13

The Disinformation Paradox

Why Regulating Online Content at Home May Make Matters Worse in the World

Bhaskar Chakravorti

13.1 INTRODUCTION

Proponents of social media and digital platforms more broadly point to the idea of societal well-being advanced through interconnectivity and collective access to vast troves of user-generated content. Individual users can discover content – ranging from deep wisdom to shallow banalities to everything else in between. Each participant on a platform can draw upon insights from anywhere in the world; conversely, in principle, a single user's voice can reach well beyond their immediate circles. There are so many striking examples of such collective sharing of content, from the experience of humanity communing digitally through a shared period of physical isolation during the COVID-19 pandemic to receiving news and updates about unfolding events, wars, elections, or sports. Our IDEA 2030 research by The Fletcher School's Digital Planet program, in collaboration with Equiception, a research firm, conducted an analysis of content on digital platforms during the first seven months of 2020 and parsed the emotional states of citizens in messages posted on digital platforms across eight democracies during an unprecedented time when many democratic civil liberties were suspended. The research revealed a surprising similarity of sentiment, primarily of positivity – presumably in solidarity during a period of collective adversity. This shared experience may have helped many get through one of the most extraordinary periods of uncertainty and anxiety for citizens everywhere of the kind that few had experienced in their lifetimes.[1]

Digital platforms had been designed with the objective of sharing in mind. The founder and CEO of Meta (formerly Facebook), Mark Zuckerberg, famously wrote in the opening of his 2012 Founder's Letter: "Facebook was not originally created to

[1] Auret van Heerden, Bhaskar Chakravorti, Ravi Chaturvedi & Ravi Sreenath, *Inclusive Crisis Management by Governments: Using Digital Ethnography and Sentiment Analysis as a Sensing Function and Policy Tool*, IDEA 2030, Digital Planet, The Fletcher School, Tufts University (Oct. 2021), https://sites.tufts.edu/digitalplanet/files/2021/Sentiment-Analysis-During-the-Pandemic.pdf (last visited Feb. 18, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

be a company. It was built to accomplish a social mission – to make the world more open and connected." The high-minded vision and the imperatives of running a company aside, it is now clear that the very openness and connectivity of digital platforms can cut both ways. Harmful content – whether it is disinformation or false, misleading, and inaccurate information and news, hate speech, conspiracy theories, and posts planted for purposes of political or commercial gain or inciting violence or with the potential of undermining institutions – represents a classic negative externality as the outcome of the open and connected nature of digital platforms. While we can recognize the social ills of harmful content, its moderation is hard. The reason comes down to a single critical factor: the misalignment of incentives. It is this misalignment that gives rise to a "disinformation paradox," where attempts to regulate harmful content could give rise to even more of it. This chapter explores this phenomenon and possible resolutions.

The chapter is organized as follows. Section 13.2 highlights the role of incentives in inhibiting adequate self-moderation of content by digital platforms. This means that public oversight and regulation is needed as a corrective measure. In Section 13.3, I shall argue that even such public oversight and regulation is fraught with challenges and may give rise to unintended consequences that could result in making the disinformation problem worse across the world. Ironically, once again, incentives are at the root of the problem. Section 13.4 outlines additional factors in play stemming from the dynamic nature of the industry that add to the complications of effective content moderation. The Sections 13.5–6 offer alternative approaches to addressing the problem and their limitations.

13.2 INCENTIVE INCOMPATIBILITY AND THE PROBLEM OF UNDER-MODERATION OF CONTENT BY DIGITAL PLATFORMS

Mark Zuckerberg's company has been in the news repeatedly for its inability to moderate the content that it helps disseminate to every corner of the world. In her October 5, 2021, testimony before Congress, former Facebook employee and whistleblower Frances Haugen made a strong case for public oversight and regulation of digital platforms. She asserted that the products of her former employer (now known as Meta Inc.), "harm children, stoke division and weaken our democracy."[2]

Clearly, Zuckerberg or any of the other founders of the major digital platforms did not intend to get to such an outcome. To understand how the platform he created might have gotten to this point, we must appreciate the disinformation chain responsible for the outcome. The chain originates with the creator of content and ends with the user who views the content at the other end. The originator of the

[2] *Whistleblower Says Facebook's Choices Are "Disastrous" for Children, Democracy*, Wall Street Journal (Oct. 5, 2021), https://www.wsj.com/livecoverage/facebook-whistleblower-frances-haugen-senate-hearing (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

content may have their own reasons for creating it. In the intermediate stages of the chain, there are other users who share, endorse, and comment on the original thereby passing it along to an even larger body of users. In parallel, the platforms and their algorithms and recommender systems determine what content users encounter. At every stage in this chain, in principle, harmful content could be flagged, blocked, or de-prioritized – but it isn't. Instead, it is more likely to be amplified. The reason is that the incentives of participants involved in the disinformation chain are misaligned with appropriately moderating such content.

Users who create disinformation have different incentives – from innocent mischief to political – driving the urge. Similarly, those who share it help propagate it for different reasons; often, it is precisely because disinformation can be edgy and there is an expectation that it will command greater engagement, which is what users want from their posts. In many instances, the users in the chain may not even know that are passing along disinformation. The platforms, for their part, often cite the sheer scale of the problem as their biggest challenge in catching disinformation. In fact, Mark Zuckerberg has noted:[3]

> We have a responsibility to keep people safe on our services. That means deciding what counts as terrorist propaganda, hate speech and more. We continually review our policies with experts, but at our scale we'll always make mistakes and decisions that people disagree with.

