ADDRESSING MISINFORMATION ON WHATSAPP IN INDIA THROUGH INTERMEDIARY LIABILITY POLICY, PLATFORM DESIGN MODIFICATION, AND MEDIA LITERACY

Ben Medeiros and Pawan Singh

ABSTRACT

Through a case study of lynchings in India that are perceived to have been catalyzed by misinformation on WhatsApp, this article explores how policymakers can mitigate social media misinformation without compromising public discourse. We evaluate the costs and benefits of three approaches to managing misinformation: intermediary liability reform, changes to platform design, and public information endeavors addressing user attitudes and behaviors. We find that while current media literacy endeavors seem somewhat misdirected, more locally attuned initiatives might productively address the underlying susceptibility to misinformation while avoiding the free speech compromises that come with stringent liability rules and restrictions on anonymous speech.

Keywords: misinformation, intermediary liability, media literacy, India, WhatsApp

WhatsFake on WhatsApp

Optimism regarding social media’s inherent democratic and disruptive potential has given way to concern about the ease with which falsehoods are spread.1 While the popular term “fake news”—or “the online publica-

1. Toward the close of the first decade of the twenty-first century, the potential of Web 2.0—social media interactive technologies like YouTube, Twitter, Facebook, and WhatsApp—has been broadly understood in terms of its affordances of greater citizen participation with inherent democratic potential to transform politics. See, for example, Dahlberg, 855. However, such potential is not automatically realized: Brian D. Loader and Dan Mercea point to the disruptive potential of social media to shape social relations in ways that are “indeterminate and contingent upon a multitude of clashes between social agents, groups and institutions that have competing conceptions of networking democracy.” Loader and Mercea, 759.
tion of intentionally or knowingly false statements of fact”—receives much of the attention, an equally vexing problem concerns the susceptibility of Internet users to believing and circulating misinformation, or falsehoods that “arise from little intent to deceive.” Anthropologist Heidi Larson, for instance, has argued that “the deluge of misinformation . . . on social media should be recognized as a global public health threat.”

In 2018, this concern about online misinformation reached a fever pitch in India following numerous reports of mob lynchings in major Indian news media outlets. The incidents were reportedly catalyzed by rumors on the popular messaging service WhatsApp (which has 200 million active users in India) about the presence of child kidnappers in various Indian villages—individuals who, in some cases, were described to be medical professionals in search of children to harvest their kidneys. Compounding the complexity of the problem was WhatsApp’s use of an end-to-end encryption technology to protect the confidentiality of messages while also preventing their traceability.

Given the gravity of these incidents, it is only natural for politicians, policy bureaucrats, and the social media companies themselves to seek a swift response that reassures constituents (whether members of the public or financial stakeholders) that things are under control. This article focuses on the relative merits of arguably the three most significant approaches to managing misinformation on social media that have emerged in the Indian context: changes in the legal liability that platforms face for third-party speech, alterations in the technical design of platforms (whether voluntary or compelled by law), and public information endeavors (whether by government, advocacy organizations, or the platform companies themselves) to inculcate greater public media literacy.

Perhaps the majority of legal and popular commentary thus far has focused on regulation or technical modification of the platform itself as the preferred avenue of reform. Yet several Indian scholars have cautioned against a technocentric approach to addressing these incidents, and legal scholars like Jack Balkin have argued persuasively that “new-school speech regulation” (where governments incentivize platform companies to enact overbroad restrictions on the speech of their users) of the variety being proposed in India can undermine freedom of expression to an

---

2. Klein and Wueller, 6; Cummings and Kong, 188.
3. Larson.
4. ABP News.
5. IANS.
unacceptable degree. Building on these scholars’ general observations, we, therefore, argue that a more fruitful approach for WhatsApp, the Ministry of Electronics and Information Technology (IT), and civil society stakeholders might be to concentrate on the local cultural and institutional factors that enable these incidents of lynching and harassment to occur.

