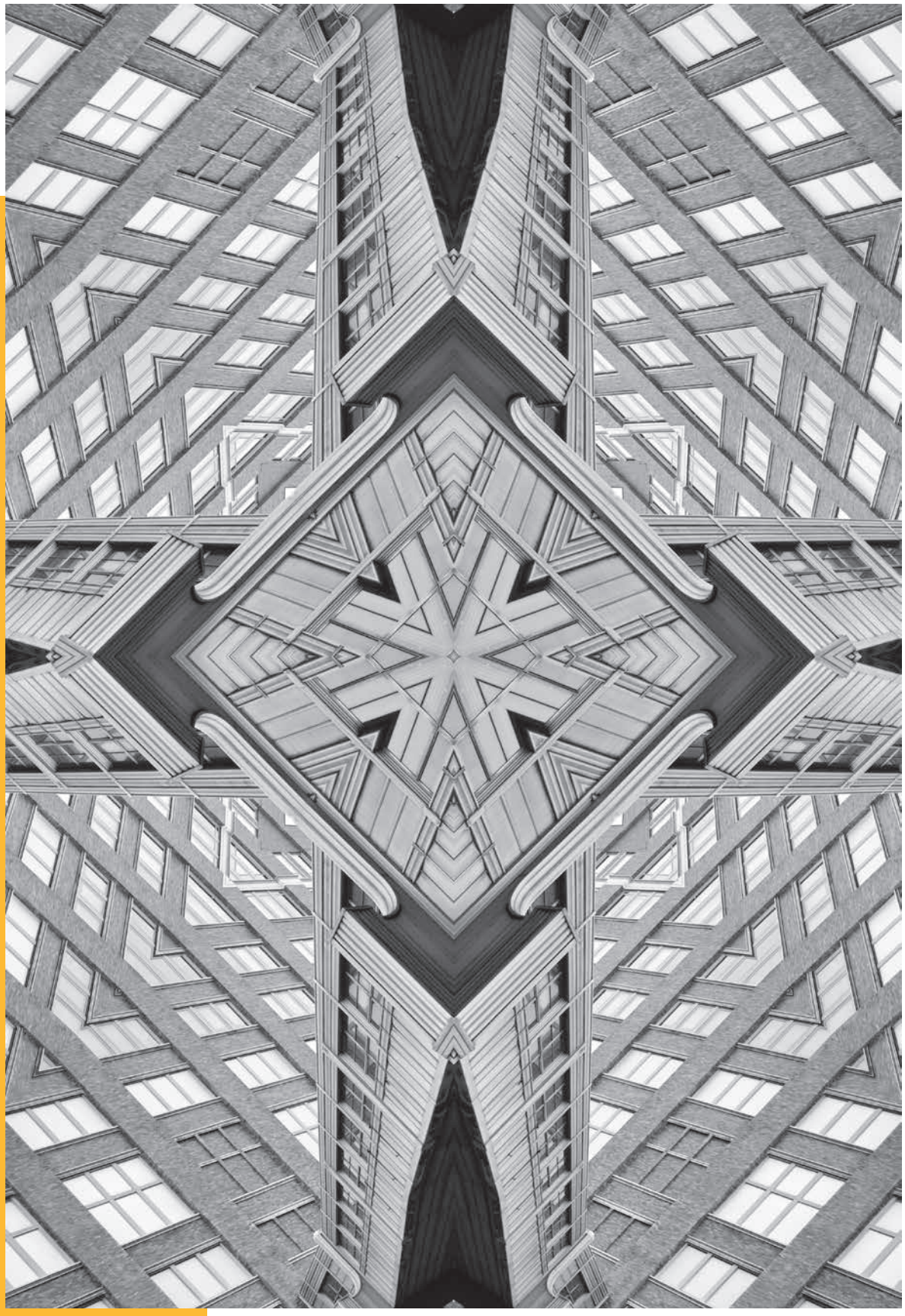
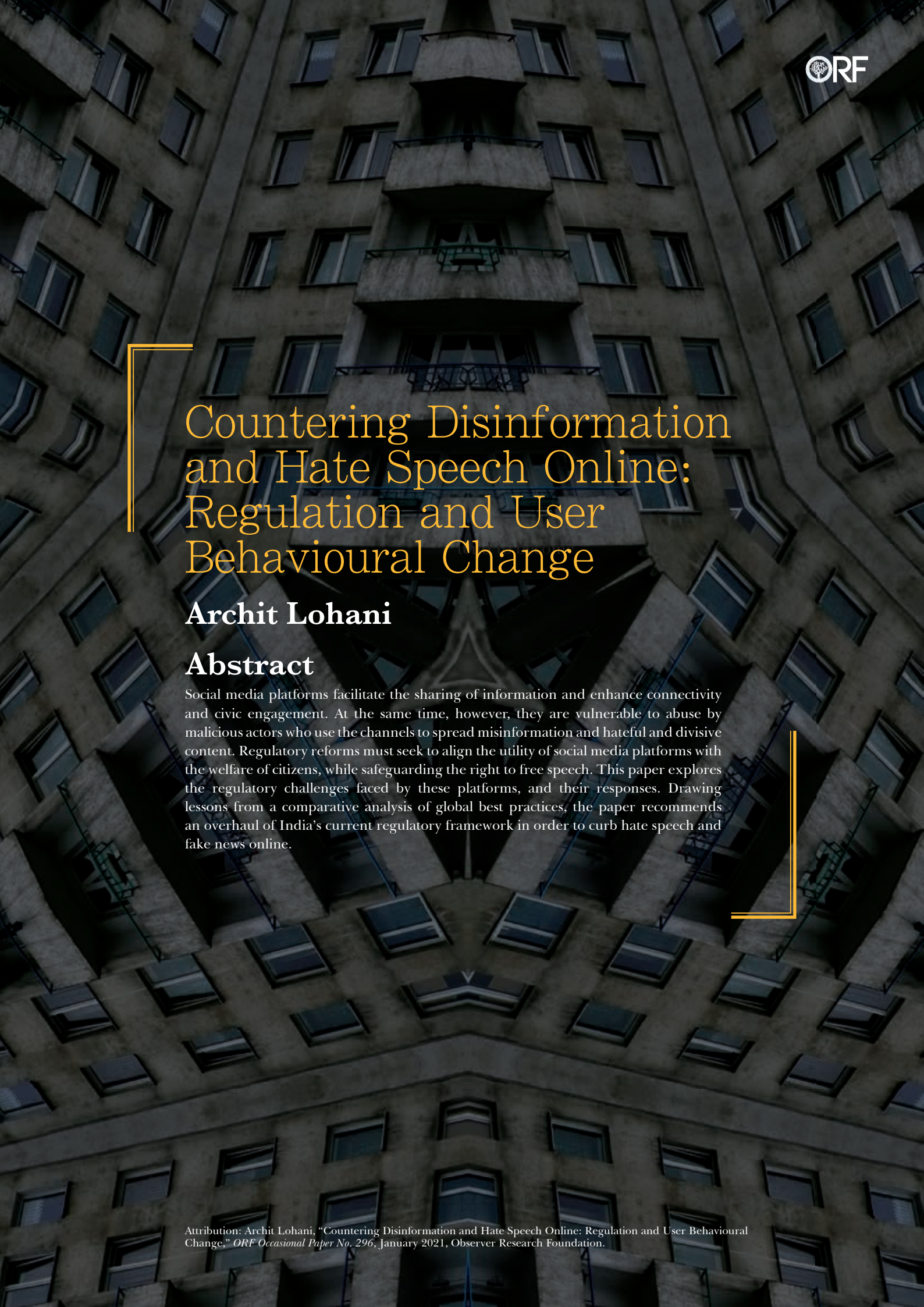
OBSERVER RESEARCH FOUNDATION

Occasional Paper



ISSUE NO. 296 JANUARY 2021



Countering Disinformation and Hate Speech Online: Regulation and User Behavioural Change

Archit Lohani

Abstract

Social media platforms facilitate the sharing of information and enhance connectivity and civic engagement. At the same time, however, they are vulnerable to abuse by malicious actors who use the channels to spread misinformation and hateful and divisive content. Regulatory reforms must seek to align the utility of social media platforms with the welfare of citizens, while safeguarding the right to free speech. This paper explores the regulatory challenges faced by these platforms, and their responses. Drawing lessons from a comparative analysis of global best practices, the paper recommends an overhaul of India's current regulatory framework in order to curb hate speech and fake news online.

Attribution: Archit Lohani, "Countering Disinformation and Hate Speech Online: Regulation and User Behavioural Change," *ORF Occasional Paper No. 296*, January 2021, Observer Research Foundation.

Introduction

The COVID-19 pandemic has shown how fast and wide information can spread: so fast, that the phenomenon was given the name, "infodemic".[1] For example, a documentary titled "Plandemic" featured the views of scientist, Dr. Judy Mikovits, who argued that the SARS-CoV-2 virus was "manufactured" and part of a scheme by pharmaceutical corporations to profit from selling vaccine. The documentary managed to rack up about 8 million views in 8 days.[2] Such disinformation compromises society's information-sharing ecosystem. In the time of a pandemic of massive proportions such as COVID-19, social media can be used as a tool of distrust to incite panic, confusion, and disharmony. The misuse of these platforms can have economic, psychological, and political impacts, both online and offline,[3] and can lead to discrimination and even violence.[4,5]

Behind the veil of protecting free speech, tech companies in India remain oblivious to such potential misuse. In the United States in early January 2021, platforms like Twitter provided a peek into their ability to counter disinformation, directing end-users to reliable sources, and suspending the account of former president Donald Trump, "due to the risk of further incitement of violence."[6] Many countries have initiated inquiries into the role played by these platforms in spreading extremist, hateful or fake content. Germany, Singapore, and France can now levy significant fines against platforms that fail to restrict illegal content after due process of notice and flagging. The United Kingdom (UK) is debating an Online Harms White Paper, while the European Commission has proposed two legislative initiatives—i.e., the Digital Services Act (DSA) and the Digital Markets Act (DMA) for the creation of regulatory mechanisms to counter online harms.

