Future of India Foundation

POLITICS OF

DISINFORMATION

(WHY THE CURRENT APPROACHES ARE GEARED TO FAIL AND POSSIBLE PATH FORWARD)

The Future of India Foundation

EMBARGOED UNTIL MAY 5 (1 PM)

Future of India Foundation

This Page Intentionally Left Blank

Future of India Foundation

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Social media platforms have effectively supplanted traditional information networks in India. While large parts of the country have come online due to one of the cheapest internet costs in the world, the dialectical relationship between online content, traditional news media (especially television) and political networks means that the messages propagated online effectively touch even that section of the population which is not yet online.

This ubiquity could have been a golden moment for India - democratizing access to information, fostering community, increasing citizen participation and reducing the distance between ordinary people and decision-makers. However, social media platforms have adopted design choices which have instead led to a proliferation and mainstreaming of misinformation while allowing themselves to be weaponized by powerful vested interests for political and commercial benefit. The consequent free flow of organised misinformation (disinformation), hate and targeted intimidation has led to real world harm and degradation of democracy in India: anti-minority hate has been mainstreamed and legitimised; communities have become divided and polarised; sowed confusion in the minds of the people; made it difficult to establish a shared foundation of truth; and led to political alienation (see "Findings from FGDs").

It is clear that **organised misinformation (disinformation) has a political and/or commercial agenda**. However, even though there is growing recognition of political motivations and impact of disinformation, the surrounding discourse in India has remained strikingly apolitical and episodic - focused on individual pieces of content and events, and generalised outrage against big tech (see "Significance of India and the Purpose of This Report") instead of locating it in the larger political context and structural design issues. The evolution of the global discourse on misinformation too has allowed itself to get mired in the details of content standards, enforcement, fact checking, takedowns, deplatforming, etc - a framework which lends itself to bitter partisan contest over individual pieces/types of content while allowing platforms to disingenuously conflate

the discourse on moderating misinformation with safeguards for freedom of expression (See "State of Play"). However, these issues are at best adjunct to the real issue of disinformation and this **report conclusively establishes that the current system of content moderation is more a public relations exercise for platforms instead of being geared to stop the spread of misinformation** (see "Politics of Misinformation").

A meaningful framework to combat disinformation at scale must be built on the understanding that disinformation is a political problem. This means that the issue is as much about bad actors as individual pieces of content, and that content moderation – as also content distribution - is an intervention in the political process. The report thus argues for a comprehensive transparency law to enforce relevant disclosures by social media platforms. Moreover, content moderation and its allied functions such as standard setting, fact-checking and deplatforming must be embedded in the sovereign bipartisan political process for democratic legitimacy. At the same time, it is important to ensure that a political approach to misinformation does not degrade into legal sanction for censorship at the behest of the government. Any regulatory body must thus be grounded in democratic principles of diversity, dissent, inclusion, transparency and accountability - its own and of the social media platforms (see "Recommendations").

Given the nature of political polarisation in our country (and most others), the constitution of such a regulator and its operational legitimacy is indeed a tall task. However, the failure of a polarized political ecosystem to come to a workable consensus is not a free pass for the platforms. It is evident that platforms are responsible for the unprecedented speed and spread of distribution of disinformation and the design choices which have made misinformation ubiquitous and indistinguishable from vetted information. It is thus the responsibility of the platforms to tamp down on the distribution of misinformation and weaponization of their platform. This report argues that platforms are sentient about the users and content they are hosting and thus must take ownership of their distribution choices. Further that, just as any action against content is seen as an intervention in the political process, the artificial increase in distribution of content (amplification) too has political and commercial value. We recommend three approaches to distribution that can be adopted by platforms: a hands-off approach to content (and by extension

content creators) to constrain distribution to organic reach (chronological feed); exercise clear editorial choice and take responsibility for amplified content; or amplify only credible sources (irrespective of ideological affiliation). Connected to the third choice, this report argues that the current approach to misinformation which relies on fact-checking a small subset of content in a vast ocean of unreviewed content is inadequate to the task and needs to be supplemented by a review of content creators itself. The report also recommends other allied measures to support and improve the overall information ecosystem in the country such as linking digital literacy outreach to platform's user base[1] such that a critical mass of users undergo digital literacy training in a fixed time period (see Recommendations).

Finally, as the country with the largest youth population in the world, it is important that we actively think of how we want our youth to engage in our democratic processes and the role of social media platforms in it. There are three notable effects of social media on our politics which require deliberation: first, social media has led to a dislocation of politics across the world but especially in India - with people weighing in on abstractions online while being disengaged from their immediate surroundings (the primary site of institutional democratic processes); social media has led to a degradation of our political discourse where serious engagement has been supplanted by "hot takes", memes and emojis; and finally, despite the apparently public nature of social media, providence of some of the most consequential interventions in the political discourse have become obscured because of opacity in technology and platform operations. Meaningful politics, especially in democracies, is rooted in local organisation, discussion and negotiation; however, the structure of social media has facilitated a perception of engagement without organization and action without consequence. This wasn't and isn't inevitable - there are more thoughtful ways to structure social media platforms which would help connect and root people in their own communities instead of isolating them locally while ostensibly "connecting" them in the virtual world. However, it is alarming that instead of moving towards more grounded communities, there is an acceleration towards greater virtuality through metaverse. Social media cannot be wished away; however, its

[1] FOI has outlined multiple approaches for scalable digital literacy programs in a separate report

structure and manner of use are choices we must make as a polity after deliberation instead of accepting as fait accompli or simply being overtaken by developments along the way.

Future of India Foundation

PART 1: SOCIAL MEDIA AND INDIA

INTRODUCTION

Pioneering social media platforms began to emerge in the early 2000s based out of US' Silicon Valley. They took off in the 2010s with users from across the globe signing up in the hundreds of millions or even billions. By offering easy ways to connect with others, share and consume content, social media democratized free expression and made it easy for everyone to access information. By being able to directly engage with events and discourse around the world, users could feel that they were citizens of a borderless world. Social movements leveraged these online social tools to drive political and social revolutions. Charismatic leaders deployed innovative social media campaigns to acquire power. It appeared that there was much to celebrate. In parallel, social media was also upending established information ecosystems. Traditional news media lost their gatekeeping powers on news and information and were further weakened with the shift towards digital advertising, with increasing share of advertising revenue going to major social media platforms. With time, the negative effects of social media platforms have come to the fore. The volume and velocity of information flow has come at the cost of quality. Over time, malicious users and organisations learned to leverage these very same social media tools to sow division, fear and confusion and undermine the integrity of democratic processes. Harmful content, misinformation and disinformation started exploding. Just as the WHO declared COVID-19 as a global pandemic, it also warned of an 'infodemic'[2].

The rise in abuse of social media forced platforms to respond with mitigation measures. The defining feature of such responses has been their American centricity. This isn't surprising since

[2] "WHO Publishes Public Health Research Agenda for Managing Infodemics," February 2, 2021. https://www.who.int/news/item/02-02-2021-who-public-health-research-agenda-for-managing-infodemics.

most of the world's dominant social media platforms are headquartered in the United States. American free speech exceptionalism approach to content moderation efforts by these platforms, the alacrity of response to developments in the US when compared to the rest of the world, the lopsided allocation of investments to combat abuse against American users vis-a-vis users elsewhere are all a manifestation of this America-first approach. In essence, these platforms have de-facto imposed American standards of speech on a diverse world including on societies with fragile social fabrics and low institutional capacity to respond to an 'infodemic' without the concomitant mitigation measures and responsiveness reserved for their home country.

SIGNIFICANCE OF INDIA AND PURPOSE OF THIS REPORT

Much of the in-depth media reportage and cutting-edge academic research of the harmful impact of social media and regulatory frameworks have emerged from Western countries. This is partly a natural consequence of the greater formalisation of economies, institutionalisation of politics and governance in these countries as well as co-location of tech, media and academic networks in the region. But India's absence from this landscape is also reflective of low political and policy prioritisation to combat disinformation in India. As countries around the world grapple with the power of social media platforms and their influence on democratic discourse, this knowledge deficiency between western countries and other countries, especially the global south, is untenable. India especially is a test case with opportunity for rich insights for other countries: world's largest democracy, high heterogeneity across multiple axes, largest youth population in the world and home to one of the largest global internet user bases in general and the largest social media user bases for several platforms specifically. It is thus important to develop a better understanding of the impact of social media platforms in India and their operations to propose solutions, which would not just deepen Indian democracy and provide a safer user experience for Indians but also provide a template for the rest of the world.

There is poor public awareness and understanding in India of how social media platforms regulate and distribute content. Any review of platform operations tends to be episodic and concentrated largely on government content takedown requests or platforms failure to police specific kinds of

content without necessarily probing the underlying structure and factors influencing platform behaviour. A lack of systemic, in-depth study of how social media platforms operate in India essentially allows platforms to escape accountability for their impact in India. This is evident in the differential response by social media platforms to elections in the United States and India[3]. Further, the discourse is overly focused on platform regulation from a legal and regulatory perspective and insufficiently informed about how platforms self-regulate (or don't).

This report seeks to provide clarity on the key issues which are important to understand the problem of contemporary disinformation today while facilitating broader public awareness of how platforms govern and distribute speech, impact on India and the conceptual constraints of the current discourse on this issue. Our hope is that this report will facilitate wider and more substantive conversations on these issues so that we can collectively find ways to improve public discourse and strengthen democracy in India and around the world.

METHODOLOGY, DEFINITIONS AND SCOPE

This report brings together ground research on the use and harmful impact of social media in India with a comprehensive review and analysis of current research and thinking on measures to mitigate harmful and false information online.

The Future of India Foundation conducted focus group discussions with youth across India to understand their social media habits and the impact of harmful content and misinformation on their

[3] Ananth, Venkat, and Bhardwaj, Deeksha. "Facebook Convinced Poll Panel to Settle on a Voluntary Code." Hindustan Times, November 22, 2021. https://www.hindustantimes.com/india-news/facebook-convinced-poll-panel-to-settle-on-a-voluntary-code-101637549263683.html.

lives. The focus on youth was deliberate: 70% of India's internet users are below the age of 35 years and this demographic also constitutes 65% of the total population in India. The experience of Indian youth on social media is thus representative not just of India's experience but will also define its future.

We also conducted a comprehensive review of the evolution of content moderation frameworks of major social media platforms based on academic research and media reportage to lay out the current state of thinking on content moderation and the emerging discourse.

Finally, we analyze the conceptual constraints in the current approach to mitigating online misinformation to make recommendations for the way forward keeping in mind the context and challenges in India. The recommendations suggest an altogether new approach to combating misinformation without impinging on freedom of expression and also make suggestions to improve efficacy of the current approach to misinformation.

