15

Conclusion

The Future of Defeating Disinformation

Bhaskar Chakravorti and Joel P. Trachtman

Even as it presents a threat to the institutions of democracy, the industry of disinformation is highly democratic: the tools to produce it and distribute it are readily accessible, whether to a humble farmer in India who has discovered WhatsApp, a teenager anywhere in the world with time and a smartphone on hand, or a former US president hyperactive on social media.

It takes little to piece together digital falsehoods. With some strategy, luck, and timing, there are tools and platforms that could spread the messages to millions, who could, in turn, spread them even further. AI has added a fresh boost of creativity to the disinformation industry. Now anyone can become a political content creator[1] thanks to generative AI tools, such as DALL-E, Reface, DeepFaceLab, and scores of others. Even before the year 2024 – declared as the "biggest election year in history" – commenced, the tools of AI have already turbocharged disinformation campaigns in Bangladesh setting the tone for all elections to come.[2]

The democratization of the disinformation process has even the experts worried: Former Google CEO Eric Schmidt warned[3] that "you can't trust anything that you see or hear" in the elections thanks to AI.[4] Sam Altman, CEO of OpenAI – the company that gave us the generative AI tool, ChatGPT – told US lawmakers that he is nervous about the future of democracy. Besides, the state of the wider environment adds to the potential for problems. Political polarization, major conflicts,

[1] Darell M. West, *Comparing Google Bard with OpenAI's ChatGPT on political bias, facts, and morality*, Brookings (Mar. 23, 2023), https://www.brookings.edu/articles/comparing-google-bard-with-openais-chatgpt-on-political-bias-facts-and-morality (last visited Feb. 16, 2024).

[2] *2024 Is the Biggest Election Year in History*, The Economist (Nov. 13, 2023), https://www.economist.com/interactive/the-world-ahead/2023/11/13/2024-is-the-biggest-election-year-in-history (last visited Feb. 16, 2024).

[3] J. Frank, *The 2024 Presidential Race Is the AI election*, Axios (June 27, 2023), https://www.axios.com/2023/06/27/artificial-intelligence-ai-2024-election-biden?utm_source=dlvr.it&utm_medium=twitter (last visited Feb. 16, 2024).

[4] S. Frenkel & K. Conger, *Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find*, Axios (June 27, 2023), https://www.axios.com/2023/06/27/artificial-intelligence-ai-2024-election-biden?utm_source=dlvr.it&utm_medium=twitter (last visited Feb. 16, 2024).

culture wars, ethnic divisions, worries about the environment, and even about technology itself have all added fuel to the disinformation machine.

One would think that social media platforms would be investing in war rooms especially in a big election year, such as 2024 – as they have done in the past[5] – and putting in place plans to catch disinformation before it spreads. Instead, companies across the tech sector caught in an industry-wide slowdown have had other pressing business to attend to, such as propping up profitability and staking out territory in the ever-widening space of emerging AI opportunities. Since the pandemic era highs, priorities have shifted to reducing staff and making cuts in parts of the company that do not directly contribute to increasing revenue, bringing in new users, or encouraging existing users to post messages that engage others.

Meta offers a case in point. CEO Mark Zuckerberg had declared 2023 as the "year of efficiency."[6] The trust and safety team that moderates content on Meta's platforms – which include Facebook, Instagram, and WhatsApp – was drastically reduced,[7] a fact-checking tool that had been in development for six months was shelved,[8] and contracts with external content moderators were canceled.[9] These developments at Meta were mirrored by the dismantling of content moderation resources elsewhere across the industry. Under Elon Musk, Twitter (now called X) decimated its content moderation teams, steadily lifted restrictions, and restored accounts that had been suspended because of their track records of spreading disinformation.[10] YouTube announced[11] its intention to lift a ban on videos making false claims over the 2020 US election. Google had cut[12] the team that monitors misinformation and toxic speech by a third by February of 2023. In fact, those who

[5] Sheera Frenkel & Mike Isaac, *Inside Facebook's Election "War Room,"* New York Times (Sept. 19, 2018), https://www.nytimes.com/2018/09/19/technology/facebook-election-war-room.html/ (last visited Feb. 16, 2024).

[6] J. Vanian, *Meta's "Year of Efficiency" Was Everything Wall Street Needed to Hear from Zuckerberg*, CNBC (Feb. 1, 2023), https://www.cnbc.com/2023/02/01/metas-year-of-efficiency-everything-wall-street-needed-to-hear.html/ (last visited Feb. 16, 2024).

