sides of the debate have extensively put forward. However, we may be failing to look at the facts and evidence available around these discussions in order to test our theories and ideas through empirical methods.

Legislators and policymakers could flip the script and take advantage of the fragmented regulation across jurisdictions to assess and determine the way forward. Empirical evaluation methods such as randomized experiments have been successfully applied to health and economic development policies, significantly advancing our understanding of how regulation could be designed. Digital markets offer unprecedented and valuable data to that effect which could allow policymakers to understand how different regimes shape the conduct of intermediaries and make policy decisions accordingly.

The following paragraphs elaborate on these ideas as follows: (i) the fragmentation of online content moderation regulation; (ii) moving beyond assumptions and theoretical debates; and (iii) randomized experiments for the design of content moderation policies.

02 FRAGMENTATION OF REGULATION OVER ONLINE CONTENT MODERATION

I have written separately about online content regulation and competition policy, arguing that the push for regulation over content moderation around the world is creating a stringent and divergent legal environment with increased liability for firms operating in the digital landscape, including the mandatory use of technology, the establishment of substantial fines and even criminal responsibility.[3]

We have seen regulations establishing increasingly strict intermediary liability in several countries, including Germany, the United Kingdom, India, Thailand, and Australia. More recently, the European Union reached an agreement on the DSA establishing, among others, the possibility for users to challenge, either judicially or through an out-of-court mechanism, content moderation decisions taken by intermediaries and providing for significant fines in case of noncompliance.[4]

[2] See Joris van Hoboken and Daphne Keller, Design Principles for Intermediary Liability Laws, Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression (Oct. 8, 2019) at 2-3. Available at https://www.ivir.nl/publicaties/download/Intermediary_liability_Oct_2019.pdf.

[3] See Imanol Ramírez, *Online Content Regulation and Competition Policy*, Harvard Law School Antitrust Association, Cambridge, MA, December 3, 2020. Available at https://orgs.law.harvard.edu/antitrust/files/2020/12/Imanol-Ramirez-Online-Content-Regulation-and-Competition-Policy-HLSAntitrustBlog-2020.pdf.

[4] European Commission, *Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment*, April 23, 2022. Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545.

The immediate consequence of the fragmentation of the regulatory landscape is that the cost of operating in the market increases, as companies have to deploy the technology and resources to comply with legal requirements and avoid liability, which may be very costly in some cases.[5] This is particularly relevant considering the economic characteristics of some digital platform markets where legal barriers may hinder an entrant's ability to access large data sets, obtain scale and generate its own positive network effects that are necessary to challenge large incumbents.[6]

In this way, policymakers worldwide face a collective action problem whereby increased liability imposed in each jurisdiction may strengthen the market position of large incumbents operating globally by raising the costs of entry and expansion into digital markets. Although its pending implementation, the European Union's DSA attempts to address this issue by establishing obligations depending on companies' size, role, and impact.[7] Government interventions on content moderation need to use a mix of strategies to take advantage of both market forces and state regulation to effectively tackle online harms, promote free speech, and guarantee competition, among others.

Evidently, international law could set a better framework to face this collective action problem. Nonetheless, countries are not close enough with respect to the legal standards applicable to content moderation, including intermediary liability/immunity and the actual rules governing users' content, i.e. what should be considered illegal content and what should not.

On one side of the spectrum, the United States' Section 230 of the Communications Decency Act of 1996 ("CDA") is generally viewed as landmark legislation granting broad protections for intermediaries with respect to harms arising from user-generated content and the moderation decisions taken by intermediaries to this effect. This more laissez-faire oriented approach to intermediary liability is also reflected in Article 19.7 of the USMCA, which establishes a similar (although not identical)[8] provision to Section 230 containing broad protections for intermediaries.

On the other hand, the European Union and its member countries have been adamant about their attempt to impose a responsibilities-based regime on intermediaries. In Germany, the Network Enforcement Law establishes fines up to €5 million for internet companies with at least 2 million users that fail to remove manifestly unlawful speech within 24 hours and all illegal content within seven days of receiving a complaint, as well as reporting obligations on how complaints are handled.[9] More recently, in April 2022, the DSA approved by the European Union 27-member countries established landmark legislation for intermediary liability, including the possibility for users and civil society to challenge moderation decisions and seek redress, as well as transparency obligations, with fines up to 6% of the companies' global turnover for the failure of compliance.[10]

Just as with historical attempts to create an international antitrust regime,[11] the fact that the United States and the European Union remain too far from each other concerning the applicable standards to regulate online content moderation most likely puts them in a deadlock where the two blocks perceive that the benefits from international laws on content moderation would not exceed its costs.