No doubt, the difficulties of content moderation at scale are considerable, given that tens of billions of posts – and a billion stories alone across the Meta platforms – are created every day.[4] That said, there is another, more structural reason as to why this is hard: attempts to moderate harmful content would run counter to the platform company's commercial interests and the incentives offered to key employees. Content of any kind leads to user engagement, which, in turn, can be monetized by selling spots to advertisers. Product managers at Facebook, for example, have compensation structures that are tied to getting users to maximize their time and engagement on the platforms. Indeed, these incentives work: platforms, such as those owned by Meta, have been successful at extending the time and engagement by users. Furthermore, both go up when the content is harmful. A study by researchers at New York University and the Université Grenoble Alpes found that disinformation generated six times the volume of engagement as did posts from

[3] Mark Zuckerberg, *The Internet Needs New Rules. Let's Start in These Four Areas*, Washington Post (Mar. 30, 2019), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html (last visited Feb. 19, 2024).

[4] Facebook Earnings Call, Q4, 2018, https://www.facebook.com/business/ads/stories-ad-format# (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

trustworthy news and information sources.[5] In turn, this greater engagement and time spent on the platform pays off for the company because over 97 percent of the company's revenues are derived from advertising, which increases with exposure.[6] Even if there were concerns about the content, procedures at the platforms are optimized to put such concerns aside. For example, user interactions that were once analyzed by engineers are increasingly being analyzed by algorithms that create personalized feedback loops for tweaking and tailoring each user's news feed to keep pushing up their engagement.[7] Simultaneously, hiring content moderation teams costs money and do not generate revenue. During times when business slows, such cost centers are among the ones that get cut. We saw this play out across multiple platforms. The most extreme case was the company formerly known as Twitter, which decimated its content moderation teams – both in-house employees and outside contractors – after the takeover by Elon Musk. As a result, there was a spike in the volume of harmful content on the platform. This also has the effect of driving advertisers and some users away, thereby cutting into revenues, but Twitter – now called X – has not made many changes in response.[8] Indeed, the entire process of how content is circulated on the platform is caught in a vicious cycle of incentives incompatible with appropriately moderating harmful content.

There is, thus, an inherent contradiction in platforms self-moderating content. Self-moderation systems and rules will be incomplete or are likely to be self-serving or ad hoc. This is a classic example of market failure and calls for regulators or lawmakers to step in. However, external oversight and regulation is vulnerable to its own incentive incompatibility challenge.

13.3 INCENTIVE INCOMPATIBILITY AND MODERATION OF CONTENT BY REGULATORS

The chapters in this book authored by Eric Goldman, Cristoph Busch, Jhalak Kakkar et. al, Jufang Wang and Artur Pericles cover the state of content moderation regulation across a wide range of jurisdictions – US, EU, India, China, and Brazil – and the varying states of such regulations across them. As the chapters collectively

[5] Laura Edelson et al., *Understanding Engagement with Us (Mis)Information Sources on Facebook*, Association For Computing Machinery (Nov. 2021), https://lig-membres.imag.fr/gogao/papers/news-interactions-imc2021.pdf (last visited Feb. 19, 2024).

[6] *Meta Reports Fourth Quarter and Full Year 2021 Results*, Meta Investor Relations (2021), https://investor.fb.com/investor-news/press-release-details/2022/Meta-Reports-Fourth-Quarter-and-Full-Year-2021-Results/default.aspx (last visited Feb. 19, 2024).

[7] Karen Hao, *Facebook Whistleblower Hearing: Frances Haugen Testified before Senate Panel*, MIT Technology Review (Mar. 11, 2021), https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-disinformation/ (last visited Feb. 19, 2024).

[8] A. Counts & E. Nakano, *Twitter's Surge in Harmful Content Keeps Advertiser Away*, Time (July 19, 2023), https://time.com/6295711/twitters-hate-content-advertisers/ (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

highlight, there is no harmonization across these jurisdictions, China being the most stringent about its rules and the US the least at the federal level, with the EU as an early mover to put the most holistic set of laws in place to create the guardrails for the digital economy.

Moreover, the state of regulation itself is constantly in flux as the regulators respond to political pressures and platforms respond to the regulatory actions. In some jurisdictions, the regulations and restrictions are being legally challenged by the platforms; the lawsuit by Twitter against the government of India is one such example.[9] In other instances, the government went so far as to block platforms altogether, as was the case in Russia after its invasion of its neighbor, Ukraine.[10] For their part, the platforms themselves have adopted aggressive tactics to build leverage in their negotiations with governments and regulators, knowing fully well how reliant users are on the platforms themselves. This was the case in Australia, in March 2021, when Facebook caused significant havoc by blocking Australian users from news pages to preempt legislation that would make Facebook pay for news content.[11] After five days of chaos in the country, the Australian Parliament caved in.[12]

All of these differences and shifting boundaries of regulatory reach means that a platform that operates in a global market must conform to a highly fragmented set of regulations. The fragmentation is not just because of the lack of standardization across countries. In a jurisdiction, such as the US, the regulations are uneven across states. While there has been little movement at the federal level, several states have taken more activist positions. Consider the law AB 587, enacted in California, on September 14, 2022.[13] The law requires that each social media company above a certain size, "must provide detailed description of content moderation practices used by the social media company for that platform, including how automated content moderation systems enforce terms of service of the social media platform and when these systems involve human review."[14] The penalties for violating the rules appear small when compared to the revenues of the social media companies: the company

[9] K. D. Singh & K. Conger, *Twitter, Challenging Orders to Remove Content, Sues India's Government*, New York Times (July 8, 2022), https://www.nytimes.com/2022/07/05/business/twitter-india-lawsuit.html (last visited Feb. 19, 2024).