After reviewing the variability in intermediary liability policies across different national contexts (including its history in India), we first demonstrate the drawbacks of more stringent intermediary liability rules and compelled modifications of the platform to restrict anonymous speech. Then, after explicated the somewhat misdirected focus of WhatsApp’s own media literacy efforts, we explore some more locally attuned media literacy and community initiatives that we argue could help ameliorate the underlying susceptibility to misinformation. Ultimately, our argument favors an approach that addresses the hazards of misinformation with minimal cost to democratic freedoms and free speech and recognizes the respective constraints and responsibilities of all stakeholders (i.e., the government, platforms, and users). In addition to providing recommendations for the Indian context, we intend our analysis to also serve as a case study that can be applied to other scenarios in which policymakers must address a consequential but highly specific problem of misinformation without also compromising free speech on social media platforms generally.

**Platform Regulation**

This section first introduces the main approaches to the legal regulation of digital platforms that have been adopted in different national contexts. It then considers the possible ramifications of adopting a more stringent approach to intermediary liability to address the concern about viral misinformation in India—specifically in the form of a recent proposal for changes to Section 79 of the IT Act.⁶

Legal scholar Jack Balkin has provided a useful framework for understanding how the Internet has changed the regulation of speech. Specifically, he differentiates between the prevailing system of speech regulation in 20th century—where governments largely punished speakers directly—and “new school speech regulation,” where governments

---

6. S.79 provides immunity to intermediaries for any illegal content posted by third parties. The intermediaries are obligated to remove such content with 36 hours from the time of notification under the Act (Intermediary Guidelines) 2011. Tech2 News Staff.
deputize or incentivize **private** intermediaries to regulate the speech of their users. Applying Balkin’s distinction between legally mandated and voluntary initiatives, we can plot social media regulations across national contexts on a continuum according to the degree to which information intermediaries are made legally liable for the content that third parties post. The nomenclature used to evaluate what role platforms do or do not play in the publication of speech varies across national contexts (e.g., platforms are evaluated in terms of whether they become “publishers” in the United States, and they are further differentiated as “conduits” or “hosts” in the European Union [EU]), but the core conceptual question is similar: when does the platform itself incur liability for the speech of those who use it, even if the platform is not the originator of the content?

In response to the perception that platforms are not doing enough to combat particular kinds of problematic content, a wave of more stringent intermediary liability laws has gained traction across the globe in recent years. The recent German legislation colloquially referred to as NetzDG, for instance, imposes steep fines on social media platforms that are alerted to the presence of “manifestly unlawful” content (a category that includes “violation of intimate privacy by taking photographs” and defamation, as defined within the German criminal code). Further, the European Parliament has recently approved additional legislation requiring any website accessible in the EU to “remove any content deemed ‘terrorist’ content” by any vaguely defined “competent authority” within one hour of being notified.” Other countries have applied similar approaches to tackling the issue of “fake news” specifically. Singapore, for instance, has passed legislation that “will require online sites to show corrections to false or misleading claims and take down falsehoods” while creating criminal and civil penalties for “those who spread an online falsehood with intent to prejudice the public interest, and those who make a bot to spread an online falsehood.” For Indian legal scholar Prashant Reddy, such developments represent “a sign that governments are increasingly exasperated by a

---

8. See the description of “interactive computer services” in 47 U.S.C §230, or the designation of different types of intermediaries in European Parliament Directive 2000/31/EC.
9. Center for Democracy and Technology.
10. Masnick, “EU Parliament Votes To Require Internet Sites.” It should also be noted, however, that the provisions of this regulation are not yet in force. Trilogue negotiations commenced in the fall of 2019, which would be followed by second reading. See Theron.
certain class of intermediaries, especially social media companies that are performing a poor job of moderating content.”