For the purpose of this report, we define

- **Harmful content** as a category of content that is generally understood to be harmful and can clearly be defined. Examples of such content would include graphic violence, child abuse material, nudity, harassment, etc.
- **Misinformation** as a category of content that contains false or misleading information that may or may not contain harmful content and may or may not have been distributed with intent to mislead
- **Disinformation** as deliberate misinformation that is propagated in an organized manner

This report is focused on user-generated content and does not look at advertisements. Social media advertising and issues around it - such as micro-targeting, privacy concerns, lack of fact-

checking of political advertisements and lack of overall transparency in this space - are a separate area of research and have not been covered in this report.

FOCUS GROUP DISCUSSIONS

The Future of India Foundation conducted focus group discussions (FGD) with youth across India[4] to understand how youth in India engages with and consumes information (including misinformation and disinformation) online. The discussion focused on ascertaining the following:

1. how young people get and consume information
2. how they determine which information is trustworthy
3. how they sift between competing narratives on the same event/issue
4. do they care to ascertain whether a piece of information is accurate
5. the purpose and use of information
6. awareness of and reliance on fact-checking sites
7. impact of online misinformation

The topic used to guide the discussion was COVID. This topic was selected for three reasons: near universal awareness of the topic; high level of reportage and engagement on this topic both online

[4] The FGDs were done with participants from 25 districts in the states of Uttar Pradesh, Rajasthan, Uttarakhand, Haryana, Delhi, Madhya Pradesh, Andhra Pradesh and Telangana between the months of October 2021 to January 2022. The population covered was both urban and rural youth between the ages of 18-35. Our partner organisations include Awadh People Forum, Badlav, Bewajah, Gramya, Vanangana, Yeh Ek Soch Foundation, Azad Shiksha Kendra, Swatantra Talim, PACE, Mahila Swarozgar Samiti

and offline; and lack of clearcut partisan affiliation which would derail discussions on other mainstream political issues.

Each meeting lasted between 1-2 hours depending on the number of participants. The questions for discussion are included in the annexure.

FINDINGS FROM FOCUS GROUP DISCUSSIONS (FGDS)

The key takeaway from the focus group discussions is that not only have social media platforms disrupted the information ecosystem in India, but that they have allowed themselves to be weaponized by vested interests in ways which are leading to real world harm without investing in meaningful safeguards.

SOCIAL MEDIA PLATFORMS: FROM DISTRIBUTORS TO SOURCE OF INFORMATION[5]

Social media platforms have become portals to the internet and are the dominant source for information for all users who have smartphones. TV news in regional languages is the only source of information and news which is more dominant than social media. Majority of users are passive consumers of information who respond to notifications for content curated for them in their feed (by the platform) or through posts in their network. Even active users who demonstrate considerable user agency in their information consumption habits invariably lapse into passive consumption of information through their platform feed. Most users do not differentiate between

[5] For Users with Smartphones

"news[6]" and personal/partisan posts due to low digital literacy and low credibility of institutional media. Participants repeatedly responded to questions about the source of some referenced information by saying, "<Name of Platform> par dekha tha" ("I read/saw it on the <name of platform>") indicating their inability to differentiate between the platform and source of information.

RESIGNATION AND POLITICAL ALIENATION

There's a sense of being overwhelmed with competing narratives and resignation that truth about a contested issue cannot be ascertained. One user encapsulated this with, "sach batayein, hum to confuse hokar reh jaate hai" ("we are unable to determine what is the truth, we remain confused"). Moreover, there's a widespread sense that even if truth could be established, it would neither change outcome nor users' own behavior in many instances due to limited options for recourse. This sense of helplessness is leading to disregard for truth itself and/or alienation from the larger political eco-system.

NARRATIVES AND DIRECTIONAL TRUTH STAND-IN FOR ABSOLUTE TRUTH

As a coping mechanism to the general miasma of uncertainty, large numbers of users are interested only in confirming and/or verifying only as much information which has narrative value without being too bothered about absolute facts. This was seen in responses related to COVID mortality numbers where there was little interest in ascertaining exact numbers beyond establishing or refuting the narrative that official numbers were false and artificially deflated. This was evident too in various anti-minority narratives (such as Muslim food and vegetable vendors spitting on food) doing the rounds where alleged individual events were less important than the overall narrative.

[6] Reported content with two characteristics: clear identification and ownership by reporting agency; and produced through an editorial process based on accepted standards

DIGITAL LITERACY NOT ENOUGH IMPETUS TO FACT-CHECK

User information consumption behavior can be categorised as passive (where source and/or content of information is pushed to the recipient) and active (where information is sought by user through proactive selection of source and/or topic). Most users fall in the passive user category and demonstrate little impetus to fact-check received information irrespective of their level of digital literacy. This manner of passive consumption of information reinforces broad narratives with limited attention to detail. Source of content is a significant driver of how the information is received - users are more receptive to content, especially which runs counter to their worldview - when received through trusted affiliates. This was seen in responses such as "agar meri vichaar dhaara se jude huye vyakti se news aati hai, to main usse zyada importance deta hoon" ("I prioritise information from people aligned to my way of thinking and/or worldview")

MISINFORMATION HAS A POLITICAL AND/OR COMMERCIAL AGENDA

During the FGDs, we identified multiple kinds of misinformation about COVID such as how COVID spread, efficacy of vaccines, cures for COVID, effectiveness of Government response and deaths due to COVID. Virtually all misinformation could be linked to narratives propagated by organised political and business entities instead of existing in isolation. We also observed a dialectical relationship between narratives fueled by misinformation, representatives of political organisations' and allied information ecosystems such that the foundation set by misinformation was used for additional layers of misinformation.

NATIONALISED DISCOURSE; LOCALISED HATE AND DIVISION

Disinformation narratives have displaced local/personal issues as the agenda for discussion within communities leading to palpable tensions between opposing groups in a local area. For instance, participants cited how manufactured narratives against Muslims pertaining to their role in spreading COVID-19 had permeated local discourse to create divisions between Muslims and

other communities. Examples include the demonisation of minorities by alleging that Tablighi Jamaat, a Muslim religious group was responsible for spreading COVID because of their congregation in Delhi before India's first COVID-19 lockdown. Participants in Haryana also reported that a Muslim vegetable vendor was not allowed to sell vegetables because of online messages alleging that Muslim food vendors were spitting on food, etc. Other non-COVID examples include narratives based on hyper-nationalism which similarly paint Muslims as lacking in patriotism (various versions of Indian Muslims supporting Pakistan, an Islamic republic at odds with India).

SOCIAL MEDIA PLATFORMS LAG IN DIGITAL LITERACY AND FACT-CHECKING EFFORTS

In all the discussions, participants who demonstrated above-average digital literacy skills acquired the relevant context and skills through their education (such as journalism students) and/or group affiliation (participation in civic groups) instead of any initiative by social media platforms. This is unsurprising since social media platforms have not created mass-touch digital literacy initiatives. Digitally literate and politically engaged users conduct their own fact-checking instead of relying on social media platforms. One reason for this is because fact-checking efforts are confined to a small subset of content and there is little fact-checked content in many Indian regional languages either because of a lack of enough professional fact-checking organizations or because of a lack of investments by platforms in the fact-checker ecosystem.

Users conduct their own fact-checking through multiple ways[7]: triangulation - the most common way to verify information was to seek multiple reports about the same topic through a variety of sources for additional information and context. Some common ways included reading multiple news reports, reading comments for counterviews, looking to affiliated groups for additional perspective etc; lived experience/context - if the received information is related to one's own

[7] No assertion is being made on the efficacy or accuracy of independent fact-checking efforts by users other than to point out that the subset of users interested in ascertaining the truth largely have to rely on personal initiative

experience or context of the individual, the veracity of information is ascertained through extrapolation. For instance, users repeatedly said that the reason they believed government numbers on COVID were false was because official data widely suppressed official deaths in their own neighborhood; personal contacts - for news reports about the same or neighbouring states, people often relied on personal contacts to get additional information indicating low trust in intermediary news sources. This method was also relied upon for hyper local news which had limited coverage in bigger news outlets. This reliance on own efforts is indicative of lack of institutional channels where users can reliably seek fact-check/verification for particular information

IMPACT OF SOCIAL MEDIA IN INDIA

Social media usage has grown exponentially in India over the last decade. Aggressive growth tactics adopted by social media platforms such as Facebook's abandoned 'Free Basics' program coupled with India's booming telecommunication industry that lowered data and smartphone device costs to affordable levels have provided the necessary tailwinds. Most social media platforms including Facebook, Youtube, Instagram, and Twitter registered impressive increases in user base especially since 2014. By 2020, over 50% of India's population was accessing social networks[8].

Extensive use of social media for political campaigns, especially in the 2014 General Elections[9] established these platforms as de facto public squares. However, abuse of social media in India took off in parallel with its growth and increasing prominence. Abuse against women to organized

[8] Statista. "India: Social Network Penetration 2025." Accessed March 14, 2022. https://www.statista.com/statistics/240960/share-of-indian-population-using-social-networks/.

[9] Financial Times. "Narendra Modi to Be India's First Social Media Prime Minister," May 23, 2014. https://www.ft.com/content/e347de5c-e088-11e3-9534-00144feabdc0.

attacks by partisan groups against dissidents or opponents became common[10]. More recently, misinformation spread on social media platforms and private messaging services has led to a spate of mob lynchings[11]. Moreover, organised political entities have mobilised online through dedicated political party "IT cells" in effect hijacking the online discourse for partisan ends. Islamophobic narratives like 'Corona Jihad', 'Love Jihad' are now part of frequently trending conversations on social media platforms.

Social media platforms are cognizant of the impact. For instance, Facebook's internal documents leaked by whistleblower Frances Haugen confirm that Facebook's internal research showed increased prevalence of inflammatory content in late 2019 and early 2020. This was around the same time as anti-CAA protests, Delhi elections and the subsequent Delhi riots that led to over 53 deaths[12] indicating a dialectical relationship between anti-minoritism online and majoritarian activity offline.

However, platforms' response to growing harms in India have been marked by a lack of urgency and response concomitant to the seriousness of impact. Over the years, Facebook announced the addition of 'caste' as a protected characteristic to its hate speech definition and offered users the ability to lock their profile from being publicly accessible. However, these cosmetic changes to policies and products do not compare favourably with the alacrity and breadth of response by platforms when it comes to harms in the US. Platforms rolled out various election integrity measures ahead of US elections in 2020, adapted and rolled out new tools and policies

[10] BBC News. "Why Are Indian Women Being Attacked on Social Media?," May 8, 2013, sec. India. https://www.bbc.com/news/world-asia-india-22378366; "The Power of Social Media: Emboldened Right-Wing Trolls Who Are Attempting an Internet Purge." Accessed March 14, 2022. https://caravanmagazine.in/vantage/power-social-media-emboldened-right-wing-trolls.

[11] "On WhatsApp, Rumours, and Lynchings | Economic and Political Weekly," February 9, 2019. https://www.epw.in/journal/2019/6/insight/whatsapp-rumours-and-lynchings.html.