[10] D. Milmo, *Russia Blocks Access to Facebook and Twitter*, The Guardian (Mar. 4, 2022), https://www.theguardian.com/world/2022/mar/04/russia-completely-blocks-access-to-facebook-and-twitter (last visited Feb. 19, 2024).

[11] K. Hagey, M. Cherney & J. Horwitz, *Facebook Deliberately Caused Havoc in Australia to Influence New Law, Whistleblowers Say*, Wall Street Journal (May 5, 2022), https://www.wsj.com/articles/facebook-deliberately-caused-havoc-in-australia-to-influence-new-law-whistleblowers-say-11651768302 (last visited Feb. 19, 2024).

[12] *Id.*

[13] Assembly Bill No. 587, California Legislative Information (2021–22), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220AB587 (last visited Feb. 19, 2024).

[14] *Id.*

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

is liable for a civil penalty not to exceed $15,000 per violation per day, but they can add up to a much larger political and regulatory liability. Besides California, there are parallel actions being taken elsewhere by lawmakers on the other end of the political spectrum: Florida and Texas are two major states attempting to pass legislation that would have an impact on online content moderation. In June 2021, the 11th Circuit Court of Appeals blocked major provisions of a Florida law that would penalize social media companies for blocking a politician's posts. The court cited infringement of the companies' First Amendment rights as the argument for its decision.[15] As for Texas, the US Court of Appeals for the 5th Circuit upheld its law that would prevent social media companies from blocking or taking down posts based on the poster's political ideology.[16]

It is notable that the state of California is among those leading the charge on setting content moderation regulations in place, given the technology industry's economic significance to the state. Supporters of such laws would argue as follows:[17]

> Efforts by social media companies to self-police [problematic] content have been opaque, arbitrary, biased, and inadequate. While some platforms share limited information about their efforts, the current lack of transparency has exacerbated concerns about the intent, enforcement, and impact of corporate policies, and deprived policymakers and the general public of critical data and metrics regarding the scope and scale of online hate and disinformation. Additional transparency is needed to allow consumers to make informed choices about the impact of these products (including on their children) and so that researchers, civil society leaders, and policymakers can determine the best means to address this growing threat to our democracy.

Meanwhile, not everyone agrees that a law such as AB587 is the answer. Eric Goldman offers a forceful critique.[18] He argues that AB587 amounts to censorship by placing regulators in the middle of the editorial process and second-guessing the platforms' editors; moreover, he is concerned that it places too heavy a burden on the platforms – up to more than 161+ different statistical disclosures and reporting systems that are incompatible with those required elsewhere and will impose

[15] Cat Zakrzewski, *Federal Judge Blocks Florida Law That Would Penalize Social Media Companies*, Washington Post (June 30, 2021), https://www.washingtonpost.com/technology/2021/06/30/florida-social-media-law-trump/ (last visited Feb. 19, 2024).

[16] Cat Zakrzewski, *5th Circuit Upholds Texas Social Media Law*, Washington Post (Sept. 16, 2022), https://www.washingtonpost.com/technology/2022/09/16/5th-circuit-texas-social-media-law/ (last visited Feb. 19, 2024).

[17] *Bill Analysis – AB-587 Social Media Companies: Terms of Service (2021–22)*, California Legislative Information, https://leginfo.legislature.ca.gov/faces/billAnalysisClient.xhtml?bill_id=20212022oAB587 (last visited Feb. 19, 2024).

[18] E. Goldman, *A Short Explainer of Why California's Mandatory Transparency Bill (AB 587) Is Terrible*, Technology & Marketing Law Blog (Aug. 9, 2022), https://blog.ericgoldman.org/archives/2022/08/a-short-explainer-of-why-californias-mandatory-transparency-bill-ab-587-is-terrible.htm (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

substantial extra costs. Goldman is concerned that, in the end, the information will not result in meaningful action that benefit users.

Considering such arguments both for and against such laws, the California law and similar laws from other states are destined for further debate and appeals. Eventually, they are likely to wind up in the Supreme Court. From the perspective of the platforms, all of this adds uncertainty on top of fragmentation.