Even in the United States, where the legal regime is comparatively deferential to technology companies, some scholars and policymakers have also endorsed reforms to create more accountability for harmful speech. In their recent book *The Misinformation Age: How False Beliefs Spread*, for instance, social scientists Cailin O’Connor and James Weatherall call for a “new regulatory framework to penalize the intentional creation and distribution of fake news, similar to laws recently adopted in Germany to control hate speech on social media.” Legal scholars Benjamin Wittes and Danielle Keats Citron have proposed modifications to Section 230 of the Communications Decency Act to make absolution from liability for third-party content contingent on a demonstration that platforms make a good faith effort to remove inappropriate content when notified by a user that it fits particular categorical criteria. As they put it, “platforms would enjoy immunity from liability if they could show that their response to unlawful uses of their services in general was reasonable.” The recent SESTA/FOSTA legislation manifests a similar goal, modifying Section 230 to create potential criminal liability for platforms that knowingly facilitate activity that violates sex trafficking laws.

In the Indian context, the core intermediary liability regulation found in Section 79 of the IT Act has undergone several noteworthy revisions and reinterpretations since its passage in 2000. The original statute created a safe harbor from liability for third-party content absent “actual knowledge” of its illegality, but this was modified in 2008 amendments to the Act implementing a notice-and-takedown system that conditioned immunity on removal of illegal content within 36 hours following receipt of a complaint. The 2015 *Shreya Singhal* case, however, was celebrated by free speech advocates for the ways in which it pared down and clarified the

15. SESTA stands for ‘Stop Enabling Sex Traffickers Act’ and FOSTA stands for ‘Fight Online Sex Trafficking Act’.
16. The amendment is considered a reaction to a case pertaining to the criminal liability of an ecommerce platform, baazee.com, on which a student from the Indian Institute of Technology, Kharagpur, posted a listing containing a pornographic MMS clip of teenage students of the Delhi Public School. The Delhi High Court rejected the petition for annulling the criminal prosecution of the managing director of the platform, who in the court’s view, failed to exercise due diligence in filtering pornographic content. See Advani, 121.
17. Sinha, location 2051.
application of the “actual knowledge” standard in Section 79. Specifically, the ruling mitigated an intermediary’s responsibility for determining the illegality of content on its own by decreeing that “for an intermediary to have knowledge, there must be a judicial order for removal of content.”

In December 2018, a proposal was introduced in the Indian Parliament to amend the statutory regulations of information intermediaries found in Section 79 of the IT Act. The proposal suggested sweeping modifications to both the responsibilities of intermediaries to actively monitor and remove content as well as to regularly remind users of these legal requirements, while also attempting to make the communications that occur on their platforms “traceable.” More specifically, it would require intermediaries to “proactively identif[y] and remov[e] or disabl[e] public access to unlawful information or content,” which suggests that intermediaries would assume the responsibility of screening the speech of their users in order to make such determinations (rather than simply waiting for notification of, say, a court order to remove speech). The draft additionally specifies that this “proactive” responsibility should be carried out by “deploy[ing] technology-based automated tools or appropriate mechanisms.” According to sources who spoke with one journalist, the impetus for the rule change was that “the government [was] keen to be seen to be acting before the general elections on the proliferation of social media and its connection to mob violence seen in the recent past.”

Some have cheered the proposal as a necessary adaptation to the current social media landscape. In his detailed analysis of the proposal, Prashant Reddy has characterized it as a positive development toward a policy regime in India that treats “safe harbor from legal liability [as] an extraordinary subsidy and not a right,” and more generally recognizes that “it is in public interest for the state to intervene and regulate the manner in which these platforms moderate content.” In his view, requiring proactive filtering on the part of intermediaries like WhatsApp would ameliorate the way in which Shreya Singhal “transferred the cost of regulation from