> "In a time of a crisis such as the COVID-19 pandemic, social media can be used to incite panic, confusion, and disharmony."

Introduction

This paper seeks to provide an understanding of the Indian government's approach to the massive regulatory challenges posed by social media platforms. A key question is whether the big corporations can be trusted to self-adjudicate while performing state-like functions to "reasonably" restrict speech. It is a relevant question, as Microsoft's Digital Civility Index has found that Indians are most likely to encounter misinformation online.[7] Other studies have found substantial growth in hate speech online.[8]

This paper analyses key issues with India's current regulatory framework to counter fake news and hate speech online, and draws lessons from the practices in other democracies. It argues that there is a need to modify the Indian approach by adopting a co-regulatory model. Governmental assistance through a penal and ethics code can complement the platform's capacity to restrict online harms. Strictly independent obligations are recommended to advance the current content moderation practices. A multi-stakeholder collaborative approach is imperative to counter the threat of even worse infodemics.

> "A crucial question is whether tech companies can be trusted to self-adjudicate."

Disinformation amid the Pandemic

The use of social media for peddling fake news and hate speech is not a new phenomenon. Before the pandemic, episodes of information dumping peaked during elections,[9] socio-political movements,[10] or to manipulate financial markets.[11] Amidst the COVID-19 crisis, it has become apparent that widespread fake news can threaten public health.[12,13,14] Public awareness is key in battling a health crisis. However, if the regulation of misinformation is concentrated in the hands of platforms or government agencies, it becomes susceptible to perception-alteration tactics.[15]

Early reportage of the pandemic in India tended to generalise Indian Muslims as wilful carriers of COVID-19.[16,17] The biased reportage began after a religious gathering turned out to be a super-spreader. A countrywide search for the attendees began, and authorities discovered 4,291 affected; there were 27 deaths. Soon after, the hashtag '#CoronaJihad' started trending in social media, in effect labelling a particular religious community as a collective danger to health and society.[18] To be sure, the global cyber-space is infested with all types of hateful narratives that are successfully fostered within society's digital communities.[19]

At the same time, social media platforms serve a function in creating a more Covid-literate Indian society. Health and executive authorities use these platforms to spread awareness and direct end-users to reliable sources. Facebook's MyGov Corona Hub, WHO Chatbot, and Corona Helpdesk Chatbot are a few examples of channels that have worked to counter misinformation and give access to first-hand awareness.

Facebook, for one, can be a highly powerful tool, with over 290 million users in India—its highest in the world.[20] In recent times, however, various governments have begun scrutinising the platform for what they allege to be its lackadaisical approach to hate speech.[21] In April 2020, Facebook flagged 50 million posts with warning labels; it argued that once a content is flagged, 95 percent of end-users do not access it.[22] Fact-checking organisations are also working to counter fake news campaigns, including, in India—reports about purported "cures" against COVID-19.[23,24,25] According to a *Reuters* report, between January and March 2020, there was a 900-percent increase in fact-checks related to Covid-19.[26] The same report indicates that a mere 20-percent of the total misleading content in that period had come from prominent public figures and enjoyed 69 percent of all engagement.

Social media platforms may have democratised the internet, but the same technology can create conflicts as it enables the proliferation of erroneous information at an unprecedented pace.[27] In a 2017 tudy of the US, a team

from the Massachusetts Institute of Technology found that fake news spreads, "farther, faster, and deeper" on these platforms.[28] The companies do not have adequate resources to quickly identify such content and remove them. The algorithms of these platforms work in such a manner that they record the user's past interactions and fill their feed with their identified interests; this facilitates targeted advertisements, from where the platforms earn their incomes.[29]

Of all the content in these platforms, those that are extremist, fake and populist are found to often garner high "interaction" numbers.[30] Facebook, for example, took down 40 million misleading posts in March 2020 alone, and another 50 million the following month.[31] For its part, Twitter challenged more than 1.5 million accounts from mid- to end-March.[32] However, *Network Contagion Research Institute* (NCRI) highlighted the role of smaller online communities and groups, that have become active avenues for targeted divisive content. While studying the "China-led bioweapon controversy", NCRI found profound anti-Asian and Sinophobic conspiracy theories on a small website, often being the source of disinformation and propaganda on other platforms like Reddit and Twitter.[33]

The question therefore, is if these platforms are plagued with manipulative and unethical content, can they still democratise? In theory, the principle invoking the so-called "*marketplace of ideas*" is a bedrock of free speech laws; it presumes that "truth" would prevail in a level playing field. In the marketplace, however, of platform-based ideas, the theory fails—[34] it seems social media is neither equal nor fair. A platform's design to maximise financial gains through data monetisation techniques can overwhelm "truth" with inbuilt susceptibility to sensationalist, viral, curated campaigns. Problematic speech is heightened due to the asymmetry of information and polarisation online.[35] To operationalise a model with necessary safeguards, the Indian approach must depart from excessive criminalisation or take-downs and instead adopt a holistic approach.

> "In theory, the 'truth' prevails in the marketplace of ideas. In social media, this assumption often fails."

Fake News

There is inadequate regulation of fake news under Indian law.[a] Due to the various types of fake news, their motivations, and the ways they are shared, the regulatory challenge is daunting. To combat fake news, the first imperative is to identify the different forms: '*misinformation*' is the inadvertent sharing of false content, while '*disinformation*' is deliberate sharing with an intent to deceive. Its sub-types, according to Claire Warlde of US-based First Draft, are misleading content; imposter content; fabricated content; false connection; false context; manipulated content; and satire or parody.[36] The misuse of platforms is also aided by so-called "flooding tactics", enabled through inorganic sharing or misusing emerging tools, like Deepfakes.[b] In a case study conducted by Blackbird. AI, a US-based company, a dangerous propaganda campaign was unveiled wherein 99.4 percent of 927,908 tweets in 47 different languages were found to have been inorganically disseminated in early February 2020.[37] A widely followed online religious leader and his followers waged a hashtag campaign, #NoMeat_NoCoronavirus, which sought to exploit religious bias through bots or impersonating user profiles.

Fake news thrives on dissemination through surplus or deficit information models. Under the surplus model, if enough users share the same information, it validates itself by a sheer numerical advantage, including when the gatekeepers of information (like journalists or politicians) validate it. A deficit information model suffers from a shortage of correct information, showcasing a lack of successful transmission between government, media, experts, and citizens. Information fails to trickle down and is manipulated en-route. Its impact is enhanced due to lack of access to correct information, limited prominence of fact-checking mediums, overwhelming nature, or the user's inability to comprehend its consequence. It also garners authenticity and virality through self-authentication,[38] and stylistic elements (like graphic content, alarmism, and imitation of authority).[39,40]

a Various Penal Code provisions like Sedition, promotion of religious enmity, defamation, public mischief, criminal intimidation etc. form the criminal jurisprudence of Fake News. The Information Technology Act lays down Cybercrime offences under Chapter XI of the Act. Section 79 of the Act grants intermediaries (social media platforms) limited immunity, under the Intermediary Rules an intermediary can only be held liable if it fails to act on a government or court order to take down the content in a period of 36 hours or aids the unlawful act. Section 66A which was earlier applicable to instances of fake news, was struck down by the Supreme court in 2015.

b Deepfakes are manipulated imagery or video wherein the subject's body or face is digitally altered to appear as someone else. These are typically used to spread malicious or false content in celebrity pornographic videos, revenge porn, fake news, hoaxes, and financial fraud.

The Indian Ministry of Electronics and Information Technology (MeitY) has recognised the potential for misuse of platforms and even broadly defined 'disinformation'.[c,41] However, the term is yet to be adopted under the IT Act or any provisions of the penal code. Section 505(1)(b) of the Indian Penal Code or Section 54 of Disaster Management Act, 2005, both provide broad recourse against cases which have severe consequences on public wellbeing; they are shorthanded, however, against the rapid pace of social media.[42] These regulations also lack precedent or uniform application against multiple types of fake news.

Hate Speech

Absolute free speech laws that protect against any type of censorship inadvertently render protection to hate speech as well. In India, hate speech is not profusely restricted, it remains undefined with appropriate IT Act provisions or a regulatory mechanism for online content. Absent appropriate codes or regulations for intermediaries,[d] those who tend to have a louder voice—such as politicians or celebrities—can harness this capacity to incite anger or divide communities without being threatened by any form of liability.