Given differing demands from regulators in different jurisdictions and the changing nature of the legal landscapes from country to country and, in countries such as the US, from state to state, the social media companies will, no doubt, make choices in terms of where to allocate from their shrinking reservoir of content moderation efforts and resources. These resource allocation decisions are crucial as there is no globally deployable algorithmic solution that can moderate content worldwide, do it consistently, and at scale. According to UCLA's Sarah T. Roberts, author of *Behind the Screen: Content Moderation in the Shadows of Social Media*, "if you talk to actual industry insiders who will speak candidly and who are actually working directly in this area, they will tell you that there is no time that they can envision taking humans entirely out of this loop."[19] According to a study by the Transatlantic Working Group at the University of Pennsylvania's Annenberg Public Policy Center, algorithms are neither reliable nor effective in content moderation for many reasons: the absence of context, potential for bias, linguistic barriers, etc.[20] Moreover, content creators have also become creative about the way they can bypass automated blockers and filters through "algospeak": code words, mis-spellings, and veiled language to avoid algorithms that detect blocked hashtags and keywords. In some instances, the algorithm may over-correct; an example arose in the case of the Israel-Hamas war where many Palestinian posts were blocked as they referenced – not necessarily in support of – controversial topics or banned organizations. This means that there is no easy way to automate content moderation and the process must involve human intervention. Moreover, the content moderation workforce is not globally fungible; to be effective in any given geography, it must be trained in local languages, colloquialisms, locally relevant dog whistles, coded language, mores, and contexts.

Content moderation, in other words, requires platforms to allocate a budget to the activity, and the size and allocation of the budget would, naturally, follow the push and pull of regulators. Among the many revelations in Frances Haugen's testimony to Congress, 87 percent of Facebook's global content moderation budget was dedicated to identifying disinformation solely in the US, leaving 13 percent for all others. Notably, at the time of Haugen's testimony, 52.5 percent of Facebook's

[19] Z. Mack, *Why AI Can't Fix Content Moderation*, The Verge (July 2, 2019), https://www.theverge.com/2019/7/2/20679102/content-moderation-ai-social-media-behind-the-screen-sarah-t-roberts-vergecast (last visited Feb. 19, 2024).

[20] E. Llansó et al., *Artificial Intelligence, Content Moderation and Freedom of Expression*, The Transatlantic Working Group (Feb. 26, 2020), https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/06/Artificial_Intelligence_TWG_Llanso_Feb_2020.pdf (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

revenues (data from the first quarter of 2021) came from outside the US and Canada and 90 percent of users of Facebook were from outside the US.

To get a sense of how out of proportion the budget allocation was, consider the fact that besides the US, Facebook's own team had placed Brazil and India in its "tier zero" of highest priority of countries to monitor for hate speech and elections-related harmful content. Germany, Indonesia, Iran, Israel, and Italy were in next-highest priority: tier one.[21] The differences in terms of content moderation resources allocated were significant and notable in the disproportionality of the allocation. For the higher priority countries, Facebook's community standards were available in local languages, there were AI classifiers to detect hate speech and disinformation in these languages, and a larger team of human moderators deployed. In the rest of the world, few of these amenities were available: there were fewer translations, AI classifiers, or fact checkers; in some cases, none were available, even though the potential for harm was high.

As demands for regulatory enforcement of content moderation increase in the US and states take the lead in imposing different regulations on the platforms, the incentives for the companies are clear: to dedicate a disproportionate amount of content moderation resources to the US. In addition, the EU has taken on a more activist regulatory position. Its Digital Services Act contains more rigorous content moderation requirements and calls for greater transparency about the content on digital platforms. The Act also requires labeling of promoted content so users can place it in context. No doubt, the EU too will garner a significant share of content moderation resources.

In addition to the regulatory demands, there are economic incentives driving the resource allocation decisions of platforms. Consider the case of Meta and the origins of its revenues. The US and Canada generated $50.25 of revenue per user as compared to $15.64 in the next-highest market, Europe, and an average of $9.82 worldwide.[22] It is no wonder that Meta allocates resources disproportionately in favor of the US.

In the meantime, the technology sector itself has to contend with economic pressures and there is a need to keep costs under control. As the regulatory pressures in the US and EU add up, combined with the fact that the economic returns are so much more attractive in these markets, the rational decision for digital platforms would be to over-allocate content moderation resources in these jurisdictions at the expense of the rest of the world. The consequence of these decisions is a sobering one: with fewer resources allocated to content moderation, particularly in the developing world, the volume and intensity of disinformation is likely to only

[21] Casey Newton, *Facebook's Leaked Tier List: How the Company Decides Which Countries Need Protection*, The Verge (Oct. 25, 2021), https://www.theverge.com/22743753/facebook-tier-list-countries-leaked-documents-content-moderation (last visited Feb. 19, 2024).

[22] *Facebook's Average Revenue per User as of 2nd Quarter, 2022, by Region*, Statista, https://www.statista.com/statistics/251328/facebooks-average-revenue-per-user-by-region/ (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

increase – with potential for greater harm as many of the institutional safety nets present in the US and EU may be absent in these other parts of the world. Regulation in the global north, unintentionally, has a spillover effect on the global south by shifting the presence of harmful content from the north to the south. At the heart of this dynamic, once again, is the incentive system that drives the decision-making in the digital platforms: resource allocation decisions are driven by a combination of market and regulatory forces.

This is, in essence, what I would characterize as the "disinformation paradox": the desire to regulate content in one part of the world leads to less content moderation elsewhere; since the "elsewhere" is the larger user base with fewer institutional safeguards and users with fewer options for fact-checking or seeking alternative sources of news and information, the volume and intensity of disinformation elsewhere could be greater and the net impact on the world could be worse than the status quo.