---

18. Arun, for instance, states that the problem with the notice and takedown system was “mitigated by Shreya Singhal.” Arun, “Gatekeeper Liability,” 83.
19. Sinha, location 2683.
20. The Information Technology [Intermediaries Guidelines (Amendment) Rules] 2018, Section 3(9). https://drive.google.com/file/d/1ong2rw-WeYzS-9eMNCPoR3yn9rsfIXBb/view
21. Gupta, Section 3(9).
22. Chishti.
24. Ibid., 38.
intermediaries . . . to the individual Indian citizen who would now have to secure a judicial order, and spend significant money and time in doing so, before getting content taken down.”\(^{25}\) He further dismisses claims that the new rules would cause “over-censorship by private intermediaries [as] over-hyped” given the fact that they already perform content moderation selectively,\(^{26}\) and argues that the demand for traceability in the proposal is justified because platforms like WhatsApp can still provide encrypted services if they deem it integral to their value to the user—they will simply lose safe harbor from liability in the process.\(^{27}\)

Despite Reddy and other scholars’ sanguine outlook on more stringent intermediary liability standards, however, they do not adequately address the countervailing concern that such reforms incentivize intermediaries to overzealously moderate under threat of liability. As Balkin cautions, such an approach “raises problems of collateral censorship and digital prior restraint,” because “[an infrastructure provider] will tend to over-block and over-censor to avoid liability . . . because it is not [their] speech that is at stake, but that of [a] stranger.”\(^{28}\) Mandates that an intermediary restrict particular kinds of speech content (e.g., “hate speech”) are always somewhat imprecise and subjective, and a platform faced with the task of monitoring a huge volume of content will (so the logic goes) either be tempted to create overbroad rules that inevitably filter out speech that would not otherwise be punishable or simply prove unable to vet all of the relevant content. For instance, the predicted chilling effects of SESTA-FOSTA have been felt already in the platform Tumblr’s decision to ban adult content starting in December 2018. As one journalist explained, the logic was likely that “it’s just a lot cleaner and easier for them to remove adult content to make sure they get rid of any sex work related advertising.”\(^{29}\) Likewise, Chinmayi Arun offers a similar argument about overbroad moderation in the Indian context, and cites a 2011 study in which deliberately flawed takedown notices sent to intermediaries were nonetheless successful in prompting the intermediary to remove the content.\(^{30}\)

\(^{25}\) Ibid., 50.

\(^{26}\) Ibid., 53.

\(^{27}\) Ibid., 56.

\(^{28}\) Balkin, 1207; 1176.

\(^{29}\) Cyboid.

\(^{30}\) As she writes, “Private censorship tends to be invisible and over-broad since the intermediary is incentivised to avoid expensive litigation, whether or not the speech is lawful.” Arun, “Gatekeeper Liability,” 85. See also Dara.
The looming possibility of liability for content not removed expeditiously also runs the risk of leading platforms either to be overly aggressive in their own moderation efforts or to create procedures that naturally skew in favor of complainants. For example, legal scholar Jeffrey Cobia has outlined how “priority is given to the copyright over the free speech” in the actual operation of the safe harbor provisions found in Section 512 of the American Digital Millennium Copyright Act (DMCA).\textsuperscript{31} When the purported copyright holder sends a complaint to a host of third-party content like YouTube, the provider must immediately remove the material and wait to receive a counterclaim by the third-party poster. In most cases, users do not file counterclaims, often because they are unaware of the opportunity.\textsuperscript{32} Even in the event of a counterclaim, the DMCA mandates that the material cannot be reposted for 10–14 days in order to allow the copyright holder to weigh whether to file suit—at which point “the damage might already be done,” as the gap “might cause the issue to pass out of the public spectrum before the criticism is reinstated.”