India's multiple laws on sedition, public order, enmity between groups, and decency and morality, broadly form the country's jurisprudence on what is known as "hate speech", without using the term itself. Following the unconstitutionality of Section 66A of the IT Act,[e] no provision under the IT Act currently aims to curtail either online or offline 'Hate Speech'. The most employed sections 153A and 295A of the Indian Penal Code (IPC) are also inadequate to deal with the barrage of online hate content. The Parliamentary Standing Committee has recommended changes to the IT Act by incorporating the essence of the Section 153A.[43] The report also suggests stricter penalties than prescribed under Section 153A due to the faster and wider spread of information in online

c Union Minister Law & Justice, Communications and MeitY, Ravi Shankar Prasad has said: "Fake news is a type of propaganda that consists of deliberate misinformation or hoax that is spread via traditional print and broadcast media or online social media. It can include text, visual, audio, data reports etc. Fake news is written and published with intent to mislead to damage an agency, an entity or a person to create disturbance and unrest often using sensational dishonest or outright fabricated headlines to increase readership, online sharing, and internet revenue. The typical attributes of fake news are that it spreads fast, is doctored, is incorrect, manipulated, intentional and unverified."

d 'Intermediaries' here refers to social media platforms.

e Section 66A criminalised sharing information online if it "causes annoyance, inconvenience or insult". In 2015, the Supreme court while adjudicating on the landmark Shreya Singhal v. Union of India case, struck down Section 66A as it imposed arbitrary and unreasonable standards on freedom of speech online.

spaces. It advocates criminalising "innocent forwards", for example, with the same strictness as the originator of the content.

This approach has the potential to create a "chilling effect": the overcriminalisation of online speech.[f] These provisions remain vaguely worded, and lack consistent interpretation of existing framework across courts. They also lead to the filing of minor cases that only overburden the country's courts and fail to act as redressal.[44] Unfortunately, unconstitutional provisions like Section 66A of IT Act continue to be invoked for prosecuting individuals. This provision is often misused due to its cognizable nature (allows arrest without a warrant) and highlights a signal failure between branches of the government.[45]

To address the proliferation of hate speech on social media, criminal law should not be the first resort, but the last. The promotion of non-regulatory tools—i.e., counter speech, fact-checking, and digital education—is an imperative.

> "The impact of 'fake news' is enhanced by lack of access to correct information, and the user's inability to comprehend its consequence."

f Although an abundance of legal obligations that vaguely restrict hate speech have failed to reduce the quantity of it.

Both government authorities and social media platforms alike, have been criticised for their failure to secure data and effectively regulate content. Many platforms, experts, and politicians have welcomed a government-led moderation of illicit content, with ample checks and balances against arbitrary imposition.[46,47] Human right groups and activists express skepticism against allowing any avenue for governmental intervention through either the arbitrary imposition of bans, content moderation, or internet shutdowns.[48] Another paradigm champions the principle of "self-regulation"—where the platform itself adjudicates on their user-policy and community guidelines. Self-regulation has largely been ineffective in preventing abuse of the platform and has garnered criticism in various democracies.[49,50,51,52] Indeed, both government-led moderation and self-regulation models have been operational worldwide and in India since the inception of social media. For example, more than 17,444 websites were blocked for promoting obscene content until 2019 by the IT Ministry[53] and similarly, Twitter took down 636,248 accounts in 2015-16 alone for disseminating extremist content.[54]

The difficult question concerning hate speech or fake news legislation pertains to the existing ethical-legal gap, the executive response departing from conservative understanding of online spaces and data. While disruptive technologies are evolving at a faster rate, the regulations fail to address gaps to deter unethical behaviour. The platforms alone are not equipped to oversee the task for a remodelled approach to counter manipulation and hate speech. Due to the overarching jurisdictional nature of these acts and easy multiplication, taking down content is not a silver bullet in countering hate speech and fake news.

> "Simply taking down content cannot be a silver bullet in countering hate speech and fake news"

While intensifying fact-checking and taking down inflammatory and fake content is a necessity, complex content-driven issues have emerged. For example,[55] Facebook in India has been accused of "ideological bias" by both Left- and Right-wing groups.[56] The Union Minister for Information Technology has called it "inherently biased" against people who support right-of-centre ideology and a "latest tool to stoke internal divisions and social disturbances."[57] Even as Facebook merely repeats its response—that it has an intermediary role as a free speech enabler—it has come under scrutiny again, most recently by the Parliamentary standing committee.[58] Moreover, a group of Facebook employees who identify themselves as Muslim@ wrote an open letter to the Facebook

leadership in August 2020 demanding greater transparency in taking down content; they also questioned why anti-Muslim and hateful content continue to find space on their platform.[59]

The lack of accountability and transparency calls for a rethinking of social media platforms' role and structure in order to counter their misuse.

Structure

The overregulation vs. under-regulation debate tends to overshadow the deeper and more inherent structural problems in the tech platforms themselves. The platform structure is driven by exploiting the disparities of wealth and power, as algorithms reward virality and interactions for monetary gains, even though they might be "divisive, misleading or false".[60] Platforms are also known to amplify certain types of users and content over others.[61,62] Platforms decentralise free speech, but "special" megaphones are provided to sensationalist ideologies or popular content. Its algorithmic nature creates and perpetuates an information divide, alienating communities with different subscriptions through echo-chambers and information silos. This has become obvious with the platform's incentive structure, which is driven by monetisation of user data, advertisement money, and constant engagement. For example, a few popular YouTube channels that earlier achieved "Creator Award" were inciting violence including rape but suffered less takedowns.[63] Platforms conveniently hide behind the garb of free speech enablers, with little responsibility, if at all.[64] Even as xenophobia, communalisation and racism have long existed in the real world, the susceptibility of social media platforms to misuse has magnified such ill-speech at a faster pace.

In India, social media platforms are not liable under any rules or regulations. They function under a regulatory vacuum and are not bound by any industry regulatory standards for the functions they dispense. None of existing news agency regulations, consumer protection laws, data privacy or other traditional sectoral regulations apply to these intermediaries. The bigger the platform, the less resources are available for user redressal against online harms. Functions like reporting or flagging: inflammatory and fake content are opaque, ineffective and inconsistent.[65] In some cases, the lack of competitors offers no incentive to provide better user protection, with majority of the targeted population already operating on the platform. Yet, the platforms continue to fail to recognise their public utility functions. In various countries, companies like Facebook, Google, and Twitter have succumbed to governmental crackdown, acknowledging the lack of preparedness and instituting duties to protect users.[66]

Role

The diminishing optimism amongst policymakers that tech companies can self-regulate has raised concerns on the unintended yet significant consequences that accrue from the current understanding of a platform's role. Under Section 79 of the IT Act, a government order can direct an intermediary to take down problematic content and not face any liability as publisher. Other than complying with state-determined policy, platforms also moderate harmful and unethical use through their own policies.[67]

Many argue that social media's role has evolved, as these platforms perform more than an intermediary role as content host.[68] Through the "community guidelines" that they set for their users, they are in effect performing the state-like legislative function of outlining rules for netizens. Similarly, they conduct state-like content moderation functions by taking down objectionable content.

The need for safeguards against problematic speech remains imperative and the role played by intermediaries must advance in tandem while revising its fiscal incentive structure and its public utility role.

> "Social media algorithms reward virality, even though such content may be divisive or false."

Best Practices in Regulation: Key Lessons

India can look to other countries for lessons on regulating social media platforms. Countries like Germany, the US, UK, Singapore, and Russia, for example, have put in place specific accountability standards assigned to these platforms. The approaches taken by these countries, however, are varied.

Australia

Australia's response to the proliferation of fake news is focused on countering foreign interference during elections. This was set up in the backdrop of growing concerns of potential Chinese interference and 2016 Russian interference in US elections. A task force was created—called the Electoral Integrity Assurance Task Force and led by the Home Affairs Department—to strengthen cybersecurity.[69] Two key frameworks are under discussion—the Fake News code[70] and the News Media Bargaining Code. The News Media Bargaining Code is highly controversial from the perspective of copyright and competition law, as it forces big companies such as Google and Facebook to realign market power by sharing multi-million-dollar revenues with newspaper and media companies for using their news content, to aid their financial futures.[71] Among the strictest penal provisions (in terms of fines) against extremist content—the Abhorrent Violent Material Act, 2019—make tech executives liable for imprisonment up to three years and fines for the platform can run up to 10 percent of a company's global turnover.[72]

France

French President Emmanuel Macron stated in November 2018, "if we do not regulate Internet, there is the risk that the foundations of democracy will be shaken."[73] France applies broadcasting standards for TV and radio stations to social media through the Higher Audiovisual Council (CSA), the independent media regulator. It aims to assure broadcasting communication freedom in the country and has been tasked with the responsibility of enforcing state policies on fake news and hate speech. The French government has identified terrorist content and child pornography as non-negotiables, demanding platforms to take down such content within one hour of being notified. France has initiated a crackdown on electoral misinformation to avoid foreign interference by empowering the CSA to unilaterally revoke broadcasting rights of news and radio outlets that function "under the control or influence of a foreign state" to disseminate misinformation.[74] At the same time, however, the law has faced criticism for granting executive and administrative powers to CSA.[75]

Table 1:
Regulatory standards for illicit content

	Australia	France	Germany	Singapore	UK	US
Speech a fundamental right?	No, but implied	Yes, with restrictions	Yes, with restrictions	Yes, with restrictions	Qualified Right	Absolute right
Intermediary obligations for enforced take-downs	48-hour take-down notices for extremist content	- Terrorist content or Child pornogrphy to be taken down in 1 hour of notification. - Hate speech, violent, racist, and sexual harassment-related content to be taken down in 24hour period post notification	24 hours to take-down "blatantly illegal content" and "other illegal content" requires deletion within a week of notice	As soon as possible, upon issuance of a direction	NA	No liability to take down hate speech or fake news
Enforcement Authority	E-Safety Commissioner and Electoral Integrity Task Force	Higher Audio-visual Council (CSA)	German Ministries and Federal Office of Justice	Ministers and Infocomm Media Development Authority (IMDA)	Independent Regulator	Copyright Office, USA (only under S. 512 Digital Millennium Copyright Act)