[12] Srivas, Anuj. "Facebook Saw Spikes in Hate Speech in India After CAA Protests and During First Covid Lockdown." The Wire, September 10, 2021. https://thewire.in/tech/facebook-saw-spikes-in-hate-speech-in-india-after-caa-protests-and-COVID-19-lockdown.

continuously during and after the elections. The storming of the US Capitol on Jan 6th elicited an instant response with a catalogue of actions and changes from platforms[13] including the suspension and subsequent deplatforming of former US President Trump and his allies from major social media platforms[14]. One would be hard pressed to find instances of similar proactivity, alacrity and sustained response by platforms to crises in Global South countries. The lack of prioritization by platforms to harms suffered by users in the Global South became apparent through Haugen leaks. Middle Eastern and Central Asian countries had minimal or faulty integrity measures in place[15]. 87% percent of Facebook's global budget for classifying misinformation is earmarked for the United States, while only 13% is allocated for the rest of the world even though North American users make up only 10% of the social network's daily active users[16]. Facebook rolled out machine learning classifiers to detect hate speech in Hindi and Bengali only in 2018 and 2021 respectively while similar technology for English was in force much earlier[17]. YouTube has been accused of taking a hands-off approach to Myanmar 2020 elections even as it rolled out mitigation measures against harmful content proactively and reactively for the US 2020 elections held just 5 days earlier[18].

Failure of platforms to moderate harmful content and misinformation is being made worse by their capitulation to political pressures to gain or protect access to prized markets around the world.

---

[13] Rosen, Guy and Bickert, Monika. "Our Response to the Violence in Washington." *Meta* (blog), January 7, 2021. https://about.fb.com/news/2021/01/responding-to-the-violence-in-washington-dc/.

[14] Washington Post. "These Are the Platforms That Have Banned Trump and His Allies." Accessed March 14, 2022. https://www.washingtonpost.com/technology/2021/01/11/trump-banned-social-media/.

[15] Scott, Mark . "Facebook Did Little to Moderate Posts in the World's Most Violent Countries- POLITICO," October 25, 2021. https://www.politico.com/news/2021/10/25/facebook-moderate-posts-violent-countries-517050.

[16] Frenkel, Sheera, and Davey Alba. "In India, Facebook Grapples With an Amplified Version of Its Problems." *The New York Times*, October 23, 2021, sec. Technology. https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html.

[17] "Facebook Failing to Check Hate Speech, Fake News in India: Report," October 25, 2021. https://www.aljazeera.com/news/2021/10/25/facebook-india-hate-speech-misinformation-muslims-social-media.

[18] Potkin, Fanny. "YouTube Faces Complaints of Lax Approach on Overseas Election Misinformation." *Reuters*, December 18, 2020, sec. Media and Telecoms. https://www.reuters.com/article/us-youtube-myanmar-misinformation-idUSKBN28S0QE.

Facebook has repeatedly given into demands from governments in Asia on content restrictions[19]. Its tools have been used by government allies to push propaganda and target dissenters. News reports indicate that Facebook did not enforce its own policies against hate speech and dangerous individuals in India when actors from the ruling party were implicated[20]. In the run up to 2019 Indian national elections, Facebook removed accounts belonging to both BJP and Congress parties. However, the announcement mentioned only the opposition Congress while omitting reference to the ruling party, the BJP[21].

It is clear that social media platforms have an outsized political impact on democratic processes in India and we must ensure that this impact deepens our democracy instead of detracting from it.

PART 2: STATE OF PLAY

CURRENT APPROACH

FREE SPEECH AS A BUSINESS MODEL

Social media platforms achieved unbridled growth by adopting a laissez faire approach to user-generated content. This approach was underpinned by legislative frameworks around the world which sought to indemnify content hosting internet intermediaries from liability arising out of

[19] Yifan Yu, Kiran Sharma, Lien Hoang, Cliff Venzon, and Erwida Maulia. "Losing Face: The Perils of Facebook's Asia Strategy." Nikkei Asia, July 28, 2021. https://asia.nikkei.com/Spotlight/The-Big-Story/Losing-Face-the-perils-of-Facebook-s-Asia-strategy.

[20] Horwitz, Newley Purnell and Jeff. "Facebook's Hate-Speech Rules Collide With Indian Politics." *Wall Street Journal*, August 14, 2020, sec. Tech. https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346.

[21] Sircar, Sushovan. "Facebook Removes 702 Pages Related To Cong and BJP-Linked IT Firm." Accessed March 14, 2022. https://www.thequint.com/elections/facebook-removes-702-pages-congress-bjp-linked-it-firm-silver-touch.

user-generated content to varying extents subject to conditions[2223]. In several developing and underdeveloped countries, social media ended up becoming the internet[2425] with the majority of users first coming online through devices and connectivity facilitated by Big Tech companies. With increasing size and socio-political influence, platforms have come under pressure to expand their content moderation operations[26].

The earliest approach to moderating content by platforms was to remove content based on a set of high-level standards set by the companies themselves[27]. These have evolved over the years into detailed rules spanning a multitude of categories of content such as nudity, graphic violence, hate speech, etc. Content that falls foul of these rules is removed through a combination of technology and human review[28]. Given that all the major social media platforms were primarily based in the US, they adopted an American approach to content moderation. While Section 230 of the US' Communication Decency Act (CDA) protects platforms from liability for content posted by their users, it also has a "good samaritan provision" that provides the underlying legal justification for platforms to place restrictions on harmful speech on their platform. However, platforms have used this provision to restrict only stringent and narrowly defined categories of content[29] for two

---

[22] Douek, Evelyn. "Governing Online Speech" Columbia Law Review, April 2021. https://columbialawreview.org/wp-content/uploads/2021/04/Douek-Governing_Online_Speech-from_Posts_As-Trumps_To_Proportionality_And_Probability.pdf

[23] Ashley Johnson, Daniel Castro. "How Other Countries Have Dealt With Intermediary Liability". ITIF, February 21 2021 https://itif.org/publications/2021/02/22/how-other-countries-have-dealt-intermediary-liability

[24] Mirani, Leo. "Millions of Facebook users have no idea they're using the internet: Quartz, February 2015. https://qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/

[25] Asher, Saira. "Myanmar coup: How Facebook became the 'digital tea shop'". BBC News, February 2021. https://www.bbc.com/news/world-asia-55929654

[26] Klonick, Kate. "The New Governors: The People, Rules and Processes Governing Online Speech". Harvard Law Review, April 2018. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

[27] Ibid

[28] Ibid; Jeong, Sarah. "The History of Twitter's Rules." *Vice* (blog), January 14, 2016. https://www.vice.com/en/article/z43xw3/the-history-of-twitters-rules.

[29] Douek, Evelyn. "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability." *SSRN Electronic Journal*, Arpil 2021. https://doi.org/10.2139/ssrn.3679607.

reasons: a pro-speech culture fostered by the American First Amendment; and the inevitable politics and difficulty of content moderation as a whole. Platforms have similarly sought recourse to First Amendment principles to reject calls for a more interventionist approach towards misinformation.

Twitter, YouTube, Facebook are all on the record stating their aversion to be the 'arbiters of truth' and that the platforms should be a marketplace of ideas. This is in line with the US Supreme Court's pronouncements on First Amendment related cases such as "some false statements are inevitable if there is to be an open and vigorous expression of views" [United States v. Alvarez, 567 U.S. 709, 718 (2012)] and "under the First Amendment there is no such thing as a false idea" [Gertz v. Robert Welch, Inc., 418 U.S. 323 (1974)]. Others have contested this approach arguing that social media platforms are private companies and not public squares and removal of content by them is not akin to state censorship [30]. There is also criticism that platforms have opportunistically used the restrictions placed by the First Amendment on government speech regulation and the protection against liability offered by Sec 230 to advance their business models while failing to ensure a good information ecosystem[31].

However, various socio-political developments around the world which were seen to be fueled by misinformation such as the election of Rodrigo Duterte in the Philippines, Rohingya genocide in Myanmar, Brexit and most prominently, the election of Donald Trump in America, put platforms on the backfoot. Platforms were forced to significantly expand their content moderation operations to act on misinformation which would not be removed as per their content standards and rules up until recently. The ensuing public scrutiny and pressure led to platforms quickly scaling up the

[30] Swisher, Kara. "It's Time for Social Media Platforms to Permanently Ban Trump." Intelligencer, January 7, 2021. https://nymag.com/intelligencer/2021/01/its-time-for-social-media-platforms-to-ban-trump-forever.html.

[31] P. Goodman, Ellen. "The First Amendment Opportunism of Digital Platforms." GMFUS, February 11, 2019. https://www.gmfus.org/news/first-amendment-opportunism-digital-platforms.

then nascent fact-checking programs as a way to address the problem of misinformation[32]. Platforms partnered with independent fact-checking organizations to avoid being the 'arbiters of truth'. This served to buffer platforms from charges of bias while helping to create new institutions and/or empower existing institutions that people/users would trust to provide credibility to fact-checking on platforms. Platforms have for long used algorithms to determine content distribution based on engagement[33]; however they now tweaked their algorithms to reduce distribution for the subset of content selected for fact-checking and found to be substantively false and/or misleading. This approach to moderating misinformation ironically underscores algorithmic distribution as a key vector for amplification of misinformation. Besides reduction in distribution, adding labels to alert users and surfacing additional fact-checked information alongside content deemed as misinformation were some additional measures adopted by platforms to combat misinformation. **Thus the content moderation toolkit cleaved into two approaches: outright removal of harmful content and fact-checking with consequent mitigating measures for misinformation**. This toolkit succinctly referred to as 'Remove, Reduce, Inform' by Facebook[34] has become the current content moderation approach by other big platforms like YouTube[35] and Twitter as well. Platforms continue to remain diffident when it comes to misinformation that may not violate their rules on harmful content though the COVID-19 pandemic and subsequent global public health emergency seems to be changing this to an extent. A short review of each aspect of this approach and the discourse around them is given below.

APPROACH TO HARMFUL CONTENT

[32] Bell, Emily. "The Fact-Check Industry." Columbia Journalism Review, 2019. https://www.cjr.org/special_report/fact-check-industry-twitter.php/.

[33] Meta. "Our Approach to Ranking | Transparency Centre," January 19, 2022. https://transparency.fb.com/en-gb/features/ranking-and-content/.

[34] Lyons, Tessa. "The Three-Part Recipe for Cleaning up Your News Feed." *Meta* (blog), May 22, 2018. https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/.

[35] YouTube. "The Four Rs of Responsibility, Part 1: Removing Harmful Content." blog.youtube, September 3, 2019. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-remove/.

Platforms take a relatively hardline approach to content they consider to be harmful or not in line with their value proposition[36]. Such content is removed and penalties are imposed on responsible users.

- **Content Removals:** Platforms have public facing rules that describe the kind of speech and behaviour prohibited by them. These rules evolve rapidly to keep pace with the changing norms of their vast user bases and adversarial tactics of malicious actors[373839]. The changes are influenced by multiple interest groups including government pressure, media coverage, civil society advocacy and user feedback[40]. Repeated violations of these rules by users lead to content removal and escalating penalties beginning with a mere warning and leading all the way up to removing the ability monetise content or advertise to even account removals[41].