13.4 SHIFTING INDUSTRY STRUCTURE, POLITICAL DYNAMICS, AND THEIR IMPACT ON CONTENT MODERATION

The imperatives that drive content moderation resource allocations also have roots in the changes in the digital industry structure and the political drivers that affect the platform companies. It is natural, as I have done so far, to center the discussion on content moderation using the example of the predominant platform company, Meta, whose issues in many ways are quintessential to those of the wider industry. That said, there are several other digital platforms that must contend with challenges that are idiosyncratic to their own circumstances. These idiosyncratic elements also influence how these platforms respond to the call for content moderation. Consider the examples of two sets of platforms, the first includes prominent ones, such as TikTok and X, formerly Twitter, and the second includes newer entrants in the industry that are positioned as havens for free-speech and, in some ways, free of content moderation responsibilities.

13.4.1 *TikTok*

If there is any platform that is giving Meta a run for its money, across all its platforms, it is TikTok, the video sharing app. TikTok is notable not only because it has been the most-downloaded app in recent years but has the potential for greater harm from hosting unmoderated content. Its users skew younger (it is more popular among Gen Z users than Instagram), representing a particularly vulnerable demographic, which often becomes a highly prized target of disinformation creators. TikTok holds the distinction of being the most engaging of all social media apps in the US – by far: an average user spends over ten minutes per session, which is twice that of the next most engaging app, Pinterest.[23] This, too, makes it a haven for manipulating users

[23] *Most Popular Social Media Apps in the US, as of September 2019, by Average User Session*, Statista, https://www.statista.com/statistics/579411/top-us-social-networking-apps-ranked-by-session-length/ (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

through content. Moreover, TikTok is very popular in the developing world where institutional and regulatory safeguards are under-developed: besides the US at number one, it is striking to note that Indonesia, Brazil, Mexico, Russia, Vietnam, the Philippines, Thailand, Turkey, and Saudi Arabia constitute the remaining nine countries in the top ten of TikTok users.[24] Next, consider the format of TikTok, which itself is another contributor to the challenges: content moderation is inherently hard because it is more difficult to filter video than it is to sort through textual content. On top of this, disinformation creators have found ways of bypassing roadblocks through deliberate misspellings of hashtags or innuendo, and other hacks of the moderation systems.

TikTok's influence is now so widely felt that even the search engine powerhouse, Google, is turning to it for inspiration and new ideas as it finds a growing proportion of Gen Z users are turning to TikTok for search and away from Google.[25] This means that TikTok is encroaching into territory that is well beyond that of social media and is under even greater scrutiny from both industry competitors and lawmakers.

But TikTok has another – more structural – organizational problem to contend with: it is owned by China-based ByteDance. With geopolitical tensions rising between the US and China, American policymakers have concerns about the ByteDance connection and the potential data vulnerability this represents as the company might give the Chinese government access to US user data. Others worry that TikTok itself could become a tool for content manipulation by Chinese officials. TikTok management is, of course, walking a fine line as the US is its most valuable market. This means, much like Meta but for reasons that extend into the geopolitical, it is likely to devote its limited content moderation resources disproportionately to the US. Particularly because the platform and its Chinese holding company is in the crosshairs of US lawmakers, we should expect it to tend to the squeaky wheels in the US. Thanks to both business and political pressures, the rest of the world should expect to be left under-resourced for content moderation by TikTok. Indeed, a Mozilla Foundation report finds that TikTok has become the force to reckon with for fueling disinformation and hosting "some of most dramatic disinformation campaigns" in countries such as Kenya.[26]

13.4.1 X, *Formerly Twitter*

For some platforms, content moderation rules are ad hoc and their organizations and processes haven't kept up with the evolving needs and the circumstances in

[24] *Which Countries Use TikTok the Most? TikTok User Data for 2022*, TikTok U.S. Data Security, https://www.houseofmarketers.com/which-countries-use-tiktok-the-most-tiktok-user-data-for-2022/ (last visited Feb. 19, 2024).

[25] *Google Borrows from TikTok to Keep Gen Z Searching*, WIRED (Sept. 28, 2022), https://www.wired.com/story/google-borrows-from-tiktok-to-keep-gen-z-searching/ (last visited Feb. 19, 2024).

[26] *From Dance App to Political Mercenary*, Mozilla Foundation, https://assets.mofoprod.net/network/documents/From_Dance_App_to_Political_Mercenary.pdf (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

which content is produced. X is an interesting case in point. While it is a highly influential platform, its user base is small relative to the Meta platforms or TikTok or YouTube, and the company often flies underneath the radar as much of the external scrutiny is on the behemoths. Recently, the spotlight turned on the platform because of the saga of Elon Musk staging and then reneging on a takeover of the company and then finally completing the acquisition. Twitter, as X used to be known, had struggled to introduce new features and innovations on its platform and its management has been under pressure to roll out new features. When it does, it seems the features run the risk of exacerbating disinformation problems. Prior to the Musk acquisition, in an explosive revelation, Twitter's former head of security Peiter "Mudge" Zatko alleged that its Spaces feature suffers from very poor content moderation. "About half of Spaces content flagged for review was in a language that the moderators did not speak, and there was little to no moderation happening," according to Zatko's whistleblower complaint filed with the US Securities and Exchange Commission, the Federal Trade Commission, and the Justice Department.