Liability-Fines or imprisonment for breach	- Imprisonment up to three years and financial penalties worth up to 10% of a company's global turnover. - Platforms can be fined AUD 525,000 and individuals can also be fined up to AUD 105,000.	- Fines up to 4% of global revenue upon failure to take down notified content or fine of up to €1.25m	- Social networks with more than 2 million users in Germany. -Individuals may be fined up to €5m and companies up to €50m	- Fake info. shared by a "malicious actor", fine of up to $37,000 or five years in prison. - If "an inauthentic online account or a bot" shares it fine up to $74,000 and potential 10-year jail term. - Platforms liable for fine up to $740,000 and jail sentences up to 10 years.	Substantial fines, appropriate to turnovers and code violation	NA
Voluntary Transparency mechanisms	Reports from Australian Communications and Media Authority and E-Safety Commissioner[76]	Independent reports on content moderated and advertisements from both CSA and Platforms	Publish semi-annual reports that detail content moderation procedures and their statistical analysis	Publications of directions online or in government gazettes	Publication of Reports by platforms and Regulator Scrutiny by the parliament	NA
Penal code legislations on Hate Speech or Fake News	Abhorrent Violent Material Act, 2019 Enhancing Online Safety Act, 2015	Lutte contre la haine sur internet (Fighting Hate on the internet), 2020	German Criminal Code defines and penalises "incitement to hatred" and other 21 statutes that now fall under the Network Enforcement Act, 2018	Protection from Online Falsehoods and Manipulation Act, 2019 (POFMA)	Proposed framework under-Online Harms White Paper, 2019	Honest Ads Act (Proposed)

Non-legal Media ethics or guidelines on Fake News or Hate Speech	MEAA *Journalist Code of Ethics*	Yes, various ethics codes broadly discussing fake and provocative content	NA	Yes, broadly covers ethical reporting. Journalists' Code of Professional Conduct	National Union of Journalists Code of Conduct	Society of Professional Journalists Code of Ethics
Advertisement Regulation or Fact-checking	News Media Bargaining Code[77] (Proposed Code to share advertisement revenue)	Compulsory disclosure of price and promoter details for sponsored content and campaign ads	NA	Disclose sources of political advertisements	Proposed under Online Harms White Paper	Under proposed Honest Ads Act, platforms to keep copies of ads, publish them and maintain details of publisher
Signatory to Global or Regional Commitments	Yes	Yes	Yes	No	Yes	No

Source: Author's own, using various open sources.

France has had a strict Hate Speech law in effect since 2005 and has passed a stricter legislation for online application against hateful speech (Lutte contre la haine sur internet- An Act to Combat Hateful Content on the Internet- Avia law)—it requires social media platforms to take down objectionable content within 24 hours of being notified. Once a political party, public authority, or an individual has filed a specific request, a judge is authorised to act "proportionally" but "with any means" to halt the dissemination of the fake news in question. The request is to be acted upon within 48 hours and a similar timeframe is applicable during an appellate request. France has also demanded transparency reports about the algorithmic functions of social media platforms, as well as financial transparency for sponsored content by disclosing the identity of promoters. The CSA is also tasked with publishing reports on data, mechanisms adopted, and their effectiveness.

Under France's Digital Republic Act, "a decision taken based on an algorithmic treatment" by public sector bodies falls under a transparency rule called "right to explainability". The citizens upon request can seek an explanation on the rules and its "principal characteristics."[78]

Germany

Germany's NetzDG has jurisdiction over high-user platforms; it can order these companies to take down illegal content or else face significant financial liabilities.[79] Human rights activists, however, are concerned about "over-removal" and privatised enforcement. The group, Human Rights Watch, has warned that the current regulatory framework "turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[80] Indeed, while the law was aimed at incentivising the quick take-down of illegal content, it has led to censorship. To begin with, the definition of "unlawful content" remains highly debated in instances of blasphemy and hate speech; "defamation" and "insult" are vaguely defined as well.[81] Another counterproductive outcome can be explained as the 'Streisand Effect':[g] prominent cases of deletion may fuel anti-government sentiments by garnering more publicity for the deleted content.[82] Moreover, any extremist or banned user can still easily migrate to smaller platforms who are not liable under NetzDG.[83]

g Streisand effect refers to a phenomenon, often via the internet, whereby an attempt to supress or censor information has a paradoxical effect and draws more attention to the censored content.

Singapore

Singapore's POFMA Act is arguably one of the most comprehensive misinformation laws globally. It is widely criticised by human rights groups and free speech advocates due to its vulnerability to governmental interference.[84,85] POFMA covers "false facts" or "misleading statements", but not satire, parody, opinions, and criticisms. Under POFMA, various directions can be issued to platforms and users before dispensing fines and imprisonment. These directions include the following:[86] (i) Correction Direction: the party who communicated the falsehood is tasked to issue a public notice stating the false nature of information. The notice is published in a newspaper or a designated area on the platform. (ii) Stop Communication Direction: the publisher is directed to ensure the information is no longer available for proliferation, with the direction being published in a government gazette. (iii) Targeted Correction Direction: the platform is directed to send a correction notice to all end-users who accessed the falsehood. (iii) Disabling Direction: the internet service provider is directed to disable access to the falsehood.

Another key aspect of the law is gross penalisation of inorganic proliferation methods like the use of bots.[87] It is strictly prohibited to allow another person to manipulate or create bots to spread misinformation online, and violators face imprisonment of up to 10 years. Amendments under the Protection from Harassment Act (POHA) are under discussion to provide recourse to victims of harassment due to falsehoods.

United Kingdom

The UK released its Online Harms White Paper[88] in April 2019, which is based on installing an overarching principle of "duty of care" towards end-users. The proposed Online Harms Paper is not limited to social media platforms; it extends to file-sharing sites, discussion forums, and e-commerce websites– which are mandated to take "responsible steps" in the direction of user safety, transparency and tackling harms. It calls attention to the need to empower users by developing their digital skills for countering misinformation and other forms of online harm. An independent regulator will be appointed with the power to issue substantial fines for social media platforms and their senior members. The paper dictates a risk-based approach, which aims proportionate regulatory action to counter harms with greatest impact on individuals. It is overwhelmingly positive and has initiated a discourse to prioritise before hurried legislative battles. However, the scope of the paper remains wide enough to cover various legal and ethical norms. At present, the National Security Communications Unit is tasked to combat disinformation campaigns by state actors and others during elections.[89]

United States

The US guarantees absolute freedom of speech and has no legislation on hate speech. Through various judicial pronouncements, courts have consistently argued that even hate speech is protected under the First Amendment.[90] The Russian interference during the 2016 presidential elections served as a turning point in the threat perception towards misinformation and manipulative sponsored content.[91] In the 2020 presidential elections, big platforms like Instagram, Facebook, and Twitter took proactive steps to flag questionable content and direct end-users to reliable sources.

The Honest Ads Bill, introduced in 2017, seeks to extend the purview of ad regulations for broadcast TV and radio on social media platforms.[92] It expands source disclosure requirements as "public communications" or "electioneering communications", demanding source disclosures and the information contained within ads. The bill also requires social media platforms to maintain publicly available records about qualified political advertisements. Various state governments across the US have initiated compulsory digital and media literacy initiatives for schools. The security implications arising from deepfakes has also been highlighted by members of both the Republican and Democratic parties.[93] Section 230 of the Communications Decency Act (1996) provides the intermediaries with broad immunity: they are not liable to take down any offensive content that upholds free speech. Only in cases of copyright infringement under Section 512 of the Digital Millennium Copyright Act (1998), can the state employ a "safe harbour" approach where it provides the conventional "notice and takedown" method.

In different parts of the world, disinformation and hate speech related to the COVID-19 pandemic have been addressed with specific responses. The pandemic serves as another case study for Indian jurisprudence, and can assist in reducing the ethical-legal gap. Online speech regulation necessitates improved understanding of online harms, their ill-effects on a democracy to sharpen the local response.

The Indian response must be driven by four guiding principles: (1) Accountability and transparency over decision-making by tech platforms, state and non-state actors; (2) Ensure consistency and collective will by encouraging inclusive stakeholder engagement for all decision-making processes; (3) Respect human rights standards and habituate humane application of tech. Incentivise innovative adoption of redesigned tech products that pre-empt and provide safeguards from online harms; (4) Legal certainty for consistent application and execution of duties and rights of stakeholders.

The lessons from the global best practices discussed in this paper should serve as basic elements for India's regulatory mechanism: (i) Institute an independent regulator to oversee compliance with fake news and hate speech codes that will be adopted; (ii) Proportional, necessary and minimal interventions from the government and platforms with effective and consistent application of their duties; (iii) An inclusive and ethical Code of Conduct developed in consultation with all stakeholders to realign platform's fiscal-driven-incentives with public interest; (iv) Democratic application of penal and non-penal standards of existing laws; (v) Periodic review policies to improve effectiveness; (vi) Encourage transparency by commissioning open-source research with periodic reports from regulator, platforms, civil society organisations and academia; (vii) Avoid creating any barriers or strengthening any dominant positions by large incumbents; (viii) Promote digital education initiatives and workshops to acquire necessary skills from a young age; (ix) Redressal and appellate mechanisms to provide support to any wrongful application of standards, take-downs or breach.