The consequences can thus be significant for free speech. Stanford law professor Mark Lemley has offered the following explanation of the impetus for granting immunity from liability for third-party content (with the exception of some crimes and intellectual property violations) that is found in Section 230 of the Communications Decency Act in the United States: “The reasoning behind these immunities is impeccable: if internet intermediaries were liable every time someone posted problematic content on the internet, the resulting threat of liability and effort at rights clearance would debilitate the Internet.”\textsuperscript{33} Balkin, correspondingly, ultimately suggests a largely voluntary approach in which “large international infrastructure owners and social media platforms to change their self-conception . . . to understand themselves as a new kind of media company to protect the global public good of a free Internet.”\textsuperscript{34}

Likewise, legally compelled use of artificial intelligence (AI) in order to handle the volume of material that platforms host triggers additional concerns. In its condemnation of the proposal, for instance, an Indian advocacy organization called the Internet Freedom Foundation (IFF) astutely references the conclusions drawn by United Nations (UN) Special Rapporteur on Freedom of Expression David Kaye in his review of the

\begin{itemize}
  \item Cobia, 400.
  \item Ibid., 395–96.
  \item Lemley.
  \item Balkin, 1209.
\end{itemize}
compatibility of AI systems with the freedom of expression provisions found in the Universal Declaration of Human Rights. As Kaye sees it, problems specifically arise when AI systems must engage with subtextual elements of human language: “Unlike humans, algorithms are today not capable of evaluating cultural context, detecting irony or conducting the critical analysis necessary to accurately identify, for example, ‘extremist’ content or hate speech.” As a result, they “are thus more likely to default to content blocking and restriction, undermining the rights of individual users to be heard as well as their right to access information without restriction or censorship.”35

Digital communication scholar Sarah Myers West additionally observes that the use of AI methods is likely necessary in order to comply with “increasing expectations by government regulators that companies remove illegal content, including hate speech and violent extremist content, within predetermined time periods.”36 Yet such automated tools are often not particularly reliable and can render the moderation process more opaque. Her research on user perceptions of content policies, for instance, highlights how “[t]he perceived absence of a real person on the other side of the computer screen [is] a particular source of frustration for many users,” and thus “use of automation and limited opportunity for human interaction in content moderation systems likely served to increase users’ frustration with the process” overall.37 And while algorithmic filtering has arguably succeeded in cases like YouTube’s screening for copyrighted material, these approaches often misrecognize content as well: Tumblr’s attempts to root out pornography with automated image screening software, for instance, resulted in comical mistakes like flagging an image of a man’s chest but not flagging the same image when a small picture of an owl was added.38

While Reddy concedes that concerns about the accuracy of AI filtering are “fair,”39 he nonetheless glosses over how integral it would be to the new system outlined in the reform proposal and correspondingly, just how problematic that could be.

Due to both the general responsibilities they stipulate and the manner in which their implementation is mandated, the proposed changes to the IT Act thus run the risk of affecting a much broader range of speech than, say, the

---

36. West, 4368.
37. Ibid., 4377.
38. Cyboid.
speech which would legally qualify as incitement to violence that ostensibly lies at the heart of the lynching incidents seen in recent years. Consequently, they could well do enough collateral harm to freedom of expression that their potential to reduce the spread of misinformation is outweighed.

_Design Modifications_

In order to attempt to implement the monitoring schemes discussed earlier, WhatsApp would likely need to suspend the end-to-end encryption of messages that represents a major feature of the platform. In the initial attempt to address concern about the lynching incidents, Information and Technology Minister Ravi Shankar Prasad asked WhatsApp to “devise a process by which messages could be traced back to their origin”—though WhatsApp declined, citing user privacy.⁴⁰ Correspondingly, the IT Act amendments contain a requirement that platforms decrypt communications by end users. Specifically, this provision is couched in a mandate that communications be “traceable” so that intermediaries can comply with government investigations: “The intermediary shall enable tracing out of such originator of information on its platform as may be required by government agencies who are legally authorized.”⁴¹ Such a modification would indeed allow the company to engage in much more robust policing of the speech that is circulated on the platform, as the company in fact currently does not possess the ability to read the communications exchanged between end users. In his analysis of the proposal, Reddy defends the traceability requirement both on the grounds that WhatsApp and other platforms can technically offer encryption still (as cited earlier), and because “traceability has been given in most modes of mass communication[,] so it is understandable for the government to request encrypted platforms like WhatsApp to make their content traceable.”⁴²