What was even more damning was that Twitter executives were, reportedly, aware of the potential for abuse but refused to slow the roll-out of Spaces, despite its use by white nationalists, Taliban supporters, and anti-vaxxers, each being notorious for their reliance on creating and spreading harmful content. Employees who raised concerns were told that given the volume of conversations and multiplicity of languages, the technology necessary to properly moderate Spaces simply did not exist.[27]

For another "innovation" on Twitter that could complicate the process of content moderation, consider the Edit Tweet feature: it lets users edit their posts for a fixed time period after posting. This feature could open the door to harmful content as miscreants could edit in language after a harmless tweet has gone out and has been retweeted and circulated. The edit could allow disinformation creators to bypass the initial content moderation screens and gain a wide audience.

The biggest blow to the platform's content moderation capabilities came after Elon Musk's acquisition. The Fletcher School's Digital Planet found that hate speech and toxic narratives have surged in the weeks immediately after the takeover. The analysis focused on three hateful narratives: anti-Semitism, anti-LGBTQ+ rhetoric, and racial/ethnic hate speech, and found a marked increase in all three narratives.[28] Subsequently, Elon Musk, the owner of the platform himself contributed to harmful speech directed at Yoel Roth, the person who had been responsible for Trust and Safety on the platform after Roth had been fired by Musk. Roth was forced to flee his home after he received

[27] Aisha Malik, *Twitter Ex-security Head Says the Social Network Has "Deficient Moderation" for Spaces*, TechCrunch (Aug. 23, 2022), https://techcrunch.com/2022/08/23/twitter-mudge-deficient-moderation-spaces/ (last visited Feb. 19, 2024).

[28] *Hate Speech on Elon Musk's Twitter: Under Musk, Hate Speech Is Rising*, Digital Planet (Dec. 8, 2022), https://digitalplanet.tufts.edu/musk-monitor-under-musk-hate-speech-is-rising/ (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

death threats as a direct outcome of Musk's posts. It is hard to find a more egregious example of under-moderation of content by digital platforms anywhere.

The dismantling of content moderation at X has had consequences. The platform run afoul of the EU's Digital Services Act, it has received a stern rebuke from Thierry Breton, the European commissioner who oversees the law, and advertisers have fled the platform in droves.

13.4.3 *Niche Players Championing Free Speech*

The heightened scrutiny of the major platforms has created a market entry opportunity for newer platforms that can cater to the needs of niche groups, disinformers, and others who wish to fly under the radar. Consider platforms such as BitChute, Odysee, and Rumble that had grown as the major incumbent platforms instituted more content rules. The newer platforms were positioned in the market as champions of free speech and, as a result, had become concentrators of disinformation, from conspiracies to hate speech. In an increasingly polarized climate in countries like the US, such platforms have gained an audience quickly. According to data from digital intelligence firm Similarweb, BitChute's online traffic grew 63 percent in 2021; it commanded an audience more than double that of MSNBC.com at that time.

To add to the regulatory challenges, it is not just classic social media platforms where harmful content could find a home. Twitch, the Amazon subsidiary, where users gather to watch skilled gamers play popular games, such as Fortnite and Minecraft, has become open territory for predators who pursue young users to exploit them. The moderation rules at Twitch have not succeeded in stopping children from broadcasting their presence and blocking predators.[29]

13.5 POTENTIAL SOLUTIONS

The earlier discussion highlights the difficulties of effective content moderation either initiated by the companies themselves or by activist regulators. In fact, regulatory activism could make the problem worse globally. This raises the question: What are possible alternative routes to a solution?

Consider some options.

13.5.1 *Letting Regulators Focus on What's Best for Their Own Jurisdictions*

This appears to be the default approach based on the actions that have been taken thus far. As discussed earlier, a regulator in a substantive market, such as the US or

[29] C. D'Anastasio, *Child Predators Use Amazon's Twitch to Systematically Track Kids Who Stream*, Bloomberg (Sept. 21, 2022), https://www.bloomberg.com/graphics/2022-twitch-problem-with-child-predators/?cmpid=BBD092222_TECH&utm_medium=email&utm_source=newsletter&utm_term=220922&utm_campaign=tech&leadSource=uverify%20wall (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

EU, could take actions that have a spillover effect on other jurisdictions by increasing the amount of disinformation in these latter regions, thereby giving rise to the disinformation paradox.

The complication is that it is hard to imagine that lawmakers and regulators in the parts of the world where disinformation spillover occurs will sit still. They are likely to respond by setting up their own restrictions and regulations and, like their US and EU counterparts, likely to do so without regard for the potential for harm to those on the outside. The regulators put their own regulations and restrictions in place for both defensive and offensive reasons. The defensive action is understandable as these jurisdictions feel there is insufficient content moderation occurring in their jurisdictions, so they feel the need to step in. The offensive action could be because governments around the world see platforms as powerful mechanisms for shaping political narratives; hence regulating and exerting control over them becomes important for existential reasons.

The collective outcome is that regulators across the world engage in a noncooperative game and are trapped in a cycle of mutually reinforcing "beggar-thy-neighbor" actions tantamount to a classic prisoners' dilemma problem. We can also think of this situation as a tragedy of the commons, where authenticity of information as a public good is under threat. Everyone is worse off because the platforms are left balancing competing regulations and rules, making them less effective in moderating content overall. Misinformation content creators take advantage of the fragmentation and lack of coordination across regulatory regimes to do their mischief. The costs of content moderation to the platforms go up and their incentives are to pare back content moderation resources to the extent they can.