Specific Recommendations

In any strategic intervention against misinformation and hate speech, the tech platforms are bound to play the biggest role. There should be continuous collaborative engagements within the industry, along with state and non-state actors. While the creation of charters or codes that define each stakeholder's duties and rights will be a lengthy process, a pre-emptive plan cannot be delayed further. This can enable the creation of voluntary multi-platform and multi-stakeholder initiatives. The Code of ethics and voluntary audits are another welcome by-product of these collaborative measures. Issue-specific methods of advertisement rules for transparency and media guidelines or ethical codes also

aim to strengthen industry standards. Some shared responsibilities between the stakeholders have already been outlined but limited action has been taken to counter online harms.[94] Platforms have deployed minimal resources to take down blatantly illegal content, as they lack real-time local responders who are well-versed in Indian languages.[95] Even their community guidelines are globally uniform and limited due to implementational and definitional challenges locally. Therefore, the government and the tech platforms should complement other information gatekeepers like media and politicians.

First, the government must speed up defining and paving a way for consensus towards a legal framework against problematic social media content.

Definitions

The Indian challenge to garner consensus and counter 'hate speech' and 'fake news' extends to their understanding in real/offline world. Both remain undefined under any domestic legal mandate, including the IT Act. By building consensus on key elements, the legislature can assist the platforms in initiating a countermovement by consistent interpretation and implementation of the law. It will also discourage platforms from adjudicating on what is acceptable speech and avoid faulty implementation. The Hate Speech Law Commission report, which suggests that the scope of regulation should not be limited to "incitement of violence" but also prohibit advocacy of hate; and incitement to hostility or discrimination.[96] To determine penal implications for 'disinformation' and 'hate speech', 'impact' and 'intention' are key elements to discourage its proliferation. However, the process of defining and introducing penal provisions must avoid ambiguous terminologies. Cohesive definitions can complement adoption of voluntary codes amongst platforms, as well as update media code and Representation of the People Act (1951) against information manipulation during elections. Potential overcriminalisation must be prevented, the legislature can identify and agree on key elements to facilitate consensus building and build safety nets around ethical codes. Criminal law should not be the first response but the last resort when state or court intervention is imperative.

Legislative Framework

A limited but necessary legislative support could help generate consensus with a much more effective regulatory mechanism.[97] Countries like France and Germany have pre-defined the limited scope of government role to avoid any arbitrary intervention and have even appointed independent regulators or independent judicial members to dispense moderation objectives. Fake News

and Hate Speech codes must be founded, and authorities must explore the scope of an independent regulator. The regulator's objectives should be to assist the application of government policies and serve as a forum for redressal against the platforms for arbitrary takedowns. The regulating body by design must not be subjected to governmental or platform discretion that can have the potential to be selective, arbitrary, non-transparent, and utilised to smother dissent or the genuine exercise of freedom of speech. The appointment of personnel in the body must also be inclusive, from a wide-ranging pool of judiciary, civil society, free speech activists, and technical experts. The code must curate appropriate guidelines to define platform duties, rights, and issue take-down notices through the regulator.

The code must differentiate between infrastructure information intermediaries (such as Internet Service Providers) and content information intermediaries that host content (such as platforms). A set of content-neutral regulations must apply to both intermediaries with a duty-of-care imposed on content information intermediaries. The under-discussion Intermediary guidelines could reflect recommendations that will define the role of the regulator, but an overarching code/framework is needed to restrict fake news and hate speech. The Code must foster transparency as a pillar and enable publication of takedown notices requested by the regulator. The Code must also provide recourse for due process and charge reasonable fines upon breach by any platform or individuals, and guarantee safeguards against potential abuse of notice periods, proper procedure, and wrongful takedowns.

The regulatory body can also assist in formulating and strengthening industry standards for the under-discussion Information Trust Alliance (ITA). ITA envisions a collaborative effort to bring all platforms to assist in an inter-platform exchange, it advises effectiveness through knowledge sharing. ITA can provide invaluable contribution in creating guidelines that highlight local problems, bottlenecks, effectiveness in diverse regional languages, study local trends, revaluate right to explanation and share data in a transparent manner.

Sponsored content and political ads should also be compulsorily fact-checked while maintaining directories of promoters, amount paid, and source. The use of inorganic amplification methods like bots to propagate hate agendas must be charged with fines in case of severe social impact. Gatekeepers of information like media houses and politicians owe a higher 'duty-of-care' as they yield a significant impact on local perceptions. Penal fines that are proportionate and consistent against repeat offenders must be employed.

However, learning from Singapore's example, inconsistent and vaguely worded law with strict government-centric censorship will be a step in the opposite

direction. Singapore's fake news law is unconstitutional by Indian standards and criminalisation-heavy. Indian policy and law should instead reflect the necessary behavioural change that is required to address the awareness with a modified platform-incentive model. Internet shutdowns are not feasible measures as they have economic and social implications, worsening enmity and lack of correct information.

Social Media Platforms

To effectively safeguard against and mitigate the impacts of ill-speech, countries should have better fact-checkers and authorities committed to respond to public interest. Real-time takedowns can have a positive impact as it significantly develops resistance and reduces impact of misinformation to percolate into local, deeper, and different forms. However, taking down problematic content is not an absolute safeguard in itself as the nature of platforms dictate that time is not the only determinant when problematic speech is published online. For example, a tweet might be taken down but still may have a direct and viral effect through its screenshots on different platforms. Ill-speech's omnipresence and negative impact remain an issue but the lack of awareness amongst end-users showcases disengagement with the primary stakeholder.

Other than adopting voluntary responsible measures, this paper suggests a four-step model to be implemented by social media platforms to counter problematic content online.

- **Step 1-** Identification: Identify fake news, extremist content, or hate speech according to the definitions or elements. While upholding anonymity, the platform must flag it in a specific manner which communicates its problematic/unreliable nature to the end-user.
- **Step 2-** Disallow Proliferation: While the content may continue to exist online, it should not only be flagged but platforms should disable any type of proliferation further, which includes content's algorithmic prioritisation it may usually employ. Any blatantly problematic content should not be promoted for interactions—i.e. such content should not be available on user feeds. Disable or provide safeguards for like, share, retweet, upvote, or other interaction methods for the same.
- **Step 3-** Issue interaction warnings: One of the most important criticisms against platforms is related to their collection of data. However, this issue can be employed to initiate a vital process of digital education and awareness. Since platforms employ interaction data, they must issue warnings to all end-users who have encountered problematic content

before it was flagged or identified. All end-users who shared or promoted such content must be sent personal notifications on the respective platforms, about the problematic nature of the content. Directing them to more credible sources and encourage corrections on smaller groups. Similarly, the publishing end-user must be provided with necessary reasons for flagging or taking down the published content. They, along with other interactors, should be encouraged to fact-check before sharing published content. This can help in disseminating digital education and necessary online skills to empower the end-users themselves.

- **Step 4 -** Provide a better recourse mechanism: In terms of reporting fake and hateful content, platforms should be user-friendly with timely action and response. This would require a wide expansion of resources employed at the intermediary's offices. Recourse against wrongful takedowns should be formalised and direct end-users to such mechanism if their content is taken down.

Publishing and providing access to relevant data is key for studying local trends in speech manipulations. Through transparency reports, data should be made publicly available for better preparedness and constant engagement with all stakeholders to improve effectiveness. Platforms should also use disruptive technologies like artificial intelligence, which have been touted to industrialise mass-level identification of such content. They can also assist political institutions and media houses to countercheck information before posting it. For example, Facebook had introduced a Corona Helpdesk Chatbot in messenger to counter COVID-19 related misinformation online.[98] Although AI has its limitations as well, for example, it has been unable to identify AI-generated misinformation itself and its use by platforms to flag problematic content, has often flagged non-spam or news content.[99] This highlights technological gaps while employing AI. Various fact-checking initiatives exist outside social media platforms but their use and inculcation on platforms can be made user-friendly through building partnerships. Lastly, platforms must be incentivised to adopt innovative redesigned tech products which prioritise humane use, prevention of harm, and provide safeguards from harms that may accrue through their misuse.

The end-user

The end-user who is exposed to problematic speech is the most important, yet, disengaged stakeholder. As social media platforms aim to ensure their risk-averseness, they must be a part of the safeguarding process. Inculcating anti-fake and anti-hate behaviour is key in empowering individual entities to discredit and stifle such speech as the first response. Counter-speech means

to engage directly against the negative speech using positive expression. In online discourse, multiple end-users already assist towards "toning down the rhetoric"[100] by providing clarifications to dubious claims, directly engaging with facts to counter hateful messages, call out fake reportage, employ humour or dissent through memes and caricatures. Most importantly, counter-speech is also a decentralised method for engaging with ill-speech, it reduces negative impact and counters the intention behind it. Radical change in user behaviour incentives is imperative to promote counter speech, disincentivise blind sharing, inculcate fact-checking, and calling-out disfavoured speech. There is no countermeasure against blind sharing but to effectively reduce its impact, it is necessary to empower individuals with easy access mechanisms to halt negative proliferation. A red flag should be easy to raise against extremist speech while imparting understanding against different types of ill-intentioned speech policies, potential impact, and its omnipresence.

Influential political, religious, or social leaders often misuse the significant online traffic while disseminating information. Trends suggest they enjoy widest amplification powers. For counter speech, incentivise best-placed "credible messengers" like politicians, media, celebrities, experts, and other individuals (with a lot of followers or authenticating blue ticks) to educate against fake news. The impact of their sharing ecosystems has a higher impact on end-users. For example, Facebook's CEO Mark Zuckerberg justified the need to take down objectionable speech employed by Indian politician Kapil Mishra[h] that intentionally "incites violence".[101] Hate speech originating from a politician has even more detrimental impact in a volatile environment. Political institutions must recognise and impose a 'duty-of-care' amongst politicians for humane and ethical use of these platforms.