In this context, it is unlikely that larger agreements at a multilateral or international level will be achieved, and even when international law may be the right avenue to create a coherent regime for issues with global dimensions, international law would not be available for content moderation at least in the short term. Moreover, a universal answer to online content moderation is unlikely, considering that views on freedom of speech vary across the spectrum, including

[5] See Imanol Ramírez, *Online Content Regulation and Competition Policy*, Harvard Law School Antitrust Association, Cambridge, MA, December 3, 2020. Available at https://orgs.law.harvard.edu/antitrust/files/2020/12/Imanol-Ramirez-Online-Content-Regulation-and-Competition-Policy-HLSAntitrustBlog-2020.pdf.

[6] For a brief description of the economics of digital platform markets, see Imanol Ramírez, Merger Thresholds in the Digital Economy, Delaware Journal of Corporate Law,

[7] European Commission, *Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment*, April 23, 2022. Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545.

[8] See Vivek Krishnamurthy and Jessica Fjeld, *CDA 230 Goes North American? Examining the Impacts of the USMCA's Intermediary Liability Provisions in Canada and the United States*, Harvard University - Berkman Klein Center for Internet & Society, July 7, 2020. Available at http://dx.doi.org/10.2139/ssrn.3645462.

[9] Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG), (Federal Law Gazette I, p. 3352 ff. Valid as from 1 October 2017) https://germanlawarchive.iuscomp.org/?p=1245.

[10] European Commission, *Digital Services Act: Commission welcomes political agreement on rules ensuring a safe and accountable online environment*, April 23, 2022. Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_22_2545.

[11] See Anu Bradford, *International Antitrust Negotiations and the False Hope of the WTO*, 48 Harvard International Law Journal, 383, 2007. Available at https://scholarship.law.columbia.edu/faculty_scholarship/561.

for each individual, the leadership within each tech company, and societies as a whole.

03

BEYOND INTUITIVE ASSUMPTIONS AND THEORETICAL DEBATES

Today, there is a lengthy debate around the regulation of online platforms, including moderation of user-generated content. Despite the extent of the discussion, opinions tend to be based on intuitive assumptions or even anecdotal evidence, with little or no support from empirical evidence.

For instance, one of the main ideas behind the rationale of CDA Section 230 that academics, courts, and organizations have persistently put forward argues that increased liability over online intermediaries would generate an over-removal or chilling effect on speech. The argument goes that when facing moderation decisions of third-party content in a stringent liability regime, intermediaries will most likely err on the side of caution to avoid liability and thus systematically censor content that eventually will hurt free speech.

Even when this is a strong logical argument, there is little empirical analysis of the validity of this idea that allows us to really understand the effects of increased regulation on speech. Most studies that have attempted to test this idea have focused on surveys without really measuring how people behave after actual changes in law since they are limited to people's claims on how hypothetical regulatory changes will affect them.[12]

At a theoretical or logical level, we can further develop more arguments about the validity or invalidity of the chilling effect principle. For example, it can be argued that the chilling effect idea fails to recognize the economic incentives that the industry has to maintain content online, which could offset or at least be in direct tension with the possibility of over-removal of speech.

Considering that user-generated content is at the core of some digital businesses and that it is central to profits in many cases, one could argue that it is possible to introduce monetary sanctions for the failure to remove harmful content without having a significant chilling effect on desirable speech. In this way, the underdeterrence concern of existing harmful speech can be weighed in the balance.

A rational company whose objective is to maximize shareholders' profits would balance the potential economic loss derived from a fine against the loss from removing profitable content, which is particularly important in industries with network effects, and will choose the smaller loss. Although other elements should be considered, such as reputational damages and the companies' need for good content, it can be argued that this is one reason why the "chilling-effect" argument, while powerful, is incomplete. The economic incentives of an intermediary will at least counter its incentives for over-removal of speech and could create a mix of incentives that results in better content moderation tools and efforts. Therefore, when facing a decision of potential removal of content, it should not be taken as a given that intermediaries will err on the side of caution to avoid fines.

This is even more compelling in cases where it is easy for intermediaries to identify harmful content and there is not much discussion about its adverse effects. Establishing fines for failing to eliminate the easy cases could suppress any incentives of companies to leave harmful content online that is highly profitable. It is also likely that a positive correlation exists between the degree of harm and the obviousness of harmful content. Thus, monetary penalties may be more effective in preventing and eliminating harmful content in these cases.