Yet the call for traceability also ignores the legitimate role of encryption in modern computing and (perhaps more importantly) overlooks the importance of anonymous speech in a democratic society. As the IFF noted, such changes would come at a cost, as encryption is used “to prevent identity theft [and] code injection attacks,” which are perhaps even “more important as more of life now involves our personal data.”⁴³ Further,

---

⁴⁰. Gill.
⁴¹. Gupta, Section 3(5).
⁴². Reddy, 57.
⁴³. Gupta, “India Must Resist.”
retaining unencrypted personal data and communications on WhatsApp could open the door for greater government surveillance of dissidents, or a situation that the IFF describes as “eerily reminiscent of China’s blocking and breaking of user encryption to surveil its citizens.” The Electronic Frontier Foundation (EFF), an American digital civil liberties organization, has noted how anonymity enables “whistleblowers [to] report news that companies and governments would prefer to suppress” and “human rights workers [to] struggle against repressive governments”—a function that potentially would be compromised if the government could easily learn the identities of WhatsApp users. Such modifications would also perhaps undermine the Indian Supreme Court judgment of 2017 upholding privacy to be a fundamental right of Indian citizens. Overall, then, the alteration of platform design to enable greater scrutiny of user speech in the interest of mitigating the spread of misinformation could potentially lead to chilling effects of its own, as users of WhatsApp might become understandably cautious about using the platform for otherwise legitimate government criticism or discussion of controversial topics.

In response to pressure from the government, WhatsApp (and its parent company Facebook) have also voluntarily explored how some other modifications of the platform’s design and mechanics might help to stem the viral spread of misinformation. The changes it has implemented regarding the forwarding of messages have received perhaps the most attention. In July 2018, the company announced a new limit on the number of times that an individual user can forward a particular message and eliminated the function of “quick forwarding”—a change that was perceived to be particularly urgent in India, where according to a Facebook source who spoke with the Guardian newspaper, “people forward more messages, photos, and videos than any other country in the world.” Yet according to Amber Sinha, “[t]hese strategies have had little impact, as political campaigns have found ways to circumvent the limitations posed by them.”

The platform has also cultivated methods of using message and user metadata to determine when an account is engaging in spam-like forwarding behavior even without seeing the content of the messages themselves.

44. Ibid.
45. EFF.
46. See, Justice K.S. Puttaswamy (Retd.) and Anr. Vs Union of India and Ors. WRIT PETITION (CIVIL) NO. 494 OF 2012.
47. Sinha, location 2375.
48. Hern.
49. Sinha, location 2386.
For instance, if an account has only been used to send group messages (rather than messages to individuals), then WhatsApp can surmise that it is possibly engaging in prohibited mass-messaging and remove the account. A different method of detecting likely automated messaging behavior involves the presence of a “typing indicator” (the bubble containing an ellipsis that appears when one’s interlocutor is typing in most chat applications). According to Matt Jones, an engineer at WhatsApp, if the code that a spammer has used to automate messaging has never sent a typing indicator before a message, then the company will ban the account.50 Such design-modification tactics thus arguably represent a more targeted response to the misinformation problems that have triggered concern in recent years because they concentrate on the specific features of virality (such as automation and volume of messages) that social media platforms can be said to exacerbate, therefore, yet they may also strike policymakers as insufficient in that they do not address the actual content being shared and (assuming Sinha is correct) seem relatively easy to circumvent.