Media

In countries like the US and France, some broadcast media regulations are made applicable to advertisement and sponsored content through declaration of sources to avoid foreign intervention in local politics. In India, independent bodies like the Editor's Guild and Press Trust, should reset industrial standards and wield best practices to discourage competitors from providing space to fake, propaganda, and hateful narratives**.** A fine structure against repeat offenders both online and in broadcast media should be applicable, with terms of suspension and even revocation of broadcasting license in extreme cases.

h During anti-Citizenship protests in 2019, politician Kapil Mishra from the Bhartiya Janta Party made comments that were inciteful and allegedly resulted in the Delhi riots. Viral posts and videos were deleted from Twitter and Facebook.

Unlike social media platforms, in case of offline broadcast media a key feature of being able to compartmentalise news through subscription is absent. In a recent development, Supreme Court issued a pre-broadcast injunction against Sudarshan News for communally charged and inciting news show titled 'UPSC Jihad'. SC expressed its concern of an ineffective regulatory model for broadcast media and questioned lapses by Press Trust of India to successfully restrict hate. The overlapping nature of broadcast and digital media also resulted in the hashtag 'UPSCJihad' trending on Twitter. Such lapses can be pre-empted within a co-regulatory platform model, that ensures not just compliance with ethical standards but is complemented by an incentive structure to disseminate facts-based reportage. The right to reply should not be absent in case of sensationalised speech.[102]

> "Any effective intervention against hate speech and disinformation will need collaboration between tech companies, the state, and non-state actors."

Conclusion

Archit Lohani *is a Research Assistant at ORF.*

The proliferation of hateful and harmful information online has become so widespread, that the United Nations, through Secretary-General António Guterres has underlined the need to counter what he called a "tsunami".[103] Guterres has also called attention to the scapegoating and "disproportionate effects" on vulnerable groups: they lose access to healthcare, and become victims of targeted violence, rising stigmatisation, and heavy-handed security responses. Proportionate restrictions and guidelines have not developed in tandem with the expansion of the power of the social media platforms. Users are exposed to problematic content without any warning, awareness, and skills to counter what they are against. A social media platform's primary incentives remain their massive user base, increased sharing and connectivity, and profits.

The evolving nature of online harm necessitates an appropriate response from the regulatory bodies. Additionally, the dissimilar nature of the pandemic, compounded by the weaponisation of information-sharing models, benefit few and negatively affect large populations.

Intervention in this regard is necessary. However, any restriction cannot be vaguely or hastily drafted to allow selective and arbitrary application by either the tech companies or government authorities. A balance must be found in this regard, defining the roles of various stakeholders in a co-regulatory model. Principles of proportionality, accountability and transparency must serve as cornerstones of any speech regulation. The pandemic case study offers another learning curve for collaborative countermeasures.

Hate speech is provocative and divisive, and in extreme scenarios where it has remained unchecked, has been responsible for terrorism and genocide. With newer tools to weaponise and sensationalise enmity, it must not be protected under the realm of free speech doctrine. Similarly, misinformation ("fake news") also has the potential to affect human safety and public health, and instigate violence. If fake news and hate speech continue to proliferate at the current rate, they pose threats to the democratic ecosystem. India must work to devise an all-stakeholder model to counter the weaponisation of online content, before it further widens societal faultlines.

Endnotes

1 World Health Organisation, *Novel Coronavirus(2019-nCoV) Situation Report,* World Health Organisation, February 2, 2020, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf.

2 Daniel Funke, "Fact-checking 'Plandemic': A documentary full of false conspiracy theories about the coronavirus," *PolitiFact.com,* May 7, 2020, https://www.politifact.com/article/2020/may/08/fact-checking-plandemic-documentary-full-false-con/.

3 Annie Gowen and Manas Sharma, "Rising Hate in India", *The Washington Post*, October 31, 2020, https://www.washingtonpost.com/graphics/2018/world/reports-of-hate-crime-cases-have-spiked-in-india/?utm_term=.2712573765d9.

4 James Weinstein, "Hate Speech, Pornography, And Radical Attacks On Free Speech Doctrine," *Routledge*, September 16, 1999, https://www.routledge.com/Hate-Speech-Pornography-And-Radical-Attacks-On-Free-Speech-Doctrine/Weinstein/p/book/9780813327099.

5 UN News, "Hate speech exacerbating societal, racial tensions with 'deadly consequences around the world', say UN experts," September 23, 2019 https://news.un.org/en/story/2019/09/1047102.

6 Twitter Inc., "Permanent suspension of @realDonaldTrump," January 8, 2021, https://blog.twitter.com/en_us/topics/company/2020/suspension.html

7 Microsoft, "*Microsoft Releases Digital Civility Index on Safer Internet Day*," Microsoft News Center India, February 5, 2019, https://news.microsoft.com/en-in/microsoft-digital-civility-index-safer-internet-day-2019/.

8 Maya Mirchandani, Dhananjay Sahai and Ojasvi Goel, "Encouraging counter-speech by mapping the contours of hate speech on Facebook in India," *Observer Research Foundation*, March 13, 2018.

9 Sidharth A., "How misinformation was weaponized in 2019 Lok Sabha election - A compilation", *ALT News,* May 20, 2019, https://www.altnews.in/how-misinformation-was-weaponized-in-2019-lok-sabha-election-a-compilation/.

10 Snigdha Poonam & Samarth Bansal, "*Misinformation Is Endangering India's Election*", *The Atlantic*, April 1, 2019, https://perma.cc/Y39M-KGWU.

11 Shimon Kogan, Tobias J. Moskowitz, and Marina Niessner, "Fake News in Financial Markets," November 11, 2020, *SSRN,* https://ssrn.com/abstract=3237763.

12 University of East Anglia, "Fake news makes disease outbreaks worse, research shows," *University of East Anglia*, February 14, 2020, https://www.uea.ac.uk/news/-/article/fake-news-makes-disease-outbreaks-worse-research-shows#:~:text=The%20rise%20of%20fake%20news,the%20COVID%2D19%20Coronavirus%20outbreak.

13 Przemyslaw Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek, "The spread of medical fake news in social media – The pilot quantitative study," *Health Policy and Technology, Volume 7, Issue 2, June 2018, Pages 115-118,* https://www.sciencedirect.com/science/article/abs/pii/S2211883718300881.

14 Cristina M. Pulido, Laura Ruiz-Eugenio, Gisela Redondo-Sama, and Beatriz Villarejo-Carballido , "A New Application of Social Impact in Social Media for Overcoming Fake News in Health," *International Journal of Environmental Research and Public Health*, April 3, 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7177765/.

15 Bhaskar Chakravorti, "The countries that trust Facebook the most are also the most vulnerable to its mistakes", *The Conversation,* March 27, 2018, https://theconversation.com/the-countries-that-trust-facebook-the-most-are-also-the-most-vulnerable-to-its-mistakes-93706.

Endnotes

16 Jayshree Bajoria, "CoronaJihad is Only the Latest Manifestation: Islamophobia in India has Been Years in the Making," *Human Rights Watch*, May 1, 2020, www.hrw.org/news/2020/05/01/coronajihad-only-latest-manifestation-islamophobia-india-has-been-years-making.

17 Asim Ali, "Coronavirus was a test of secular nationalism. Then Tablighi Jamaat became the scapegoat," *The Print*, April 1, 2020, www.theprint.in/opinion/coronavirus-test-of-secular-nationalism-tablighi-jamaat-became-scapegoat/392764/.

18 Giovanni Luca Ciampaglia and Filippo Menczer, "Biases Make People Vulnerable to Misinformation Spread by Social Media," *The Conversation*, June 21, 2018, www.scientificamerican.com/article/biases-make-people-vulnerable-to-misinformation-spread-by-social-media/.

19 Savvas Zannettou, Jason Baumgartner, Joel Finkelstein and Alex Goldenberg, "Weaponized Information Outbreak: A Case Study on COVID-19, Bioweapon Myths, and the Asian Conspiracy Meme," *Network Contagion Research Institute*, 2020, https://ncri.io/wp-content/uploads/NCRI-White-Paper-COVID-19-13-Apr-2020.pdf.

20 J. Clement, "Facebook: Most users by country," *Statista.com*, July 24, 2020, https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

21 Billy Perrigo, "Facebook Was Used to Incite Violence in Myanmar. A New Report on Hate Speech Shows It Hasn't Learned Enough Since Then," *Time*, October 29, 2019 https://time.com/5712366/facebook-hate-speech-violence/.

22 Guy Rosen and VP Integrity, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," *Facebook*, April 16, 2020, https://about.fb.com/news/2020/04/covid-19-misinfo-update/.

23 Julie Posetti and Kalina Bontcheva, "Disinfodemic Deciphering COVID-19 disinformation," *UNESCO*, 2020, https://en.unesco.org/sites/default/files/disinfodemic_deciphering_covid19_disinformation.pdf.

24 Shachi Sutaria, "Boiled Garlic Water For Treating Coronavirus? Not Really," *Boom Live*, February 1, 2020, www.boomlive.in/health/boiled-garlic-water-for-treating-coronavirus-not-really-6737.

25 Anmol Alphonso, "FAKE: Letter Claiming Health Ministry Declared Holidays In 4 States," *Boom Live*, March 13, 2020, www.boomlive.in/fast-check/fake-letter-claiming-health-ministry-declared-holidays-in-4-states-7209.