[7] *Id.*

[8] H. F. Vanian Jonathan, *Tech Layoffs Ravage the Teams That Fight Online Misinformation and Hate Speech*, CNBC (May 26, 2023). https://www.bbc.com/news/technology-64225270 (last visited February 16, 2024).

[9] Chris Vallace, *Meta Denies African Content-Moderator Firm Exit Poses Risk*, BBC (Jan. 10, 2023), https://www.bbc.com/news/technology-64225270 (last visited Feb. 16, 2024).

[10] Sheera Frenkel & Kate Conger, *Hate Speech's Rise on Twitter Is Unprecedented, Researchers Find*, New York Times (Dec. 2, 2022), https://www.nytimes.com/2022/12/02/technology/twitter-hate-speech.html (last visited Feb. 16, 2024).

[11] *An Update on Our Approach to US Election Misinformation*, YouTube Official Blog, June 2, 2023, https://blog.youtube/inside-youtube/us-election-misinformation-update-2023/ (last visited Feb. 16, 2024)

[12] T. Brewster, *Google Cuts Company Protecting People from Surveillance to a "Skeleton Crew," Say Laid Off Workers*, Forbes (Feb. 3, 2023), https://www.forbes.com/sites/thomasbrewster/2023/02/02/jigsaw-google-alphabet-layoffs/?sh=5efab4c52d71 (last visited Feb. 16, 2024).

work in trust and safety have found that there aren't many job openings in their area of expertise.[13]

This raises the question of the pressing need for regulation and its effectiveness in controlling the spread of disinformation on digital platforms. The regulation of platforms – including but not limited to the assignment of responsibility to platforms – is, of course, secondary, "gatekeeper" regulation, intended not so much to regulate the behavior of the platforms themselves but to regulate user-generated content. Gatekeepers are regulated in other areas, where, from a regulatory standpoint, they have better access to information than government regulators would have, and where their incentives are distinct from those of the actors whose underlying behavior is sought to be addressed. For example, in the US, securities offerings underwriters are required to apply "due diligence" in ensuring that prospectuses for securities offerings are not fraudulent.

In the digital platform context, the sheer volume and velocity of speech is a challenge to regulators, and so in this relatively novel field, some governments have sought to assign responsibilities to platforms. It is noteworthy that in the European Union's Digital Services Act (DSA), the main obligation of platforms is one of "due diligence." Due diligence is not specifically defined, and moreover, the underlying "offensive" addressed in the DSA's due diligence obligation may not itself be a violation of law. And for good reason – the "offense" may be protected free expression. It is in making the trade-off between protected free expression and legally proscribed behavior that the challenge of international platform responsibility arises. Governments may cross a constitutional or human rights line when imposing responsibilities on platforms that cause them to restrict speech.

We hasten to add that platforms can themselves engage in objectionable behavior – for example, failing to monitor and remove objectionable content, amplifying the distribution of objectionable content, whether through algorithms or otherwise, or censoring benign content. As a rapid and essentially costless global medium, platforms provoke states to regulate behavior that they formerly might have ignored. In grappling with platform responsibilities, states may be drawn to regulate speech that in a different medium would have been subject to a laissez-faire approach.

The problem of platform regulation per se is a novel problem. States may develop the capacity to regulate user generated content directly, without regulation of platforms. Governments are developing this capacity at different rates. In the US, for example, users can be prosecuted for content that violates laws governing topics such as child pornography or buying illegal drugs online or content promoting banned terror organization, among others. In general, direct censorship of the internet is prohibited in the US. Citizens are protected by the First Amendment, which grants them rights to freedom of speech and expression. As is to be expected, the rules on what is considered illegal digital content vary depending on the country.

[13] *Supra* note 8.

Different countries have different rules governing what is or is not allowed for posting on social media. Many governments are experimenting with less intrusive approaches, such as parental control tools, age-rating systems, and other controls. In some cases, governments have gone to the opposite extreme by shutting down entire platforms or the internet altogether; in fact, according to Freedom House, internet freedoms have been declining for over a decade.[14] In fact, AI has been a powerful mechanism for accelerating that decline. In at least twenty-two countries, digital platforms are required to use automated systems to remove content deemed illegal under local laws; often the content that is censored is protected according to international human rights standards. In at least forty-seven countries, governments used generative AI to push state propaganda.[15]

While all such regulations and restraints are critical to content moderation and contribute to the volume of disinformation around the world, another fundamental problem – and source of complexity – stems from the intensification of cross-border activity on platforms, which increases collisions between different states' regulatory visions.