Further, fixating on technological solutions risks overlooking the societal factors that enable the lynching incidents to occur in the first place. Several Indian scholars have offered nuanced analyses of the situation that push us to think beyond technocentrism. Feeza Vasudeva and Nicholas Barkdull argue, for instance, that the government’s technocentric response thus far is attributable to the fact that the “least-cost solution is to blame technology and other factors rather than attempt to heal age-old social divides and address the discursive constructions being spread through the messages themselves.”51 For Chinmayi Arun, technology plays only a minor role, as she reframes the WhatsApp-fake news lynching nexus in terms of the tacit approval of local actors (including law enforcement officers and leaders) that sets the stage for communal violence. Maya Mirchandani of the Observer Research Foundation similarly contextualizes the public receptiveness to misinformation in relation to the narrative of majority persecution that has taken hold in India’s political and cultural imaginary.52 Finally, Rahul Mukherjee characterizes mob lynchings as resulting from a slow build-up of the everyday habit of forwarding communal (fake) messages via WhatsApp—an argument that points in the direction of inculcating awareness and understanding of WhatsApp viral content.53 We must,
therefore, consider longer-term endeavors to cultivate media literacy and more discerning user engagement with the information being shared over the platform as a more comprehensive solution to the kinds of viral incidents that have prompted concern in India.

**Media Literacy**

Media literacy has been proposed as a primary means of combatting misinformation online by a variety of commentators. As a 2017 Yale Information Society Project workshop report suggested, “[c]ontent consumers must be better educated, so that they are better able to distinguish credible sources and stories from their counterparts.”54 This could be achieved by “[c] onsumers [becoming] educated about how news information propagates in today’s world, the harms of fake news, and how to identify it.” Likewise, communication scholars Christopher Cummings and Wei Yi Kong argue that “[t]he ability to critically analyze and evaluate information quality is important in discerning fake news and in turn remove [sic] motivation to inadvertently share misinformation.”55 Pratik Sinha of Altnews (a fact-checking website in India) has argued more specifically that those in the remotest areas of India have less experience with the conventions of social media discourse and are thus especially susceptible to misinformation.56 Indeed, from a general policy perspective, emphasizing the development of such skills in users is sensible in that it addresses an underlying societal problem of the “digital divide” without requiring the kinds of problematic legally mandated moderation or liability schemes discussed earlier.

In fact, a renewed emphasis on media literacy might be especially necessary in India in order to prevent overzealous regulation of speech platforms by the government. As legal scholar Siddharth Narrain has pointed out, in the 2015 *Shreya Singhal* case, “the government attempted to justify the retention of a law meant to curb free expression online [Section 79] on the basis that there was something different about the medium, specifically pointing to the ease of use and the fact that one did not require to be literate or have specialized knowledge to use [the internet].”57 If lack of literacy

54. Yale Information Society Project, 11.
55. Cummings and Kong, 201.
57. Narrain, 394. While Section 79 was upheld, it is worth noting that Section 66A, which proscribed “grossly offensive” and “menacing” content, was struck down. See Narrain, 393.
has been rhetorically invoked as a pretext for regulation, then presumably increasing literacy would mitigate the perception that even more stringent liability rules are required for online intermediaries.

On the other hand, other scholars have cautioned against putting too much stock in simplistic conceptions of media literacy. When assessing possible solutions in their recent study *Network Propaganda*, Yochai Benkler, Robert Faris, and Hal Roberts state unequivocally that media literacy “is not a panacea and will not by itself disarm the incredibly resilient psychological and social-identity-based factors that so often lead us astray.”\(^58\) Melissa Zimdars, a communication scholar and the creator of the document called “False, Misleading, Clickbait-y, and/or Satirical ‘News’ Sources” that went viral following the 2016 US presidential election, similarly concluded in a recent keynote address that “media literacy is necessary . . . but insufficient for addressing the problem of fake news.”\(^59\) Media scholar danah boyd has more specifically argued that proposals to combat fake news focused on media literacy are “likely to fail” because they ignore how the core tenets of media literacy—such as evaluating the credibility of sources and the financial motivations at play, the general mandate to be skeptical of authority, and the elevation of expertise—do not necessarily lead to uniform conclusions about what information to trust. For instance, she notes how the dictum to “follow the money” in analyzing the credibility or motivations of sources can easily lead one to facile antisemitic ideas about the unreliability of the media because of the prevalence of Jewish-Americans in its upper echelons.\(^60\)