26 Dr. J. Scott Brennen, Felix Simon, Dr Philip N. Howard and Professor Rasmus Kleis Nielsen, "Types, sources, and claims of COVID-19 misinformation," *Reuters Institute*, April 7, 2020, https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

27 Roli Srivastava, "Facebook a 'megaphone for hate' against Indian minorities," *Reuters*, October 30, 2019, https://www.reuters.com/article/us-facebook-india-content/facebook-a-megaphone-for-hate-against-indian-minorities-idUSKBN1X929F.

28 Soroush Vosoughi, Deb Roy and Sinan Aral, "The spread of true and false news online," *Science*, Vol 359, Issue 6380, March 9, 2018, https://science.sciencemag.org/content/359/6380/1146.

29 Kevin Newman, "Facebook's Political Algorithm and Extremism Silos," *The Medium*, May 27, 2020, https://medium.com/@Touvan/facebooks-political-algorithm-error-and-tribal-extremism-c101ae0fa13.

30 United Nations, "UN chief Global Appeal to Address and Counter COVID-19 Related Hate Speech," Countering COVID-19 Hate Speech, https://www.un.org/sg/en/node/251827.

Endnotes

31 IANS, "Facebook flagged 50 million misleading COVID-19 posts in April," *IndiaTV News*, May 13, 2020, www.indiatvnews.com/technology/news-facebook-flagged-50-million-misleading-covid-19-posts-in-april-616897.

32 Vijaya and Matt Derella, "An update on our continuity strategy during COVID-19," April 1, 2020, https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html.

33 Savvas Zannettou, Jason Baumgartner, Joel Finkelstein and Alex Goldenberg, "Weaponized Information Outbreak: A Case Study on COVID-19, Bioweapon Myths, and the Asian Conspiracy Meme," *Network Contagion Research Institute*, 2020, https://ncri.io/wp-content/uploads/NCRI-White-Paper-COVID-19-13-Apr-2020.pdf.

34 Richard Stengel, "Why America needs a hate speech law," *The Washington Post*, October 29, 2019, https://www.washingtonpost.com/opinions/2019/10/29/why-america-needs-hate-speech-law/.

35 Dan M., "The 'Marketplace of Ideas' is a Failed Market," *The Medium*, February 14, 2017, https://medium.com/@danmcgee/the-marketplace-of-ideas-is-a-failed-market-5d1a7c106fb8.

36 Claire Wardle, "Fake news. It's complicated.," *First Draft*, February 16, 2017, https://firstdraftnews.org/latest/fake-news-complicated/.

37 Blackbird.ai "COVID-19 Disinformation Report- Volume 1," *Blackbird.ai,* February 19, 2020, https://www.blackbird.ai/blog/2020/02/19/covid-19-coronavirus-disinformation-report/.

38 Gautam Prakash, Ravinder Kumar Verma, P. Vigneswara Ilavarasan, and Arpan K., "Authenticating Fake News: An Empirical Study in India," *ICT Unbounded, Social Impact of Bright ICT Adoption*, May 2019, https://www.researchgate.net/publication/333700939_Authenticating_Fake_News_An_Empirical_Study_in_India.

39 Joyojeet Pal, & Syeda Zainab Akbar, "It's Essential to Sift Through Hate-Driven Misinformation on Coronavirus," *The Wire*, March 25, 2020, https://thewire.in/media/coronavirus-misinformation-hate-fake-news.

40 Devin Coldewey, "False news spreads faster than truth online thanks to human nature," *TechCrunch*, March 9, 2018, https://techcrunch.com/2018/03/08/false-news-spreads-faster-than-truth-online-thanks-to-human-nature/.

41 Anirudh Sunilkumar, "Government Defines 'fake News' In Parliament; Says Social Media Being Used for Weaponisation Of Information," *Republic World*, July 26, 2018, www.republicworld.com/india-news/general-news/government-defines-fake-news-in-parliament-says-social-media-being-used-for-weaponisation-of-information.html.

42 Subimal Bhattacharjee, "Fake news, that other pandemic," *The Economic Times*, March 18, 2020, https://economictimes.indiatimes.com/blogs/et-commentary/fake-news-that-other-pandemic/.

43 Rajya Sabha, *Action taken by the Government on the Recommendations/Observations contained in the 176th Report on the Functioning of Delhi Police*, by Department-related Parliamentary Standing Committee on Home Affairs, 189th Report, December 2, 2015, http://164.100.47.5/newcommittee/reports/EnglishCommittees/Committee%20on%20Home%20Affairs/189.pdf.

44 Anandita Yadav, "Countering Hate Speech in India: Looking for answers beyond the law," *ILI Law Review*, Volume II, Winter Issue 2018, http://ili.ac.in/pdf/csi.pdf.

45 Zombie tracker, "66A," https://zombietracker.in/recommendations/66a/.

Endnotes

46 Ajita Shashidhar, "'Think beyond sex and abuse': Netflix, Amazon Prime, other OTT platforms stare at new challenge," *Business Today,* November 12, 2020, https://www.businesstoday.in/latest/trends/govt-regulation-on-netflix-amazon-prime-to-hamper-freedom-of-storytelling/story/421729.html.

47 Amrita Nayak Dutta, "Javadekar notes absence of self-regulation by OTT platforms, says looking into suggestions," *The Print*, November 16, 2020, https://theprint.in/india/governance/javadekar-notes-absence-of-self-regulation-by-ott-platforms-says-looking-into-suggestions/545448/.

48 Ashima Obhan and Bambi Bhalla, "India: OTT Platforms Brought Under Government Regulation," *Mondaq*, November 19, 2020, https://www.mondaq.com/india/broadcasting-film-tv-radio/1007300/ott-platforms-brought-under-government-regulation.

49 Leveson Inquiry, "An Inquiry into the culture, practices and ethics of the press," *Independent report*, November 29, 2012.

50 House of Lords, "Regulating in a digital world," *Select Committee on Communications*, March 9, 2019 https://publications.parliament.uk/pa/ld201719/ldselect/ldcomuni/299/299.pdf.

51 Bundesministerium der Justiz und für Verbraucherschutz, "Act improving law enforcement on social networks [Netzdurchführungsgesetz – NetzDG]," *EU*, June 28, 2017, https://ec.europa.eu/growth/tools-databases/tris/en/search/?trisaction=search.detail&year=2017#=127.

52 Creating a French framework to make social media platforms more accountable: Acting in France with a European vision, "Regulation of social networks – Facebook experiment," *Interim mission report*, May 2019, http://thecre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf.

53 Krishnanand Tripathi, "More than 17,000 websites blocked for spreading obscene content, violating Indian values," *Financial Express,* February 13, 2019, https://www.financialexpress.com/industry/technology/the-list-of-websites-blocked-by-the-government-for-spreading-obscene-content/1486743/.

54 BBC News "Twitter shuts 377,000 'terrorism' accounts," *BBC News,* March 22, 2017, https://www.bbc.com/news/technology-39351212.

55 J. Clement, "Facebook: Most users by country," *Statista.com,* July 24, 2020 https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

56 PTI, "Parliamentary panel to discuss Facebook issue on Wednesday," *Economic Times,* September 1, 2020, https://economictimes.indiatimes.com/news/politics-and-nation/parliamentary-panel-to-discuss-facebook-issue-on-wednesday/articleshow/77876060.cms.

57 "Ravi Shankar Prasad writes to Mark Zuckerberg, accuses Facebook India of bias: Full text of letter," *India Today Web Desk*, September 01, 2020, https://www.indiatoday.in/india/story/ravi-shankar-prasad-writes-to-mark-zuckerberg-accuses-facebook-of-bias-full-text-of-letter-1717521-2020-09-01.

58 Rakesh Mohan Chaturvedi, "Facebook India chief appears before parliamentary panel; Opposition, BJP MPs allege bias," *Economic Times,* September 2, 2020, https://economictimes.indiatimes.com/news/politics-and-nation/facebook-india-chief-ajit-mohan-appears-before-par-panel-debating-misuse-of-social-media-platforms/articleshow/77894164.cms.

59 ET Bureau, "Muslim staffers at Facebook call for transparency in enforcing policies," *Economic Times, August 22, 2020,* https://economictimes.indiatimes.com/tech/internet/muslim-staffers-at-facebook-call-for-transparency-in-enforcing-policies/articleshow/77686338.cms.

Endnotes

60 Jack Nicas, "How YouTube Drives People to the Internet's Darkest Corners," *The Wall Street Journal*, February 7, 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478.

61 Sarah Koslov, "Incitement and the Geopolitical Influence of Facebook Content Moderation," *Georgetown Law Technology Review*, January 2020, https://georgetownlawtechreview.org/incitement-and-the-geopolitical-influence-of-facebook-content-moderation/GLTR-01-2020/.

62 Archis Chowdhury, "Fake News in The Time Of Coronavirus: A BOOM Study," *Boom Live*, May 8, 2020, https://www.boomlive.in/fact-file/fake-news-in-the-time-of-coronavirus-a-boom-study-8008/page-2.

63 Kunal Purohit, "YouTube Hatemongers Are India's New Stars," *FP*, August 25, 2020, https://foreignpolicy.com/2020/08/25/india-youtube-stars-hatemongers-nationalism/.

64 Jeff Horwitz and Deepa Seetharaman, "Facebook Executives Shut Down Efforts to Make the Site Less Divisive," *The Wall Street Journal*, May 26, 2020, https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499.