- **De-platforming:** Platforms may also take into account offline behaviour of certain users/organizations to ban them from using their services in perpetuity, a measure known as 'deplatforming'. Arguably the most high-profile deplatforming was the ban on US President Donald Trump by all major social media platforms after his supporters laid siege

---

[36] Facebook for instance has a hardline approach to nudity and pornography whereas Twitter is more permissive about such content

[37] 'Note' in The Twitter rules: safety, privacy, authenticity, and more. https://help.twitter.com/en/rules-and-policies/twitter-rules

[38] How we update the Facebook Community Standards. https://transparency.fb.com/en-gb/policies/improving/deciding-to-change-standards/

[39] YouTube Community Guidelines and policies - How YouTube Works. "YouTube Community Guidelines and Policies - How YouTube Works." Accessed March 14, 2022. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/

[40]Klonick, Kate. "The New Governors: The People, Rules and Processes Governing Online Speech". Harvard Law Review, April 2018. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/

[41] YouTube Community Guidelines and policies - How YouTube Works. "YouTube Community Guidelines and Policies - How YouTube Works." Accessed March 14, 2022. https://www.youtube.com/howyoutubeworks/policies/community-guidelines/.;

"Restricting Accounts | Transparency Centre," February 11, 2022. https://transparency.fb.com/en-gb/enforcement/taking-action/restricting-accounts/ ;

"Disabling Accounts | Transparency Centre," January 19, 2022. https://transparency.fb.com/en-gb/enforcement/taking-action/disabling-accounts/.

to the US Capitol in a bid to overturn the Presidential election. Trump's deplatforming had a discernible impact on election related discourse on platforms: according to Zignal Labs, a social media measurement company, there was a 73% decline in mentions of election fraud on Twitter the week after his ban when compared to the week before[42]. The efficacy of deplatforming in reducing the impact of extreme influencers and toxicity on social media platforms has been known for a few years now[43][44] and emerging research has reinforced it[45].

APPROACH TO MISINFORMATION

1. **Fact-Checking:** The most prominent effort by platforms to combat misinformation is fact-checking. It is important to note however that fact-checking as an approach to combat misinformation applies only to a tiny subset of content actually selected for fact-checking by the platforms or the independent fact-checking organisations. Content to be fact-checked could be selected by different methods including feedback from platform users, virality metrics or at the editorial discretion of independent fact-checkers. At its core, fact-checking relies on reactive measures to counter misinformation. These programs involve rating the veracity of news articles and other information based on investigation by human fact-checkers. Such ratings are then fed into the platforms to reduce the distribution of matching content and/or inform users of the veracity of the content through labels.

[42] Bloomberg.com. "Trump's Ban Has Already Had an Impact on the Social Media Landscape," January 21, 2021. https://www.bloomberg.com/news/newsletters/2021-01-21/trump-s-ban-has-already-had-an-impact-on-the-social-media-landscape.

[43] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech. Proc. ACM Hum.-Comput. Interact. 1, 2, Article 31 (November 2017). https://doi.org/10.1145/3134666

[44] Vice.com. "Deplatforming Works," October 2018. https://www.vice.com/en/article/bjbp9d/do-social-media-bans-work

[45] Jhaver, Shagun, Christian Boylston, Diyi Yang, and Amy Bruckman. "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 381:1-381:30. https://doi.org/10.1145/3479525.

- **Reduction in Distribution:** Algorithmically driven reduction in distribution is primarily discussed in the context of fact-checking programs launched by various platforms. Facebook partners with third party fact-checkers to rate the veracity of content leading to reduced distribution and addition of fact-check labels to content. The selection of such fact-checking partners has also come under some controversy [46] . Google surfaces fact-checks against news articles from algorithmically determined sources of authoritative information[47]. Twitter's fact-checking program is less well-documented [48] but after a short period of fact-checking internally, it built a tool to crowdsource authoritative information[49] before eventually outsourcing fact-checking work like Google and Facebook[50].

- **Labeling and informing:** Platforms also use fact-checks to inform users about the veracity of content through correction labels. Platforms deploy such labels only on the subset of content which is fact-checked by independent fact-checkers. Facebook claims that 95% of the time people don't click past correction labels to view the

[46] Newton, Casey. "A Partisan War over Fact-Checking Is Putting Pressure on Facebook." The Verge, September 12, 2018. https://www.theverge.com/2018/9/12/17848478/thinkprogress-weekly-standard-facebook-fact-check-false.

[47] "Our Approach to Misinformation | Transparency Centre." Accessed March 14, 2022. https://transparency.fb.com/en-gb/features/approach-to-misinformation/.; Justin Kosslyn and Cong Yu. "Fact Check Now Available in Google Search and News around the World." Google, April 7, 2017. https://blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/.

[48] Reuters. "With Fact-Checks, Twitter Takes on a New Kind of Task," May 30, 2020, sec. Media Industry. https://www.reuters.com/article/us-twitter-factcheck-idUSKBN2360U0.h Coleman. "Introducing Birdwatch, a Community-Based Approach to Misinformation," January 25, 2021. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.;

Coleman. "Introducing Birdwatch, a Community-Based Approach to Misinformation," January 25, 2021. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.;

Mantas, Harrison . "Twitter Finally Turns to the Experts on Fact-Checking." *Poynter* (blog), August 5, 2021. https://www.poynter.org/fact-checking/2021/twitter-finally-turns-to-the-experts-on-fact-checking/.

[49] Coleman. "Introducing Birdwatch, a Community-Based Approach to Misinformation," January 25, 2021. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

[50] Mantas, Harrison. "Twitter Finally Turns to the Experts on Fact-Checking." *Poynter* (blog), August 5, 2021. https://www.poynter.org/fact-checking/2021/twitter-finally-turns-to-the-experts-on-fact-checking/.

underlying content[51]. Correction labels have seen adoption by major platforms over the past few years but redirecting users to authoritative sources of information as a strategy to counter misinformation has seen adoption more recently during the ongoing COVD-19 pandemic[52], and the US2020 elections[53]. Facebook launched hubs to provide authoritative information to users on certain topics like Climate Change, COVID-19 & US 2020 Elections[54]. Facebook claims to have connected over 2 billion people from 189 countries to reliable information about the coronavirus through their COVID-19 Information Center and messages[55]. Labeling and corrections are seen as a proportionate and less restrictive means to mitigate misinformation while also staying true to the American First Amendment interpretation that the best way to counter bad speech is through more good speech[56].

2. **Prebunking Measures:** Fact-checking is by design a reactive measure. Prebunking, a proactive measure, relies on the process of inoculation. Inoculation against misinformation is similar to inoculation against diseases. Prebunking is premised on the idea that if people are forewarned that they might be misinformed and are exposed to weakened examples of the ways in which they might be misled, they will become more immune to

[51] Rosen, Guy. "How We're Tackling Misinformation Across Our Apps." *Meta* (blog), March 22, 2021. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/.

[52] Clea Skopeliti and Bethan John. "Coronavirus: How Are the Social Media Platforms Responding to the 'Infodemic'?" First Draft, March 19, 2020. https://firstdraftnews.org:443/articles/how-social-media-platforms-are-responding-to-the-coronavirus-infodemic/.

[53] Reuters. "Twitter, Facebook Outline Action on Posts Claiming Early U.S. Election Victory," November 2, 2020, sec. U.S. Legal News. https://www.reuters.com/article/us-usa-election-twitter-idUSKBN27I1UX.

[54] Foo Yun Chee and Katie Paul. "Facebook Launches Climate Science Info Center amid Fake News Criticism." *Reuters*, September 15, 2020, sec. APAC. https://www.reuters.com/article/facebook-climatechange-int-idUSKBN2660M5.

[55] Meta. "Reaching Billions of People With COVID-19 Vaccine Information," February 8, 2021. https://about.fb.com/news/2021/02/reaching-billions-of-people-with-covid-19-vaccine-information/.

[56] Douek, Evelyn. "Governing Online Speech: From 'Posts-As-Trumps' to Proportionality and Probability." *SSRN Electronic Journal*, April 2021. https://doi.org/10.2139/ssrn.3679607.

misinformation[57]. There is mixed evidence of the efficacy of prebunking measures against misinformation[58]. Twitter rolled out a pre-bunking measure against misinformation related to Climate Change during the ongoing COP26 Conference[59].

Approach to Borderline Content: Less discussed but perhaps more critical is the use of algorithmic distribution by platforms to tamp down toxic content that isn't misinformation but is on the borderline of platform rules. Facebook outlined their approach to such borderline content in a note by CEO Mark Zuckerburg in late 2018[60] wherein borderline content is penalised so it gets less distribution and engagement. YouTube announced the launch of similar interventions in early 2019 [61] with reduced recommendations for borderline content. Transparency around algorithmic interventions are far lesser than that around content standards or removals[62] and it appears that such interventions on algorithmic distribution have been largely confined to certain

[57] Lewandowsky, Stephan, and Sander van der Linden. "Countering Misinformation and Fake News Through Inoculation and Prebunking." *European Review of Social Psychology* 32, no. 2 (July 3, 2021): 348–84. https://doi.org/10.1080/10463283.2021.1876983.

[58] Coldewey, David. "Debunk, Don't 'Prebunk,' and Other Psychology Lessons for Social Media Moderation." *TechCrunch* (blog). Accessed March 14, 2022. https://social.techcrunch.com/2021/01/25/debunk-dont-prebunk-and-other-psychology-lessons-for-social-media-moderation/.; Linden, Sander van der, Jon Roozenbeek, Rakoen Maertens, Melisa Basol, Ondřej Kácha, Steve Rathje, and Cecilie Steenbuch Traberg. "How Can Psychological Science Help Counter the Spread of Fake News?" *The Spanish Journal of Psychology* 24 (April 12, 2021). https://doi.org/10.1017/SJP.2021.23.

[59] "#COP26 Is Happening on Twitter." Accessed March 14, 2022. https://blog.twitter.com/en_us/topics/company/2021/-cop26-is-happening-on-twitter.

[60] Zuckerberg ,Mark. "A Blueprint for Content Governance and Enforcement," May 6, 2021. https://www.facebook.com/notes/751449002072082/.

[61] blog.youtube. "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation." Accessed March 14, 2022. https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/.

[62] Douek, Evelyn. "What Facebook Did for Chauvin's Trial Should Happen All the Time." The Atlantic, April 21, 2021. https://www.theatlantic.com/ideas/archive/2021/04/facebook-should-dial-down-toxicity-much-more-often/618653/.

"at-risk countries" or high risk situations[6364]. Implementation too is sketchy: according to an internal Facebook memo in 2019, as much as 40 percent of "borderline content" in India that was ignored by Facebook's contractual moderators included "nudity", "violence" and "hate" and generated 10 times as much views as outright violative content[65].

Platforms also predominantly use automation technologies to remove bots and fake accounts used to spread dis/misinformation as well as other kinds of prohibited activity e.g., spam.