Not only does platform activity change the regulatory calculus in ways that challenge individual states, but it also exists in what may be or should be a global commons. Thus, the novel problem of platform governance is exacerbated by globalization. The research program that formed the basis for this volume was prompted by many changes being enacted and considered in different parts of the world, from the growing demand to amend Section 230 of the US Communications Decency Act to the EU's Digital Services Act to firewalls and unilateral steps being taken in countries such as China or India to place limits on content shared on social media platforms. When all of these regulatory moves are being enacted by governments unilaterally and in the absence of international harmonization, a basic question arises: How will changes in a single jurisdiction affect the ability of other states to achieve their regulatory goals, and vice-versa? How might platforms that operate across national borders moderate content that is borderless and is, yet, governed by different rules across different regulatory jurisdictions?

As of this writing, no amendment has been made, and the US Supreme Court has recently refrained from pronouncing on the scope of Section 230 – in particular, the extent to which platforms may be treated as "publishers" or "speakers" where they algorithmically or otherwise select which content to promote. And yet, the analyses of Brazilian, Chinese, European Union, Indian, and US approaches to platform responsibilities show wide divergence in mediating the tension between, on the one

[14] Allie Funk, Adrian Shahbaz & Kian Vesteinsson, *The Repressive Power of Artificial Intelligence, FREEDOM ON THE NET 2023*, Freedom House (Oct. 4, 2023), https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence/ (last visited Feb. 16, 2024).

[15] Allie Funk, *Advances in AI Are Compounding Internet Freedom's Decline. But They Don't Have To*, Freedom House (Oct. 4, 2023), https://freedomhouse.org/article/advances-ai-are-compounding-internet-freedoms-decline-they-dont-have/ (last visited Feb. 16, 2024).

hand, treating platforms as transparent and thereby regulating only the users who generate the content and, on the other hand, seeking to assign to platforms responsibilities to moderate user-generated content on their platforms.

Of course, one solution to this problem is a kind of platform "splinternet," in which, for example, TikTok must follow different policies in the US from those followed by its sister company, Douyin, in China, as related by Jufang Wang (Chapter 4). This is a case of two platforms with essentially the same structure, but different memberships, and most importantly, different policies. China's regulations would conflict with US constitutional free speech principles, as well as other US public policy. The Chinese state does not permit foreign platforms to operate in China, presumably because it cannot sufficiently apply its regulation to foreign platforms. As this book is finalized, the US is considering a similar approach: debating legislation to ban TikTok if its ownership structure is not changed. The EU's DSA does not bar foreign control but applies to the extent that foreign platforms provide services to EU persons.

On the other side of the jurisdictional divide, Section 230 would probably prevent enforcement in the US of platform liabilities for failure to carry out responsibilities under non-US law that are inconsistent with Section 230. But of course, a US company that provides platform services in China or the EU may find that its assets outside the US are subject to enforcement actions, or that it would be blocked from operating in those jurisdictions unless it satisfied its regulatory responsibilities there.

These jurisdictional problems operate at two levels: first, at the level of substantive law, and second, at the level of specialized platform responsibilities, such as obligations under the DSA to engage in due diligence. Global platforms increase the collision of substantive law – including law relating to elections, privacy, consumer protection, libel, securities regulation, taxation, competition, intellectual property, or others. They make it more difficult for businesses to craft a compliance strategy insofar as either one government prohibits what another requires, or insofar as a "highest common denominator" strategy that complies with all the most restrictive policies places an international platform at a competitive disadvantage vis-à-vis purely national platforms. At the level of platform due diligence, different standards also will require the platform to follow a highest common denominator strategy. On the other hand, to the extent that governments will focus on harms that occur within their own territories, platforms will have incentives to focus monitoring resources on the most restrictive and most remunerative jurisdictions, perhaps providing insufficient monitoring resources for developing countries.

In this book, we see that China has deputized platforms to assist in maintaining a state-sanctioned information environment for "public opinion management" and has excluded foreign platforms from operating in China, presumably because they would be unable or unwilling to cooperate with state control. As Jufang Wang states: "China's requirements for platforms to proactively monitor, moderate and sometimes censor content, especially politically sensitive content, make it almost

impossible for foreign social media platforms to survive without full compliance." Thus, China is distinct both in substantive law and in platform regulation, and the degree of difference makes interoperability practically impossible. This is one extreme of national behavior.