65 Equality Labs, "Facebook India- Towards a tipping point of violence caste and religious hate speech," *Equality Labs*, 2019, https://www.equalitylabs.org/facebookindiareport.

66 David Kaye, "The Clash Over Regulating Online Speech," *Slate*, June 06, 2019, https://slate.com/technology/2019/06/social-media-companies-online-speech-america-europe-world.html.

67 Megha Mandavia, "Facebook says will remove content to mitigate adverse legal or regulatory impact," *The Economic Times*, September 2, 2020, https://economictimes.indiatimes.com/tech/internet/facebook-says-will-remove-content-to-mitigate-adverse-legal-or-regulatory-impact/articleshow/77865481.cms.

68 Shuchi Bansal, "Content regulation lapses cast doubts on Facebook's biz model," *Livemint*, August 24, 2020 https://www.livemint.com/companies/people/-content-regulation-lapses-cast-doubts-on-facebook-s-biz-model-11598232566696.html.

69 Will Ziebell, "Australia forms task force to guard elections from cyber-attacks," *Reuters*, June 9, 2018, www.reuters.com/article/us-australia-security-elections/australia-forms-task-force-to-guard-elections-from-cyber-attacks-idUSKCN1J506D.

70 Australian Association of National Advertisers, "A quick guide to the Australian "Fake News" Code," *Australian Association of National Advertisers*, July 2, 2020, https://aana.com.au/2020/07/02/a-quick-guide-to-the-australian-fake-news-code/.

71 Naaman Zhou, "Google's open letter to Australians about news code contains 'misinformation', ACCC says," *The Guardian*, August, 17, 2020, https://www.theguardian.com/technology/2020/aug/17/google-open-letter-australia-news-media-bargaining-code-free-services-risk-contains-misinformation-accc-says.

72 Damien Cave, "Australia Passes Law to Punish Social Media Companies for Violent Posts," *The New York Times*, April 3, 2019, www.nytimes.com/2019/04/03/world/australia/social-media-law.html.

73 Emmanuel Macron, "Speech by M. Emmanuel Macron, President of the Republic at the Internet Governance Forum," November 12, 2018, https://www.elysee.fr/en/emmanuel-macron/2018/11/12/speech-by-m-emmanuel-macron-president-of-the-republic-at-the-internet-governance-forum.

74 Law on the fight against information manipulation, "Anti-manipulating information

manipulation law," *Fight against information manipulation*, December 22, 2018, www.senat.fr/dossier-legislatif/ppl17-623.html.

75 Alexander Damiano Ricci, "French opposition parties are taking Macron's anti-misinformation law to court," *Poynter*, December 4, 2018, https://www.poynter.org/fact-checking/2018/french-opposition-parties-are-taking-macrons-anti-misinformation-law-to-court/.

76 E-safety commissioner, Accountability reporting, *E-safety commissioner*, https://www.esafety.gov.au/about-us/corporate-documents/accountability-reporting.

77 Australian Competition and Consumer Commission, "News media bargaining code," *Australian Competition and Consumer Commission*, July 31, 2020, www.accc.gov.au/focus-areas/digital-platforms/news-media-bargaining-code/draft-legislation.

78 Lilian Edwards and Michael Veale, "Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'?," *IEEE Security & Privacy*, 2018, https://strathprints.strath.ac.uk/63317/1/Edwards_Veale_SPM_2018_Enslaving_the_algorithm_from_a_right_to_an_explanation_to_a_right_to_better_decisions.pdf.

79 Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG), October 1, 2017, https://germanlawarchive.iuscomp.org/?p=1245.

80 Human Rights Watch, "Germany: Flawed Social Media Law," February 14, 2018, *Human Rights Watch*, May 1, 2020, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law

81 Article 19, "Germany: The Act to Improve Enforcement of the Law in Social Networks," *Article 19*, August 2017, https://www.article19.org/wp-content/uploads/2017/09/170901-Legal-Analysis-German-NetzDG-Act.pdf.

82 Heidi Tworek, "An Analysis of Germany's NetzDG Law," *Transatlantic Working Group*, April 15, 2019, https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

83 Ars Technica, "Gab, the right-wing Twitter rival, just got its app banned by Google," *Ars Technica*, October 25, 2007, https://arstechnica.com/civis/viewtopic.php?t=1395877&start=520.

84 Amnesty International, "Facebook Forced to Comply with Singapore's Censoring of Critics," *Amnesty International*, February 19, 2020. https://www.amnesty.org/en/latest/news/2020/02/singapore-social-media-abusive-fake-news-law/.

85 F Kathleen, "POFMA Is Yet Another Tool by the Singapore Government to Suppress Criticism and Dissent, Said FORUM-ASIA and CIVICUS," *The Online Citizen*, April 12, 2019, https://www.onlinecitizenasia.com/2019/04/12/pofma-is-yet-another-tool-by-the-singapore-government-to-suppress-criticism-and-dissent-said-forum-asia-and-civicus/.

86 Online Citizen Asia, "POFMA Is Yet Another Tool by the Singapore Government to Suppress Criticism and Dissent, Said FORUM-ASIA and CIVICUS," *The Online Citizen*, April 12, 2019. https://www.onlinecitizenasia.com/2019/04/12/pofma-is-yet-another-tool-by-the-singapore-government-to-suppress-criticism-and-dissent-said-forum-asia-and-civicus/.

87 Singapore Statutes Online, "Protection from Online Falsehoods and Manipulation Act 2019," *Government of Singapore*, June 3, 2019, https://sso.agc.gov.sg/Act/POFMA2019.

88 UK Government, "Online Harms White Paper." Government of United Kingdom, April 2019, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf.

Endnotes

89 BBC News, "Government Announces Anti-Fake News Unit." *BBC News*, January 23, 2018, https://www.bbc.com/news/uk-politics-42791218.

90 Brandenburg v. Ohio, 395 U.S. 444 (1969).

91 CNN, "2016 Presidential Campaign Hacking Fast Facts," *CNN*, October 31, 2019, https://edition.cnn.com/2016/12/26/us/2016-presidential-campaign-hacking-fast-facts/index.html.

92 Colin Lecher, "Senators Announce New Bill That Would Regulate Online Political Ads." *The Verge*, October 19, 2017. https://www.theverge.com/2017/10/19/16502946/facebook-twitter-russia-honest-ads-act.

93 "Schiff, Murphy and Curbelo Request DNI Assess National Security Threats of 'Deep Fakes,'" *U.S. Congressman Adam Schiff of California's 28th District*, September 13, 2018, https://schiff.house.gov/news/press-releases/schiff-murphy-and-curbelo-request-dni-assess-national-security-threats-of-deep-fakes.

94 Megha Mandavia, "Social Media to Join Hands to Fight Fake News, Hate Speech," *The Economic Times*, February 19, 2020, https://economictimes.indiatimes.com/tech/internet/social-media-to-join-hands-to-fight-fake-news-hate-speech/articleshow/74200542.cms.

95 Zachary Laub, "Hate Speech on Social Media: Global Comparisons," *Council on Foreign Relations*, June 7, 2019, https://www.cfr.org/backgrounder/hate-speech-social-media-global-comparisons.

96 Law Commission of India, "Hate Speech" *Law Commission of India*, Report No.267, March 2017, http://lawcommissionofindia.nic.in/reports/Report267.pdf.

97 Article 19, "Article 19's response to recognition of IMPRESS," *Article 19*, October 23, 2016, https://www.article19.org/resources/article-19s-response-to-recognition-of-impress/.

98 Shawn Lim, "Facebook Launches Chatbot and News Hub in India to Fight against Misinformation on Coronavirus." *The Drum*, March 30, 2020, https://www.thedrum.com/news/2020/03/30/facebook-launches-chatbot-and-news-hub-india-fight-against-misinformation.

99 Rebecca Heilweil, "Facebook Is Flagging Some Coronavirus News Posts as Spam." *Vox*, March 17, 2020, https://www.vox.com/recode/2020/3/17/21183557/coronavirus-youtube-facebook-twitter-social-media.

100 Jamie Bartlett and Alex Krasodomski-Jones, "Counter-speech on Facebook," *Demos*, September 2016, https://demosuk.wpengine.com/wp-content/uploads/2016/09/Counter-speech-on-facebook-report.pdf.

101 Anam Ajmal, "Kapil Mishra's speech on CAA example of inciting violence: Zuckerberg," *Times of India*, June 7, 2020, https://timesofindia.indiatimes.com/india/kapil-mishras-speech-on-caa-example-of-inciting-violence-zuckerberg/articleshow/76240114.cms.

102 Live Law News Network, "Hate Speech Undermines Free Market Place Of Ideas," *Live Law*, September 15, 2020, https://www.livelaw.in/top-stories/hate-speech-undermines-free-market-place-of-ideas-consideration-on-prior-restraint-is-different-in-case-of-hate-speech-gautam-bhatia-tells-sc-162978.

103 Zoe Tidman, "Coronavirus has unleashed 'tsunami of hate and xenophobia' across the world, says UN chief," *Independent*, May 8, 2020, www.independent.co.uk/news/world/coronavirus-xenophobia-hate-antonio-guterres-un-cases-deaths-a9505271.html.

Images used in this paper are from Getty Images/Busà Photography.

OBSERVER RESEARCH FOUNDATION

Ideas . Forums . Leadership . Impact

20, Rouse Avenue Institutional Area,
New Delhi - 110 002, INDIA
Ph. : +91-11-35332000. Fax : +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org