GOVERNMENT REGULATION

In response to flailing efforts by platforms at reigning in online misinformation and its deleterious impact on society, governments around the world have kicked off multiple legal and regulatory initiatives comprising of different approaches to stem the tide of online misinformation[66]. These initiatives can roughly be categorized as: legislation against misinformation; media literacy to help citizens become more discerning about their news consumption; government task force to combat online misinformation; encouraging self-regulation by platforms; internet shutdowns (as ad-hoc measure); legal action against false information; high-level government reports; misinformation codes (non-statutory measures as agreed between government and online platforms). Some of the common issues with these regulatory measures include: vague or arbitrary definition of

[63] Miranda Sissons. "Our Approach to Maintaining a Safe Online Environment in Countries at Risk." *Meta* (blog), October 23, 2021. https://about.fb.com/news/2021/10/approach-to-countries-at-risk/.

[64] Constine, Josh. "Facebook Will Change Algorithm to Demote 'Borderline Content' That Almost Violates Policies." *TechCrunch* (blog), September 16, 2018. https://social.techcrunch.com/2018/11/15/facebook-borderline-content/.

[65] The Indian Express. "'Borderline' Problematic Content Faster, Slips under Radar: Facebook Memo," November 16, 2021. https://indianexpress.com/article/india/facebook-zuckerberg-hate-speech-and-violence-7624665/.

[66] Daniel Funke and Daniela Flamini. "A Guide to Anti-Misinformation Actions around the World." *Poynter* (blog). Accessed March 14, 2022. https://www.poynter.org/ifcn/anti-misinformation-actions/.

misinformation (legislation); lack of meaningful action (self-regulation); government abuse and censorship (legal action); and inability to achieve political consensus on a way forward.

LIMITATIONS OF CURRENT APPROACH TO HARMFUL CONTENT AND MISINFORMATION

HARMFUL CONTENT

There are several limitations of content removals and de-platforming as a content moderation tool. The volume of content posted to the platforms is many orders of magnitude greater than the reviewing capacity of human and existing machine content moderation tools. This makes thorough and exhaustive content moderation next to impossible. There is the added complexity of a multitude of languages and local cultural nuances leading to gaps in enforcement due to lack of technological tools, language expertise or cultural competence[67]. Platforms have also repeatedly shown their inability to enforce their own rules in a consistent, accurate and fair manner[68]. They are not immune to pressure from the powerful including influencers, media, politicians and governments. Platforms have been known to take down or block content (including critical political speech[69]) based on government requests while also making exceptions for powerful users who are at times linked to the government and its affiliates[70]. The underlying principle in many

[67] Elizabeth Culliford and Brad Heath. "Language Gaps in Facebook's Content Moderation System Allowed Abusive Posts on Platform: Report." The Wire. Accessed March 14, 2022. https://thewire.in/tech/facebook-content-moderation-language-gap-abusive-posts.

[68] Matsakis, Louise. "YouTube Doesn't Know Where Its Own Line Is." *Wired*, March 2, 2018. https://www.wired.com/story/youtube-content-moderation-inconsistent/.;Tworek, Heidi. "The Dangerous Inconsistencies of Digital Platform Policies." Centre for International Governance Innovation, January 13, 2021. https://www.cigionline.org/articles/dangerous-inconsistencies-digital-platform-policies/.;

McSherry, Jillian C. York and Corynne. "Content Moderation Is Broken. Let Us Count the Ways." Electronic Frontier Foundation, April 29, 2019. https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways.;

[69] BBC News. "Vietnam: Facebook and Google 'complicit' in Censorship," December 1, 2020, sec. Asia. https://www.bbc.com/news/world-asia-55140857.

[70] Horwitz, Newley Purnell and Jeff. "Facebook's Hate-Speech Rules Collide With Indian Politics." *Wall Street Journal*, August 14, 2020, sec. Tech. https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-

instances has been linked less to adherence to content moderation rules and more to business considerations and/or bad PR.

There is strong evidence about the efficacy of deplatforming super-spreaders of disinformation in reducing related disinformation on the platform; however deplatforming can only reduce misinformation on the platform where the action is taken and not necessarily the larger information and political ecosystem. This will require a minimum threshold level of standards to be maintained at an ecosystem level lest committed followers of deplatformed influencers migrate to other, less fettered platforms where they could be subjected to even more radical content[71].

FACT-CHECKING

A wide ranging review of social science research found that fact-checking can reduce the impact of misinformation on beliefs, although studies find varying degrees of efficacy depending on the type of misinformation and intervention[72]. Social media has dismantled gatekeeping powers of traditional news media and has enabled the rise of influencers. This has led to the decentralization of content creation and fragmentation of the audience (by influencer, even though different influencers are all on the same few centralised platforms). Some influencers are leading the charge in the spread of misinformation[73]. With public trust in governments[74] and media on the decline

11597423346.; Horwitz, Jeff. "Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt." *Wall Street Journal*, September 13, 2021, sec. Tech. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353.

[71] Colarossi, Jessica. "Banning Trump from Social Media Makes Sense. But Beware the Downside." Boston University, January 8, 2021. https://www.bu.edu/articles/2021/trump-banned-from-twitter-facebook/.

[72] Courchesne, Laura, Julia Ilhardt, and Jacob N. Shapiro. "Review of Social Science Research on the Impact of Countermeasures against Influence Operations." *Harvard Kennedy School Misinformation Review*, September 13, 2021. https://doi.org/10.37016/mr-2020-79.

[73] "The Disinformation Dozen." The Center for Countering Digital Hate, March 24, 2021. https://252f2edd-1c8b-49f5-9bb2-cb57bb47e4ba.filesusr.com/ugd/f4d9b9_b7cedc0553604720b7137f8663366ee5.pdf.

[74] UN. "Trust in Public Institutions: Trends and Implications for Economic Security | DISD," July 20, 2021. https://www.un.org/development/desa/dspd/2021/07/trust-public-institutions/.;

and even authoritative, supposedly non-partisan, sources like the WHO & US Centres for Disease Control and Prevention (CDC) struggling on the public trust front during the COVID-19 pandemic[75], **the challenge of disseminating authoritative information to successfully counter misinformation is often a bigger challenge than the fact-checking itself**.

Another key issue is the question of volume. A 2018 study[76] by the Knight Foundation on the spread of disinformation on Twitter during and after the US2016 elections found that just the top 50 fake news sites received 89% of all fake and conspiracy news links from Twitter. However, the current fact-checking approach involves assessing the veracity of individual articles/pieces of content. Since the volume of user generated content is in the order of billions of posts daily, it is impossible for human driven fact-checking to cover the entire universe of misinformation let alone motivated disinformation[77]. Automation systems that rely on human fact-checking capacities to detect and apply labels to similar content thus run into the same inherent limitations imposed by human capacity constraints. Finally, there are concerns that false headlines that are not selected for fact-checking and thus not labeled as false or misleading on the same distribution surface where fact-checked misinformation is labeled as 'false' could be considered validated by default and thus seen as more accurate due to the "implied truth effect"[78].

EMERGING DISCOURSE

Newman, Nic. "Overview and Key Findings of the 2021 Digital News Report." Reuters Institute for the Study of Journalism. Accessed March 15, 2022. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2021/dnr-executive-summary.;

[75] DiResta, Renée. "Vaccines and the Mediation of Consent." Tech Policy Press, July 19, 2021. https://techpolicy.press/vaccines-and-the-mediation-of-consent/.

[76] Knight Foundation. "Disinformation, 'Fake News' and Influence Campaigns on Twitter," 2018. https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter/.

[77] Binkowski, Brooke. "Opinion: Fact-Checking Facebook Was Like Playing A Doomed Game Of Whack-A-Mole." BuzzFeed News, February 9, 2019. https://www.buzzfeednews.com/article/brookebinkowski/fact-checking-facebook-doomed.

[78] Arun, Chinmayi. "On WhatsApp, Rumours, Lynchings, and the Indian Government." SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, January 3, 2019. https://papers.ssrn.com/abstract=3336127.

Social media platforms make money by selling advertisements to users of their services. Many commentators including Facebook whistleblower Frances Haugen have argued that this business model incentivises platforms to design algorithms to maximize engagement on their platforms leading to a prioritization of extreme, polarizing content including misinformation[79]. Platforms have rejected this causal analysis arguing that they have no economic incentives to push extreme content and misinformation as their users do not want to view such content and their clients do not want their advertisements to run alongside it[8081]. Facebook's Vice President for Global Affairs and Communications Nick Clegg went a step further and argued that the algorithms are merely showing what empowered users train them to show[82]. The counter argument criticizes the myth of an empowered social media user since the sheer amount of information online is too much for even the most educated and privileged person to keep up with[83] in order to use social media with agency.

More recently, the discourse surrounding content moderation is shifting away from content removal to focus on how harmful, mis/disinformation and borderline content is distributed. There is extensive media reportage demonstrating YouTube's recommendation algorithm's propensity to promote inflammatory content and favor extremism. There is similar reportage on Facebook's algorithms[84]. Experts are calling for platforms to "turn down the dial" on toxic discourse[85].

---

[79] Hao, Karen. "The Facebook Whistleblower Says Its Algorithms Are Dangerous. Here's Why." MIT Technology Review, October 5, 2021. https://www.technologyreview.com/2021/10/05/1036519/facebook-whistleblower-frances-haugen-algorithms/.

[80] Zuckerberg, Mark. "Facebook Note," October 6, 2021. https://www.facebook.com/zuck/posts/10113961365418581.

[81] Google.com. "Improving our Brand Safety Controls", March 2017. https://blog.google/around-the-globe/google-europe/improving-our-brand-safety-controls/

[82] Clegg, Nick. "You and the Algorithm: It Takes Two to Tango | by Nick Clegg | Medium," March 31, 2021. https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2.

[83] Carmi, Elinor. "Nick Clegg and Silicon Valley's Myth of the Empowered User." Tech Policy Press, April 7, 2021. https://techpolicy.press/nick-clegg-and-silicon-valleys-myth-of-the-empowered-user/.

[84] Tech Transparency Project. "Facebook's Militia Mess," March 24, 2021. https://www.techtransparencyproject.org/articles/facebooks-militia-mess.

[85] Douek, Evelyn. "What Facebook Did for Chauvin's Trial Should Happen All the Time." The Atlantic, April 21, 2021. https://www.theatlantic.com/ideas/archive/2021/04/facebook-should-dial-down-toxicity-much-more-often/618653/.

Rigorous research into the impact of algorithmic distribution of content is constrained by social media platforms' reticence to share their proprietary data and algorithms often on grounds of trade secrecy[86]. There are active legislative processes in the US and Europe that, if passed, will force platforms to be more transparent about their algorithms. Beyond increasingly strident calls for transparency, other demands include moving from content moderation to focusing on algorithmic amplification[87] to doing away with algorithmic amplification altogether[88].