The other extreme, perhaps, is the US, in which First Amendment fundamentalism limits the substantive regulation of speech. In the converse of China's circumstances described above, the US objects to foreign platforms such as TikTok regulating speech to US persons. Section 230 goes even further, insulating platforms from responsibilities applicable to users who generate content. This keeps the government out of the business of regulating speech but, in addition, leaves open the possibility of private sector regulation of speech: under Section 230, platforms may moderate as they see fit. One path to reform in the US would be to limit the scope of autonomy that platforms have to moderate – and it might be attractive to see platforms as a type of "common carrier" that have an obligation of political neutrality, while policing certain types of speech according to neutral principles, such as illegality or demonstrable falsity.

The EU, Brazil, and India may be viewed as operating in between these polar extremes. The EU is most interesting. First, in the EU, most regulation of content is provided by national law. Second, the EU, to some extent, prioritizes privacy over free speech, causing some friction with the US. Third, at the EU level, the DSA operates by imposing systemic risk management obligations on platforms without explicit specification of what the risk management obligation must accomplish, for example, in the area of election interference. As Christoph Busch (Chapter 3) notes, "the scope of this risk assessment includes the "dissemination of illegal content," "any actual or foreseeable negative effects for the exercise of fundamental rights," and "any actual or foreseeable negative effects on civic discourse and electoral processes, and public security." The DSA will apply to all platforms, irrespective of their place of establishment, to the extent they provide services to users in the EU.

As this volume's generative and interdisciplinary chapters suggest, there may be multiple means of improving the organization of allocation of authority over platform content. As Federico Lupo-Pasini notes (Chapter 8), platforms externalize costs to society without national regulation, and states may externalize costs to citizens of other states, or simply restrict international commerce, without international rules. There is a trade-off, as he suggests, among (i) a global market, (ii) regulatory sovereignty, and (iii) internalization of regulatory externalities. Regulatory sovereignty can only be achieved by compromising the global market or accepting regulatory externalities. One response is harmonization – giving up regulatory sovereignty to some extent. Regulatory competition can be permitted to operate, but it may be driven in part by externalization of costs to other states. Regulatory networks are an intermediate solution, utilizing communication and negotiation to minimize unnecessary differences in regulatory regimes. Soft rules also may be more acceptable but less satisfactory from a regulatory standpoint. Mutual recognition

may be useful where goals are sufficiently shared and merely implemented in different ways. The discussion above suggests that there may be important differences in goals among different leading jurisdictions.

In Chapter 9, Carlo Garbarino also focuses on the problem of regulatory externalities, from the perspective of international cooperation and diversity in the global tax system. To the extent that common standards may be achieved, these externalities can be reduced, and states may also cooperate with one another in the absence of agreement on standards by seeking to reduce externalization of costs by their actors. He evaluates policies that counteract aggressive tax strategies based on the mobility of capital as potential models for policies for platform content moderation with mobile computational capital. One way forward, as in financial regulation, is through multilateralism among like-minded states.

Mark Jit and Dominik Hofstetter (Chapter 7) derive lessons from state regulation and cooperation in managing infectious disease. While they suggest that some areas of regulation may present "low-hanging fruit" for cooperation, other areas where states have different regulatory goals will be distinct from the field of infectious disease where states tend to have more homogeneous interests. This would suggest that an internationally negotiated agreement on standards would help to support greater cooperation in managing disinformation.

In order to manage the issue of diversity of regulatory vision, states may, to some extent, harmonize substantive regulation. This is less likely than states determining unilaterally or multilaterally to develop manageable rules of jurisdiction, so that their regulation applies only in limited circumstances. The fullest realization of this "choice of law" solution would involve geo-blocking or other technology that divides up regulatory authority according to a specified, and a perhaps agreed-upon, principle. Geo-blocking may be costly and ultimately porous, but it would allow different communities to effectuate their different visions of the good in the platform context. To the extent that the principles of jurisdiction are agreed, and are structured to be exclusive, platforms would have the certainty of knowing the requirements under which they must operate in each market. Of course, different communities may remain territorial states, but given the a-territorial nature of the internet, it may be possible for other divisions of authority and responsibility to develop. Cultural affinity, or political perspective, may be more compelling as an organizational principle to some than territorial co-location.

Eventually, our ability to defeat disinformation worldwide hinges on a critical factor: the incentives of multiple stakeholders to cooperate on several fronts. These fronts could include common standards that define acceptable content, markers of deviation from standards and actions to be taken in the case of deviations, and fact-checking ecosystems using an open-source mechanism involving designated experts of different regions.