To make matters worse, some governments are likely to utilize the window of opportunity created by this lack of discipline across the world and use the regulatory cudgel to shape narratives and, potentially, get into the disinformation business themselves or stifle action that promotes the spread of authentic information. A case in point is India, where Alt News, a leading fact-checking organization on issues ranging from child kidnapping rumors to anti-Muslim rhetoric, ran afoul of the administration that was keen on pushing its own version of the facts. One of the founders of Alt News was jailed and has charges pending against him with accusations against him being leveled by the ruling BJP party.[30] In other instances, countries, such as Nigeria, blocked Twitter altogether. In other circumstances, analysts argue that in countries, such as Kenya, where, historically, there were no content moderation rules in place, the abundance of disinformation will provide governments the excuse to impose strict censorship and shut down entire platforms, leading to outright censorship. As the author of the Mozilla Foundation report on disinformation on TikTok in Kenya noted, "If the platforms don't get their act

[30] Suhasini Raj, *In India, Debunking Fake News and Running into the Authorities*, New York Times (Sept. 22, 2022), https://www.nytimes.com/2022/09/22/world/asia/india-debunking-fake-news.html (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

together, they become convenient excuses for authoritarians to clamp down on them across the continent . . .. And we all need these platforms to survive. We need them to thrive."[31]

13.5.2 *Regulatory Coordination*

There are several – imperfect – resolutions to the tragedy of the commons problem in other contexts. To consider their application in the current situation, in essence, we need the regulators in the various jurisdictions to get to a "cooperative" outcome where they harmonize or coordinate on the demands placed on the platforms. In addition to the benefits of harmonizing regulation and setting standard guidelines, such coordination by regulators and lawmakers across jurisdictions also helps build up their collective bargaining power vis-à-vis the major platforms. Such rebalancing of bargaining power can help avoid situations such as the experience of Australian lawmakers noted earlier, where Meta simply used its market power to force the lawmakers to acquiesce.

In antagonistic settings, such cooperative outcomes have arisen out of threats of noncooperation, involving a commitment to punishing any deviation from the cooperative outcome. This does not appear to be desirable or practical in the international regulatory context as it doesn't seem practical for regulators to punish other regulators or even to threaten them implicitly. Alternatively, we can consider a solution involving explicit coordination among regulators. There is a precedent for such actions in other international contexts where there is a risk of a tragedy of the commons: consider the examples of agreements over trade and monetary policies, energy use and GHG emissions reduction, public health measures, international security alliances, among many others.

Such international coordination could work if there were sufficient guarantees that all or most regulators would achieve their objectives from a coordinated solution. As has been noted in the analyses of the effectiveness of international coordination in other contexts, such coordination fails when:

- Different countries come to the table with fundamentally different models in mind which leads to a breakdown in communication. As Jeffrey Frankel has noted in his analysis of coordination over international trade and monetary policy, when two players sit down at the board, "they are unlikely to have a satisfactory game if one of them thinks they are playing checkers and the other thinks they are playing chess."[32]

[31] Neha Wadekar, *Why Dangerous Content Thrives on Facebook and TikTok in Kenya*, Washington Post (July 31, 2022), https://www.washingtonpost.com/world/2022/07/31/kenya-meta-tiktok-facebook-disinformation/ (last visited Feb. 19, 2024).

[32] Jeffrey Frankel, *International Coordination*, Faculty Research Working Paper Series, RWP-16-002, Harvard Kennedy School (Jan. 2016), https://www.hks.harvard.edu/research-insights/research-publications (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

- Different countries make assumptions about why the other parties wish to coordinate and these assumptions are based on their own models, which could be wrong and sets the coordinated agreement up for failure.
- Each country has its own set of internal political dynamics to attend to and there are interest groups with differing agendas that destabilize the cross-country agreements.
- Countries vary significantly in their interests in regulating versus letting the market do the job.
- Agreements are harder as there are more countries at the table and there are significant asymmetries across the needs and interests and the relative market and political power of the countries.

Each of these factors is present in the case of coordinating over content moderation. In the increasingly fractured and politicized world of managing online content, the chances of such misperceptions are high, especially when it is clear that the platforms only have a limited set of resources to go around.

13.5.3 *Extending a Regulator's Reach beyond the Jurisdiction*

Can a regulator hold a platform accountable for content moderation in geographies that are not part of its jurisdiction? In theory, this could be a way in which the regulators in the US or the EU could anticipate the emergence of harmful content elsewhere, especially in the most vulnerable parts of the world, and put their leverage to good use while mitigating the unintended consequences of vulnerable jurisdictions being starved of content moderation resources.

One solution is to have lawmakers in regions such as the US and the EU pass laws to regulate not only the content hosted by the platforms but also on the investments that platforms make on content moderation and how these resources are deployed across the world. For instance, platforms could be required to maintain a certain threshold of content moderation staff for countries in proportion to the number of users and the level of disinformation risk there.

While these ideas have merits, there are several problems to consider.

One is the obvious problem of going against self-interest. If it is common knowledge that there is only a limited amount of content moderation resources to go around, then there is a potential zero-sum game between such resources being deployed in the regulator's home jurisdiction versus elsewhere. By holding platforms responsible for content elsewhere, the regulator runs the risk of thinning out resources to moderate content that affects the constituency that the regulator cares about the most – in its own jurisdiction.