PART 3: POLITICS OF DISINFORMATION

The discourse around misinformation and its resolution has become excessively mired in content-moderation and its minutiae. This framework suits social media platforms but is adjunct to the core issue of misinformation, which is distribution. This section outlines the key issues essential to understand the issue of contemporary misinformation and chart a possible way forward

PLATFORMS HAVE MADE MISINFORMATION UBIQUITOUS AND INDISTINGUISHABLE FROM QUALITY INFORMATION

[86] Daphne Keller. "Has Anyone Seen Good Analysis of Platforms' Trade Secret-Based Objections to Transparency? For All but Serious Secret-Sauce Tech, This Always Sounds Bogus to Me. Many Transparency Improvements Don't Need to Reach the Secret Sauce, Anyway. Am I Missing Something?" Tweet. *@daphnehk* (blog), November 8, 2021. https://twitter.com/daphnehk/status/1457748966158712837.

[87] "Nobel-Winning Journalist Ressa: 'Tech Tearing Apart Shared Reality' | News | DW | 19.10.2021." Accessed March 15, 2022. https://www.dw.com/en/nobel-winning-journalist-ressa-tech-tearing-apart-shared-reality/a-59552866.

[88] Milmo, Dan, and Dan Milmo Global technology editor. "Five Questions in Westminster for Facebook Whistleblower Frances Haugen." *The Guardian*, October 25, 2021, sec. Technology. https://www.theguardian.com/technology/2021/oct/25/five-questions-in-westminster-for-facebook-whistleblower-frances-haugen.

Misinformation has always existed however **turbocharged distribution** through social media platforms has made misinformation and propaganda invasive[89] and pervasive. Platforms have further elided the distinction between different sources of information which has removed an important signal of credibility and ideological positioning of the consumed content. Instead engagement is perceived to be a bigger driver of the importance - and by extension - credibility of a piece of news. This equal treatment (appearance and placement of different and unequal sources of information) and making virality instead of quality the primary determinant of a source's credibility and/or a piece of content's importance has eroded the distinction between vetted information, propaganda and misinformation in the minds of the user. The impact is especially acute in India because platforms have de-facto control over distribution of the message combined with low digital literacy among users.

FALSE PLATFORM-FUELED BINARY BETWEEN MISINFORMATION AND FREE SPEECH

Social media platforms keep users engaged by constantly keeping their feeds populated with new content from sources and content creators that the user has not proactively followed. This is a deliberate boost to the organic reach of a subset of content by the platforms and is known as "amplification". Since quality and value based amplification is difficult due to the challenge of determining "quality" and "value", platforms rely on amplification based primarily on engagement signals[90]. This approach absolves platforms of the need to exclude vast swathes of content while remaining value agnostic and thus avoiding charges of editorial control. This approach is leading to two consequences: since amplification is disproportionately driven by engagement and not substantive metrics of quality of content, the entirety of content[91] on the platform is used as input

[89] Almost all users mentioned that they responded to notifications of new content posted on various social media platforms, which serve as de-facto portals to the internet

[90] In response to the criticism of this engagement-driven amplification, platforms have tried to pivot towards "quality". However two things stand out: first, there is no transparency on the differential weightage between quality and engagement signals; second, the determination of "quality" too is driven by signals of meta engagement instead of a review of actual content

[91] Excluding obvious content which violates platform guidelines and is easily caught through machine learning such as graphic violence, nudity (for Facebook) etc

inventory for amplification; and since borderline content gets more engagement (as admitted by platforms themselves), this value-neutral and engagement-driven approach is resulting in amplification of misinformation and other harmful content, including in some instances, content which violates platforms' own content policies [929394]. **The underlying issue for pervasive misinformation is thus driven more by value-neutral amplification of content propagating misinformation instead of failure to outrightly remove misinformation. However, platforms have bypassed the discussion around amplified distribution and are exclusively framing measures to reduce misinformation as being in "tension" with freedom of expression - an issue which can arise only in the case of outright removal**.

FOR PLATFORMS, FREE SPEECH IS A BUSINESS MODEL INSTEAD OF A PRINCIPLED IMPERATIVE

As private corporations, social media platforms have - and exercise - the right to decide what content they want to host. Adopting a hands-off approach to user-generated content has helped platforms achieve unbridled growth and is also operationally simpler. Moreover, it is notable that important content moderation decisions by platforms are often ad-hoc and driven by external pressure - especially government, media, PR - instead of coherent speech policies. Most importantly, platforms have opportunistically used the laudable principle of "free speech" and the protection against liability for intermediaries to advance their business models while failing to ensure a good information ecosystem. Traditional news media is liable for published content and must thus invest time and resources to vet information before publishing. Platforms compete with traditional news publishers for advertising revenue while enjoying the double advantage of speed

[92] Sonnemaker, Tyler. "Facebook Is Still Failing to Enforce Its Own Rules against Election Disinformation, Conspiracy Theories, Hate Speech, and Other Policies, Another New Analysis Finds." Business Insider, October 16, 2020. https://www.businessinsider.in/tech/news/facebook-is-still-failing-to-enforce-its-own-rules-against-election-disinformation-conspiracy-theories-hate-speech-and-other-policies-another-new-analysis-finds/articleshow/78691790.cms.

[93] Giansiracusa, Noah. "How Facebook Hides How Terrible It Is With Hate Speech." *Wired*, October 15, 2021. https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/.

[94] Stokel-Walker, Chris. "YouTube's Algorithm Recommends Videos That Violate Its Own Policies." New Scientist, July 7, 2021. https://www.newscientist.com/article/2283354-youtubes-algorithm-recommends-videos-that-violate-its-own-policies/.

(to get content to users) and protection from liability (for unvetted content). Since, advertising revenue is directly proportional to the amount of time users spend on the platforms, platforms have exploited this twin advantage to boost user engagement without caring about the deleterious impact of a surfeit of misinformation on the information ecosystem and wider democracy.

It can thus be argued that for social media platforms, "free speech" is a business model instead of a principled imperative. **As private companies, the issue in the case of outright removal of content, is not freedom of speech but political neutrality of the platform. The degree of permissiveness for misinformation, hate speech, etc is thus a political and/or commercial choice by the platforms**. The argument that this is a political and/or commercial choice is underscored by the differential treatment of the same content on different platforms. A good example of this is President Trump's tweet and Facebook post with the statement "...when the looting shoots, the shooting starts…" in May 2020 against the protests in the wake of George Flyod's death. Twitter restricted the tweet by placing a public interest notice on the relevant tweet by President Trump for breaking the platform's rules. It stopped short of outright removal of the tweet citing public interest. Facebook on the other hand, chose not to take any action on the same content posted to its platform citing a commitment to free expression and public interest[95].

PLATFORMS ARE SENTIENT AND STRADDLE THE CONTINUUM FROM DISTRIBUTORS TO PUBLISHERS

Distribution of Content on Platforms is a Choice: As private corporations, platforms have the right to decide what content they will not host and distribute on their platform. This is explicitly codified in US law as the "Good Samaritan" provision in Section 230 of the Communication Decency Act which gives platforms protection from civil liability for removal or moderation of user content. Accordingly, all major platforms have extensive terms of service and content guidelines which define what is and is not acceptable to be posted on the platform. Facebook, for instance, bans nudity and pornography while Twitter permits them. Further all platforms invest in moderating content they *privately* determine to be unsuitable for their platform. This implies that platforms

[95] https://www.politico.com/news/2020/05/29/zuckerberg-facebook-leaving-up-trumps-shooting-post-290292

are aware of the content they are hosting and that distribution of content on their platform is a choice. This point is further underscored by the fact that platforms are known to "deplatform" certain users (notably President Trump and more recently many Russian state handles during Russia's invasion of Ukraine) to showcase their *political* distance from the deplatformed user. Platforms thus prohibit distribution of the entirety of the deplatformed user's content on their platform and not just violating content by the deplatformed user.

As platforms sought to maximize user engagement (time spent on platform) and retain/increase users, platforms first started to curate the *distribution* of content (amplification) and now some platforms have moved towards directly commissioning content itself. Facebook, for instance, has pledged USD 1 billion for creator content in 2022 (for reference, its net income in 2021 was USD 33 billion) and are thus increasingly evolving to become publishers. It is therefore too simplistic to argue that platforms have no responsibility for content hosted on their platforms. **Instead platforms have a "duty of care" in proportion to the harms posed by the content they are hosting along with liability linked to their distribution choice**.

PLATFORMS ARE RESPONSIBLE FOR POLITICAL AND COMMERCIAL IMPACT OF AMPLIFICATION

The internet has led to the transition from a content scarce to "content surplus and attention scarce" economy. Therefore, amplification of user content and consequent mass engagement has obvious political and commercial value (in proportion to increased distribution[96]) for the creator of content. This is evident from the mainstreaming of individuals and narratives which may otherwise have remained on the fringes of public consciousness. Amplification is thus an intervention in socio-economic and political processes of a society. Initially platforms embraced this impact on societies and political systems by positioning themselves as harbingers of democracy and pro-people

[96] There is no transparency in how much increased distribution of a particular piece of content has resulted from platform design and amplification choices

movements especially during the Arab Spring[97]. However, the pro-David positioning of platforms is no longer compatible as platforms have become increasingly aligned with power instead of ordinary people in areas of conflict. Organized political entities have mobilized online (in India through dedicated political party "IT cells") overwhelming individual dissidents and less-resourced people's movements, and platforms have been found to be compliant with government take-down requests (of dissident speech)[98].

CONTENT MODERATION IS DRIVEN BY THE DYNAMICS OF THE POLITICAL PROCESS

Fact-checking acquires salience primarily when some issue has become politicized and the underlying (alleged) facts form the basis for competing narratives. Fact-checking and consequent content moderation decisions are thus inevitably seen as an intervention in the political process especially since social media platforms have displaced existing information systems. Platforms are more likely to intervene when there's relative political consensus on the way forward and there's an authoritative source of information which can vouch for "truth" against which content can be verified. In the two instances - the need to combat the spread of COVID with the WHO as the authoritative source; and the need to maintain integrity of the 2020 US Presidential election with authoritative sources for election results - platforms have been unusually interventionist even in instances where the counterview may be correct (e.g., the WHO's initial directive against mass use of masks[99]). However, with increasing polarization and contest over trustworthiness of all sources of information - whether primary or reported - along with politicization of "objective" fields such as competing "science" narratives, fact-checking is itself becoming politicized. This

[97] Arafa, Mohamed, and Crystal Armstrong. "'Facebook to Mobilize, Twitter to Coordinate Protests, and YouTube to Tell the World': New Media, Cyberactivism, and the Arab Spring," January 2016, 31. https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1187&context=jgi

[98] Wade, Peter. "Facebook Bowed to Vietnam Government's Censorship Demands: Report." *Rolling Stone* (blog), October 25, 2021. https://www.rollingstone.com/politics/politics-news/facebook-vietnam-censorship-1247323/.; Zia, Ather. "How Facebook Helps Silence Kashmiri Voices." The Caravan, December 1, 2019. https://caravanmagazine.in/commentary/how-facebook-helps-silence-kashmiris.; Ghaffary, Shirin. "A Major Battle over Free Speech on Social Media Is Playing out in India during the Pandemic." Vox, May 1, 2021. https://www.vox.com/recode/22410931/india-pandemic-facebook-twitter-free-speech-modi-covid-19-censorship-free-speech-takedown.