A natural question to ask is: How likely is it that multistakeholder interests will converge? The issue of misaligned incentives across regulators as well as across the

platform companies, one that Bhaskar Chakravorti returns to repeatedly, cannot be ignored. At a symposium on the subject of this book, held in December 2022 at The Fletcher School, Josephine Wolff, who contributed to this book, noted that, "While enforcing different rules and policy measures for different national versions of a platform is certainly not the same as reaching international consensus on these measures, it is no small thing that so many countries have accepted this approach as a satisfactory implementation of their domestic laws …. This widespread acceptance of a highly imperfect system of implementation and enforcement is, itself, an impressive feat of international coordination." Joel Trachtman cautions that there are many added complexities to consider; for example, even on the matter of free speech, different countries have different legal formulations of free speech. Multiple countries' laws can apply to the same platform or the same platform transaction resulting in platform-based dispersal escaping regulation altogether. Daniel Drezner, also a contributor to this book, is even less sanguine about the stakeholders getting to a state of meaningful coordination; he argues that countries have wildly divergent preferences in terms of which internet content should be regulated and notes the possibility of the creation of "sham standards" at the global level. He worries about token agreements negotiated at the global level that lack enforcement mechanisms and likely be honored only in the breach.

There are, no doubt, many gaps to be closed. We conclude our approach to the daunting task of defeating disinformation with a five closing ideas from the contributors in this book.

15.1 MIND THE POLITICAL CONTEXT

As Eric Goldman (Chapter 2) points out, online speech freedoms are intertwined with partisan politics; in the US, Democrats want more content moderation; while Republicans want fewer restrictions as they see the restrictions as disproportionately affecting those on the right of the political spectrum. He concludes that this fracture is causing the two sides to advance two radically different visions of the internet's future. This calls for finding common ground before enacting legislation – which is bound to be a slow process of negotiation. Jufang Wang (Chapter 4) offers a portrait of risks of the opposite – a single party – environment. The Chinese version of platform regulation involves heavy obligations and administrative measures for enforcement. The requirements for platforms to proactively monitor, moderate and sometimes censor content, especially political sensitive content, de facto lock out foreign social media platforms. This creates a very different kind of problem altogether.

Considering a context, such as the EU, that sits in between the US and Chinese polar extremes, Joel Trachtman (Chapter 14) notes that subsidiarity demands harmonization within the EU and possibly elsewhere. He argues that optimality depends on the value to states of diversity, and the costs of legal integration. This

might call for more specialized international bodies, or the coordination to be done within the framework of a trade agreement.

15.2 BORROW FROM OTHER REGULATORY MODELS

One problem that has been raised is that regulators are limited by their jurisdictions while digital platforms are global in their reach; so there is a mismatch between what the regulators can affect and the effects of the actions of the platforms. There are potential solutions to consider from other domains. As one such source of inspiration from other domains, Federico Lupo-Pasini (Chapter 8) offers the example of the US Dodd-Frank Act. It allows for extraterritorial jurisdiction by linking the application of domestic law to local contact points.

15.3 DEFINE STANDARDS

As Mark Jit and Dominik K. Hofstetter (Chapter 7) remind us, while there may be broad consensus on goals, there may be disagreement over how the cost of reaching them is shared or even over a prioritization of the goals. What may be "political speech" to one platform user is "hate speech" to another, and vice versa. What one state may consider interference in its electoral system, another state may consider as a normal part of the electoral process. Mutually agreed-upon definitions and standards will be critical in getting to a multistakeholder dialogue on these issues and harmonizing across countries as well as across regulators and the platforms.

15.4 INVEST IN DIGITAL LITERACY

Farah Pandith (Chapter 10) highlights the value of digital literacy and hygiene – possibly through structured education programs. She is also optimistic about the effects of generational change: "digital natives," that is, Gen Alpha, Gen Z, and Millennials, are better attuned to the workings of technology and may be more amenable to interventions that encourage digital literacy.

15.4 INNOVATE IN REGULATIONS

As Bhaskar Chakravorti (Chapter 13) has pointed out, as platforms redirect scarce content moderation resources to a few important markets, such as the US, we could end up with a situation where the rest of the world pays a high price for American democracy. This calls for regulators to think a few steps ahead and anticipate the consequences of their demands. For example, one solution is to have US lawmakers pass laws not only on the nature of the content hosted by the platforms but on the investments that platforms make on content moderation and how these resources are allocated across the world. The regulations ought to require that market-specific

resources be allocated in proportion to the potential risk of disinformation in any given market.

On balance, it is clear that the goal of defeating disinformation will not happen overnight. There is no silver bullet, and it will take a systemic approach to addressing the problem. Despite the many challenges, the components of a solution are also within our grasp. It is essential that we get to a workable set of solutions. The future of democratic societies depends on it.