A second problem is best illustrated by the example of California's AB 587, which provides a useful parallel. The California law extends its reach beyond the state and requires disclosures from entities and users from outside of California, including

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

from foreign countries. This opens up the possibility of potential conflicts with a multitude of laws and restrictions that may be incompatible with those proposed for California. As Eric Goldman has argued, this can create several "Dormant Commerce Clause" problems: he cites the example of a recent New York law that defines "hateful conduct" and specific requirements for dealing with it, which may not coincide with what California requires, or the disclosure requirements in Texas that are structured differently from those in California. When such differences are added across numerous jurisdictions, Goldman argues, the law would place an undue burden on the platform.[33]

It is conceivable that a similar argument could be applied to any regulation that affects content moderation beyond the regulator's jurisdiction and would result in an undue burden on the platform leading to an impractical solution or an unenforceable task for the regulator. The burden is likely to increase as one considers the reality that this concern extends not just to multiple states in the US but to multiple countries.

13.5.3 *Algorithmic Solutions*

In theory, the most equitable way to catch disinformation without skewing resources in the direction of the most powerful nations is by relying on automated systems that can reliably monitor and filter content worldwide. Algorithms can, in principle, perform linguistic or graphic or visual analysis, for example, and flag problematic content. Text can be interrogated for word vectors, word positioning, and connotation or reverse engineer images to catch harmful content or deepfakes. If algorithms could in the future catch most of the harmful content increases, the dilemmas highlighted here are significantly lessened as the reliance on a limited pool of human content moderators is reduced.

But, as I noted earlier, we are too far from that future. With algospeak, even simple workarounds such as misspelling hashtags or introducing punctuation marks can help disinformation get around filters and blocking systems. Moreover, there is an entire arsenal of coded language and dog whistles that vary by context, language, and culture that allow the bad stuff to slip through. It would take a long time before enough data can be collected to train algorithms to reliably recognize all such bypassing tricks and hacks.

Moreover, an over-reliance on algorithms developed in the West can create its own biases and inequities as these systems fail to recognize codes from parts of the world that are far away from the hubs where such AI-powered tools are being

[33] E. Goldman, *California Seems to Be Taking the Exact Wrong Lessons from Texas and Florida's Social Media Censorship Laws*, TECHDIRT (June 23, 2022), https://www.techdirt.com/2022/06/23/california-seems-to-be-taking-the-exact-wrong-lessons-from-texas-and-floridas-social-media-censorship-laws/ (last visited Feb. 19, 2024).

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

developed. To appreciate the difficulties, consider an example from Kenya, where a narrator in a video clip mimicked a detergent commercial and spoke of "detergents" to eliminate "madoadoa," an innocuous word meaning a spot to be removed from a piece of clothing. In reality, the word "madoadoa" in this context referred to members of Kamba, Kikuyu, Luhya, and Luo tribes and was inciting viewers to violence against them. To make matters worse, these code words shift over time and younger users are constantly developing fresh colloquialisms and neologisms, especially for sensitive topics, which would make the search for the algorithmic solution a task of Sisyphean proportions.

There is one area where one could hold out hope for algorithmic solutions. As more harmful content gets created by other algorithms, using generative AI, it is conceivable that the process of training such "bad" AI could be reverse engineered to then train "good" AI, which becomes better at catching the bad stuff.

That said, it is hard to see a future where humans are taken out of the loop altogether. Humans will be present at the creation of content. The need for and reliance on human moderators will never go away. Ironically, much of the labor for such moderation will continue to be drawn from the countries where labor is cheap and plentiful – and these are the countries that are most at risk of disinformation being rampant. It is also hard to see a future without the role of regulators taking a more active stance on content moderation. But they need to go beyond demanding that there ought not to be harmful content disseminated within their jurisdictions, as the Digital Services Act does. As noted earlier, lawmakers need to be pressed to require that platforms maintain a minimum threshold of content moderation staff for countries in proportion to the number of users and an objective evaluation of the level of disinformation risk in these countries.

13.6 IN CONCLUSION: A CASE FOR BETTER DIGITAL HYGIENE

In closing, I would like to offer a very different approach to the solution as a complement to the ideas presented earlier. While this approach in no way takes care of the problem, it is important to pay attention to an important piece of the disinformation puzzle – by taking the spotlight off of platform responsibility but pointing it at the other end: on user responsibility. There is a need for investment in better education and global standards in "digital hygiene." Much like reading, writing, social studies, and even personal hygiene, the ability to use the internet in productive ways and sort authentic information from disinformation ought to be codified and integrated into curricula in schools across the world. We have done very little in this arena since the bulk of the efforts has been directed at policing or regulating content on the platforms.

If we accept that disinformation or harmful content will never be eliminated, possibly for all the reasons discussed in this chapter, we must invest in ways to mitigate the damage wherever possible and limit its power in ways that can be shared

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013

universally. Driving on highways, for example, always carries a degree of risk, which is why we emphasize road safety rules and inculcate better driving habits. In combination with the introduction of new technologies that aid automotive safety, this will lead to fewer crashes. My final recommendation would be to apply similar principles to the problem of content moderation and invest in the area of user education and to institutionalize it in schools across the world. No matter what else is done to regulate content moderation or holding platforms to account, educated digital consumers represent the best – and most inclusive – defense against the scourge of disinformation.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 19 Nov 2026 at 04:44:21, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009438636.013