[99] Howard, Jacqueline. "WHO Stands by Recommendation to Not Wear Masks If You Are Not Sick or Not Caring for Someone Who Is Sick." CNN, March 31, 2020. https://www.cnn.com/2020/03/30/world/coronavirus-who-masks-recommendation-trnd/index.html.

problem is exacerbated when a large amount of misinformation is organized disinformation (propaganda) and linked to powerful (state) actors. It is evident that private companies cannot be relied upon to take on national governments especially in countries with weak rule of law given business considerations and the safety of local employees. Finally, because norms for acceptable speech are constantly evolving and vary widely across societies, cultures and geographies, it is difficult to achieve wide and stable consensus on standards for content moderation thus embedding content moderation in a dynamic political process.

CONTENT MODERATION IS NOT A SCALABLE RESPONSE TO MISINFORMATION

One key aspect of content moderation, especially in a country like India, is allocation of adequate resources to account for the volume of content and linguistic diversity. However, given that content moderation is inherently linked to the political process, true scalability is limited by the constraints of the political process itself since contested issues will outpace the ability to drive consensus.

This also implies that a framework of content moderation which relies on individual post-level fact-check is operationally doomed to fail by virtue of the scale and complexity of the problem given the sheer volume of posts, local context, language etc.

DISINFORMATION IS A BAD ACTOR PROBLEM NOT A DISCRETE CONTENT MODERATION ISSUE

Disinformation is the deliberate creation and propagation of false information in service of a political and/or commercial goal. This understanding of "disinformation as a bad actor problem instead of a discrete content issue" is evident in platform policies such as deplatforming based on offline activity, removal of accounts involved in "coordinated inauthentic activity", pages which build audience through "manufactured virality" etc. In each of these instances, the impetus for action against the user is not discrete content posted by them but a judgement about the user account(s) itself. Moreover labeling of user accounts affiliated with a government or state-

affiliated media entity too is acknowledgement that content is inextricably linked with the user and does not exist in isolation. Moreover studies[100][101] have found that a small subset of sites/handles are responsible for bulk of misinformation on platforms. Since disinformation is driven by motivated actors producing false information at an industrial scale, a more effective approach to resolution would be at the user level instead of focusing only on cleaning up on a discrete content by content basis. Such a user-level approach could take multiple forms including deplatforming, labeling and limiting amplification to credible users only.

PART 4: WAY FORWARD

RECOMMENDATIONS

The recommendations like the rest of the report are intended to serve as a basis for discussion and provide a directional way forward.

• BRING A COMPREHENSIVE TRANSPARENCY LAW FOR PLATFORMS

One of the biggest hurdles in being able to stem the flow of misinformation and understand its impact on our society and polity is lack of transparency by social media platforms. Even when platforms have disclosed certain kinds of information (e.g., the Ad Library by Facebook), the data is often not presented in a manner which facilitates easy analysis and prompt response. At the same time, there are reasonable privacy concerns about certain kinds of data sharing. Data transparency bills in the pipeline, notably in the United States

[100] Bond, Shannon. "Just 12 People Are Behind Most Vaccine Hoaxes On Social Media, Research Shows." *NPR*, May 14, 2021, sec. Shots - Health News. https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes.

[101] Paul, Kari. "A Few Rightwing 'super-Spreaders' Fueled Bulk of Election Falsehoods, Study Says." *The Guardian*, March 5, 2021, sec. US news. https://www.theguardian.com/us-news/2021/mar/05/election-misinformation-trump-rightwing-super-spreader-study.

and Europe seek to address these issues. However, given platforms' differential treatment of the Global South, India must enact its own comprehensive transparency law to ensure parity and relevance for India.

- ## CONSTITUTE A REGULATOR UNDER PARLIAMENTARY OVERSIGHT

Content moderation is embedded in the political process; however, the determination of the standards, evaluation and manner of content moderation have been entirely outsourced to private corporations. This is highly undesirable: public opinion is the currency of democracy and social media platforms are increasingly becoming the primary ground for public discourse and mobilization of public opinion. Giving control over the public discourse to a handful of individuals heading technology companies lacks both transparency and democratic legitimacy. Equally importantly, this approach has shown itself to be highly inefficient: platforms have shown themselves unable to work together to evolve a coherent framework to stop misinformation and have instead responded erratically to events and ad-hoc public pressure. Moreover, newer platforms are emerging where the differentiating point is not more innovative ways of engagement but more permissive standards of speech. Since misinformation exists within an ecosystem, the absence of a uniform baseline approach, enforcement and accountability has served to vitiate the entire information ecosystem. External regulation is thus desirable. However, given the highly dynamic and "of the moment" nature of this issue, any legislative route will rapidly become outdated. At the same time, bringing governance of speech under state purview is fraught with risks to free speech and democratic dissent. A way forward could be to constitute a statutory regulator under Parliamentary oversight. The Regulator would have statutory powers to lay out broad processes for governance of speech, set standards for transparency for social media platforms (within the framework of the law referenced in the point above) and audit social media platforms for compliance; and advisory powers to develop a point of view on key misinformation themes/events in the country especially on issues which have immediate public policy implications e.g., public health misinformation on COVID. This aspect of its role should derive legitimacy from the inclusiveness and credibility of its process and not coercive power. Such a model will ensure three things: first, it will increase democratic contest by moving contested speech issues into the political sphere from the

private and/or government sphere; it will enable more considered but still agile interventions in the public discourse on urgent public policy issues; and facilitate greater transparency of powerful technology platforms. The Regulator would be answerable to the Parliament and not the Executive.

- ## PLATFORMS MUST CHOOSE APPROACH TO DISTRIBUTION AND AMPLIFICATION

Amplification of user content has a significant impact on public consciousness and public discourse and is thus an intervention in the socio-economic and political processes of the country. Platforms can neither distance nor absolve themselves of responsibility by adopting a value-neutral stance towards their own actions. It is thus incumbent on platforms to define their approach to amplification. There are two possible paths: adopt a hands-off approach to content (and by extension content creators) and constrain distribution to organic reach (chronological feed); or exercise clear editorial choice and take responsibility for amplified content. However, it is notable that platforms have consistently refused to take editorial responsibility and there are numerous news reports which show that platforms amplify content which violate their own content policies[102][103][104]. A third possible via media is to amplify only those content providers[105] who have gone through a vetting process[106] to ensure that amplified content has gone through some due process for integrity

[102] Sonnemaker, Tyler. "Facebook Is Still Failing to Enforce Its Own Rules against Election Disinformation, Conspiracy Theories, Hate Speech, and Other Policies, Another New Analysis Finds." Business Insider, October 16, 2020. https://www.businessinsider.in/tech/news/facebook-is-still-failing-to-enforce-its-own-rules-against-election-disinformation-conspiracy-theories-hate-speech-and-other-policies-another-new-analysis-finds/articleshow/78691790.cms.

[103] Guy Rosen. "How We're Tackling Misinformation Across Our Apps." *Meta* (blog), March 22, 2021. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/.

[104] Stokel-Walker, Chris. "YouTube's Algorithm Recommends Videos That Violate Its Own Policies." New Scientist, July 7, 2021. https://www.newscientist.com/article/2283354-youtubes-algorithm-recommends-videos-that-violate-its-own-policies/.

[105] One concern with such a vetting process is that it will likely favour institutions over individuals. Such an approach would thus need to go hand in hand with other measures to support individuals and upcoming institutions

[106] There is a similar precedent in Apple's approach to the App Store where each application theoretically goes through a vetting process to ensure a "safe experience for users to get apps"

and quality of messaging, irrespective of ideological affiliation. This option has been explored in greater detail below.

- ### MAKE AMPLIFICATION CONTINGENT ON CREDIBILITY OF CREATORS AND SOURCES INSTEAD OF TOKEN LINKAGE WITH A NEGATIVE LIST

Volume of user generated content is several magnitudes higher than the capacity of platforms to detect and enforce their own content standards. Thus only a small subset of content is fact-checked and content within this subset which is found to be false/misleading is made ineligible for amplification. This can be thought of as eliminating a small negative list of content from serving as inventory for amplification algorithms. It is evident that the negative list generated in this manner will be incomplete and partial when compared to the total potential bad content (including outrightly violating content such as hate speech, nudity, etc as well as gray area content like misinformation) on any given platform. The rest of the content that is not fact-checked or caught for violating platform speech rules and which should have made it to the negative list but did not instead continues to be eligible for amplification based on value-neutral and predominantly engagement-driven criteria. This means that not only is bad content not being removed but it is instead being amplified and distributed more widely by platforms. **This notional linkage of distribution with a token negative list based on a very small subset of total content generated can not only not curb misinformation but is actually responsible for its amplification**.

For platforms which are committed to eliminating the distribution of misinformation on their platform but feel that amplification is an inextricable part of their business model, it is suggested that amplification of "political content" be linked to the user's trustworthiness and credibility. In this approach, users who primarily post content on topical issues will be

classified as "political users"[107] and amplification of such users would be contingent on meeting certain criteria of due process. It is argued that given the nature of political speech and its impact on public discourse and democracy, increased reach or amplification should be contingent on a certain level of responsible production instead of engagement (which has been proven to privilege misinformation and divisive content).

There are a variety of efforts around the world to rate trustworthiness and credibility of news sources and social media platforms. There are efforts too to find a more community driven approach to provide additional context around content, such as Twitter's Birdwatch[108] which is based on building out "reputation" and "consensus" systems for users and associated note-making. Platforms can draw upon these initiatives to populate a positive list of users which would be eligible for amplification. The positive list can be a graduated list with progressively higher standards of credibility linked to increased amplification. Disinformation is predominantly a bad-actor problem instead of a discrete content by content issue; such an approach will create disincentives for bad actors and will insulate platforms from getting embroiled in controversies regarding censorship and freedom of expression, difficulty of ascertaining "truth" and charges of ideological bias since individual pieces of content will not be removed. The purpose of a positive list would not be to establish ideological boundaries but to establish the principle that amplification will be restricted to only those creators which can be held accountable for due process and standards of content creation. Users and creators would still be able to increase their reach through organic outreach methods.

[107] AI can be used to identify who primarily post content on topical issues and classify them as "political". It is argued that given the nature of political speech and its impact on public discourse and democracy, increased reach or amplification be contingent on a certain level of responsible production.

[108] Coleman. "Introducing Birdwatch, a Community-Based Approach to Misinformation," January 25, 2021. https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

Operational discussion and concerns: There are some operational challenges and second-order consequences which need further discussion: such a system is likely to favor institutions and powerful actors over ordinary individuals and small local movements. Further, in a highly polarized atmosphere, implementation of such a rating system may struggle to establish legitimacy when challenged by negatively affected users; the sheer number of creators, even if limited to those posting political content, will be considerable and the system may struggle to rate this universe on an ongoing basis. However, this is a worthwhile direction for research, especially in artificial intelligence. For instance, AI is already being used to identify original news reporting.