But at this point of the debate, shouldn't we move beyond these intuitive assumptions and, if possible, question the reality and strength of these ideas and other principles at an empirical level? Although it may be complex, data is at the heart of digital markets, and business models and platforms collect and store unprecedented information that could be used to measure the effects of the different policies in place.[13] Nonetheless, numerous studies show that the data collected is never used, signaling that incentives to gather it are there but not the incentives to analyze it.[14] To the extent possible, the ideas and assumptions over content moderation regulation should be accompanied by empirical methods to increasingly employ data generated by firms and be able to test and confirm the intended effects of regulation.

[12] See Suneal Bedi, *The Myth of the Chilling Effect*, Harvard Journal of Law & Technology, Volume 35, Number 1 Fall 2021. Available at https://jolt.law.harvard.edu/assets/articlePDFs/v35/Bedi-The-Myth-of-the-Chilling-Effect.pdf.

[13] Daniel Björkegren and Chiara Farronato, *To Regulate Network-Based Platforms, Look at Their Data*, Harvard Business Review, October 18, 2021. Available at https://hbr.org/2021/10/network-based-platforms-must-be-regulated-but-how.

[14] Viktor Mayer-Schönberger and Thomas Ramge, *The Data Boom Is Here — It's Just Not Evenly Distributed*, the MIT Sloan Management Review, Spring 2022, February 9, 2002. Available at https://sloanreview.mit.edu/article/the-data-boom-is-here-its-just-not-evenly-distributed/.

04 RANDOMIZED EXPERIMENTS

Given these circumstances, how can we make something good out of the increasingly fragmented regulatory framework for online content moderation and the data richness that characterizes the industry? Due to the somewhat accidental disagreement among the international community on the regulation of content moderation, policymakers could imagine themselves in a position where only one global content moderation program exists, but some of its features are varied across jurisdictions, in a situation similar to randomized experiments.

Following the logic of randomized experiments or randomized controlled trials, which have been successfully employed for development economics and evaluation of public policies,[15] researchers could take advantage of the divergent regulatory framework to implement large-scale experiments that test the theories underlying the different legal regimes. Divergent regulation across jurisdictions will have the effect of assigning online intermediaries that operate internationally to different potential solutions. Thus, the effects in the conduct of an intermediary could be attributed to the differences in the regulation. The goal of these studies should be to understand how to better design content moderation regulation.

For instance, these empirical studies could shed some light on the intermediary liability/immunity debate. It may be possible to understand how moderation decisions of online intermediaries have been shaped by regulations that establish fines for failure to comply with the removal of harmful content, and how those decisions compare to other regimes that provide for a more laissez-faire approach. Moreover, comparisons among different intermediaries with divergent moderation policies but acting under the same legal framework could also allow policymakers to understand the effects of the rules governing users' content (i.e. companies' internal moderation policies). To achieve the preceding, regulators will require information such as the amount and type of material being taken down in each of the jurisdictions being studied, the reasons why a certain type of content is being censored or not, the results of flagged content or complaints, the differences in technology being deployed and investments, including in human resources and programs, among others.

To this effect, it should not be expected that online intermediaries provide their sensitive data voluntarily, especially proprietary information and trade secrets arising from their own data analysis. Simultaneously to the ongoing efforts, governments should consider increased transparency obligations and data access mandates for online intermediaries. Regulators would have to meet very high standards of confidentiality and protection of the information to guarantee the safety of the platform and that its competitiveness is not jeopardized, as well as to invest in teams that can analyze raw data provided by companies.

The randomization-like approach would not offer definitive answers or ultimate solutions to content moderation problems as every study group (in this case, the same companies acting in different jurisdictions) possesses intrinsic elements that would shape moderation decisions. However, no study is likely to demonstrate causality on its own. One of the main features of randomization is that it sets aside the observed and unobserved characteristics of the different study groups allowing attribution of any differences in the outcome to the intervention and thus measuring to some extent its effectiveness.[16]

Applying the rationale behind randomized experiments will help to understand the effects of the different policy ideas and theories in place without solely relying on intuitive assumptions of alleged impacts of regulation. This may bring countries' standards closer and thus real possibilities of larger consensuses that further promote innovation and competition in the digital economy. ■

> *"Given these circumstances, how can we make something good out of the increasingly fragmented regulatory framework for online content moderation and the data richness that characterizes the industry?*

[15] See Arthur Jatteau, *The Success of Randomized Controlled Trials: A Sociographical Study of the Rise of J-PAL to Scientific Excellence and Influence*, Historical Social Research, 43, no. 3, 165, 2018: 94–119. Available at https://www.jstor.org/stable/26491530.

[16] Eduardo Hariton and Joseph Locascio, *Randomised controlled trials - the gold standard for effectiveness research*, Study design: randomised controlled trials. BJOG: an international journal of obstetrics and gynaecology, 125(13), 1716, December 2018. Available at https://doi.org/10.1111/1471-0528.15199.