- ## LABEL CONTENT PRODUCERS INSTEAD OF INDIVIDUAL POSTS

Social media platforms have displaced user agency on selection of source for news and information, thus removing an important signal of credibility and ideological affiliation. This is significant because studies[109][110] have found that a small subset of sites/handles are responsible for bulk of misinformation on platforms. Platforms can provide this supplementary information by directly labeling the source of information on trustworthiness instead of placing labels on individual pieces of content. This will also bring down the universe of content which would need to be reviewed. Norms can be devised to identify such handles and/or groups. Under this proposal, users who repeatedly post borderline content, or lift original content, or post content fact-checked to be false by independent third-party fact checkers and/or other reputation and credibility related research will be labeled as 'Low Credibility Source' in addition to the content itself being labeled as false. User-level action - whether deplatforming or labeling users to provide

[109] Bond, Shannon. "Just 12 People Are Behind Most Vaccine Hoaxes On Social Media, Research Shows." *NPR*, May 14, 2021, sec. Shots - Health News. https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes.

[110] Paul, Kari. "A Few Rightwing 'super-Spreaders' Fueled Bulk of Election Falsehoods, Study Says." *The Guardian*, March 5, 2021, sec. US news. https://www.theguardian.com/us-news/2021/mar/05/election-misinformation-trump-rightwing-super-spreader-study.

additional context around the posted content (irrespective of the content itself being fact-checked or not) is already being done by many platforms. Examples include labeling labeling state users, and removing pages (comparable to users) based on how they generate content and garner audience instead of the content itself posted by them

- ### REVIEW SUPER USERS

A set of users have acquired high organic reach by propagating false and divisive content or through other inauthentic processes. These users are the fountainhead of misinformation with studies documenting their disproportionate impact in spreading misinformation online. Platforms should establish a principle of progressively more stringent content guidelines for users with high organic reach and select the top users (either top 1% of users in a geography or all users above a certain follower threshold) for review. Such accounts can be reviewed periodically and should be deplatformed for repeated violations of guidelines.

- ### SET MINIMUM STANDARDS FOR INTEGRITY INVESTMENTS, INFRASTRUCTURE AND TRANSPARENCY AT COUNTRY LEVEL

Platforms must have the ability to comprehensively moderate content (understand local language, context and political dynamics) when entering any new country. In the countries where platforms have already become entrenched, this capacity must be ramped up on a fixed timescale. Further, since complexity is often linked to size of the country, integrity investments must be linked to the size of the user base instead of the current revenue linkage which penalizes large but developing countries. This role will automatically come under the proposed Regulator's purview.

- ### RECONFIGURE DEFAULT USER FEED TO CHRONOLOGICAL FEED

Platforms are de-facto portals to the internet and control content being consumed by individuals. While theoretically, users can configure their feed to remove "amplification" and "personalisation", most users lack awareness or understanding of this option. Platforms must thus do two things: make chronological feed default thereby exposing the user to only that content which s/he has specifically opted to receive; ensure that all content which is put in the feed is controlled for source and quality. One way to do this is through the creation of a "positive list" as proposed above.

- ### REMOVE DESIGN CHOICES WHICH INCENTIVISE EXTREME CONTENT

Platforms incentivise extreme content, much of which is likely to be disinformation in two ways: through amplification algorithms, which are predicated on engagement instead of quality thus privileging borderline content (which get more organic engagement); and the signaling of high engagement as a driver of importance of a piece of content. Platforms can make two immediate improvements in this regard: hide engagement numbers[111] on content and eliminate trending topics. Trends have been increasingly hijacked by organized entities acting in coordinated manner and are no longer an indicator of popular issues animating the public.

- ### DEVELOP ECOSYSTEM APPROACH TO FACT-CHECKING (EASE, CAPACITY, AND DISTRIBUTION)

At present, fact-checking initiatives are adjunct to the actual business and design of platforms. They thus lack adequate incentives and resources, feedback loop within the platform organization and a framework for priority distribution. The latter is specially important since disinformation is targeted but fact-checked content lacks similarly targeted distribution. The following suggestions are thus aimed at addressing some of the known

[111] Instagram, for instance, provides an option to hide "like" counts on posts to reduce social media pressure

deficiencies[112]: expand fact-checking capacity to include all languages above a certain threshold of users; work with relevant stakeholders to develop specific measures to reduce disinformation during elections and ensure parity of prioritization for India and Global South with US2020 elections; translate fact-checked information into regional languages for accessibility [113]; push fact-checked information into user feeds and targeted to individuals who have shown prior interest or engagement with related topics (to increase likelihood that users exposed to disinformation will see fact-checks); recruit "Fact Ambassadors [114]" for distribution of fact-checked information. This will improve distribution and reception of fact-checked information. It will also provide social media platforms with valuable feedback on information consumption behavior; use AI to provide representative counterview by surfacing links to good quality news reports and comments. Some platforms like Facebook[115], YouTube[116] and Google[117] are already doing this and this measure can be expanded to cover all content above a certain threshold of engagement; provide crowdsourcing channels to flag content that is going viral, are time-sensitive and may lead to real world harm for real-time fact-checking and proactive notification to impacted users (this is already being done in some measure).

- SCALE DIGITAL LITERACY PROGRAMS

[112] Not all suggestions may be appropriate to all platforms. Some suggestions may already be being implemented in partial form by some platforms and thus the recommendation is to scale existing measure

[113] This is specially important for countries like India where large sections of the population do not understand English

[114] Future of India Foundation has a separate proposal on the identification, training and management of local community Fact Ambassadors

[115] Lyons, Tessa. "Replacing Disputed Flags With Related Articles." *Meta* (blog), December 21, 2017. https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/.

[116] "See Fact Checks in YouTube Search Results - YouTube Help." Accessed March 15, 2022. https://support.google.com/youtube/answer/9229632?hl=en.

[117] "Find Fact Checks in Search Results - Google Search Help." Accessed March 15, 2022. https://support.google.com/websearch/answer/7315336?hl=en.

Digital literacy as a way to reduce misinformation works only if done at scale. Social media platforms are at the forefront of distribution but have limited their digital literacy initiatives to a public relations exercise. This needs increased impetus with a target for outreach linked to the platform's user base[118] such that a critical mass of users undergo digital literacy training in a fixed time period.

CONCLUSION

This report is an attempt to cut through the morass of crosstalk and obfuscation on the issue of disinformation to make two clear points: first, disinformation is a political problem and any way forward must be located within the broader political process, involving all relevant stakeholders. Any attempt to seek resolution within a technocratic or solely government framework will lack buy-in and democratic legitimacy and is unlikely to lead to any kind of lasting resolution; second, as long as amplification is driven by engagement instead of trustworthiness of content sources, the current content-moderation driven approach to disinformation by all major social media platforms will not stop the spread of misinformation. Therefore, if we are serious about addressing disinformation and its deleterious impact on our democracy, the way forward has to be radically different from the current approach instead of getting lost in the minutiae of content moderation and its implementation.

The report suggests a framework to think about the relevant issues and suggests two possible paths forward for platforms: reversion to chronological feed or amplification of content/users based on

[118] FOI has outlined multiple approaches for scalable digital literacy programs in a separate report

quality and trustworthiness. The recommendations in the report are directional and suggest ways of approaching the problem of misinformation. We believe that much more deliberation is required and a lot more needs to be done in this space by all stakeholders invested in ensuring a healthy information ecosystem, equitable and safe access to the internet and social media in particular for Indians. Platforms especially must recognize and acknowledge the impact their products are having in India and the Global South, prioritise investments and capacity commensurate to this impact and support transparency and knowledge sharing initiatives here at par with what is happening in the United States and Western Europe. At the same time, Indian Government initiatives have primarily focused on controlling social media platforms through legalistic instruments and via threats of criminal liabilities. Instead the world's largest democracy should set an example for the rest of the world by locating its regulatory efforts in the broader democratic political process and by bringing a comprehensive transparency law to force meaningful disclosures by platforms to enable a broader community of informed stakeholders.

Finally, it is important to acknowledge that public opinion is the currency of democracy and that everything about information - from definition to regulation - is political. There can thus be no "apolitical" approach to disinformation. The question is only where do we want to draw the line.

The End

ANNEXURE

QUESTIONS ASKED IN THE FOCUS GROUP DISCUSSIONS:

1. How and where did you hear of COVID first?

2. Where do you get your regular COVID updates?
3. Who or what source do you trust most on COVID?
4. Have you discussed COVID or politics with your friends and family in the past week?
    a. How did you find this information/What has been the basis of this discussion (source of info/news)? Through what medium did you get this information?
    b. Mention some issue they talk about and ask them about alternate explanations
    c. When there are multiple versions of the same issue - how do you know which is true?
5. If you want to know something important, how will you find authoritative information? Is there an information source you trust? Online and offline
6. What are the different kinds of information related to COVID you have seen?
    a. To ascertain the different kinds of misinformation seen by users
7. What is the impact of various kinds of COVID posts in your community?
    a. To ascertain impact of misinformation in communities
8. Have you read any COVID info from fact checkers? If yes, which fact checker? Do you believe in these fact checkers?
9. People in the Opposition and media suggest that the number of people who died from COVID is 4-30 times more than official numbers. What do you think? Does it matter to you what the real number is? How will the number change your actions?
10. Did you forward COVID related information or news in your network? If yes, how frequently, to whom and why?
11. Would it make you feel good/proud if you had authoritative information that you could give to your friends on topical issues? If you were recognised as a ‘Fact Ambassador’ or ‘Ambassador of Truth’ by prestigious companies?

12. What would make you change your mind about something? Who can change your mind about something? Your friends, family, elders in your community, teacher, a political leader, newspapers, TV channels?

FUTURE OF INDIA

Future of India Foundation is a registered non-partisan foundation to develop **actionable** big picture ideas for the future of India keeping the needs, aspirations and context of its youth front and center. The Foundation works at the intersection of big-picture youth issues, politics and technology.

Founders and Directors

Ruchi Gupta is former national in-charge of the world's largest progressive student union with more than 40 lakh registered members and a global diaspora base. Ruchi has 15 years of experience working in the Indian political and social sector space and has written extensively on policy and political issues in all the English language national newspapers and journals. Ruchi is co-editing the Youth Volume in Samruddha Bharat Foundation's "Rethinking India" series being published by Penguin. Before coming back to India, she was a consultant with McKinsey & Co. in New York City. She is a Fellow at The Aspen Institute and a member of Aspen Global Leadership Network. Collected writings are available at rgupta.substack.com

Email: ruchi@futureofindia.in | Twitter: guptar

Saurabh Sharma is founder of Josh, a non-profit organization which works with youth on issues of transparency and accountability. Saurabh is the convener of Delhi's Right to Education Forum and a core member of the National Right to Education Forum. J has worked extensively on the implementation of RTE in government schools of Delhi's resettlement colonies. Saurabh ran the "Youth and MDG" Secretariat in collaboration with Unicef in India through which consultations with youth were held all over the country, and has represented Indian youth in many international forums