A 2019 study by Indian researchers Shakuntala Banaji and Ram Bhat echoes boyd’s problematizing of media literacy as a one-size-fits-all solution. Specifically, their study offers a distinction between functional and critical media literacy. Where functional media literacy refers to an individual’s skills and capacities to use various media and their technical properties, affordances of information and communication technologies such as apps, critical media literacy emphasizes the intersection of skills and capacities with “understandings of ideology, political economy and other forms of power as well as an active audience that struggles to make meaning of a text.”\(^61\) Applying this distinction, the researchers found that users shared mis/disinformation or acted to suppress its identification out of diverse

\(^{58}\) Benkler, Faris, and Roberts, 378.

\(^{59}\) Zimdars.

\(^{60}\) boyd.

\(^{61}\) Banaji and Bhat, 27.
motivations, including preexisting prejudice against specific communities, belief in a particular nationalist/religious ideology, and preexisting loyalty to political parties and ministers. Given these findings, they argue that: “functional literacy is of little help in preventing the spread of misinformation since even technically savvy users are willing to share disinformation and misinformation as long as it aligns with their values, beliefs, and ideological convictions.”

In their conclusion, correspondingly, Banaji and Bhat echo other critics that WhatsApp’s modification of its technical features in response to the pressure by the Indian government is not adequate given the possibility of circumventing the restrictions on the number of forwarded messages or members in a group using an outdated version of the messaging service such as WhatsApp GB and WhatsApp Plus. The way ahead to curb the spread of misinformation to prevent potential violence, therefore, appears to be a combination of modifying technical features and inculcating more critical media literacy instincts.

With this wide range of analytical perspectives on media literacy and misinformation in mind, we can turn to the critique of existing initiatives in the Indian context. In response to the Indian government’s calls for WhatsApp to address misinformation on its platform, WhatsApp has responded by issuing Public Service Announcements (PSAs) on mainstream television channels as well as through radio and print newspapers. National and international news channels and print media have also covered the debate, providing critical perspectives on the issue. Through a close reading of television and print PSAs and expert commentaries aired on national news networks, we can unpack the assumptions underlying broader public measures to educate users and evaluate their likely impact.

In a series of one-minute ads, WhatsApp conveys its central message through the brief story of a central character, typically a middle-class person who is part of a WhatsApp family or friends group. In the first of three ads analyzed here, we are introduced to a middle-class, Punjabi girl, Kavya, who has moved away from her family to live elsewhere and shares the events in her everyday life through a family WhatsApp group. The voiceover describes her as a happy person who shares jokes with her family but in one instance, when her uncle shares a “forwarded” message with her, the voiceover says, “today she is serious.” Then Kavya calls her uncle up and asks him if he had any proof or if he just randomly shared the message.

---

63. Scroll Staff.
Then she explains to him that fake news can cause violence and convinces him to leave the group that circulated the rumor. The next scene is that of a birthday celebration in the family that befits the tagline, *Be like Kavya: Share joy, not rumors*. The brief glimpse of the rumor in the WhatsApp message reads “Dangerous Times Ahead.”

In the second ad, we meet Rajat who creates a classroom WhatsApp group in college. Rajat is described as popular and helpful as he shares information about class schedules, notes, rehearsal updates, and surprise exams on the group. He is described as the admin of the group but he never acts like one except “today!” when he receives a message from a college mate that reads “Shocking! Spread the News!” (with emojis). Rajat clarifies the message with the sender and explains that such messages could create tensions and turn people against each other. He warns him to not do this again or else he’d be removed from the group. The ad ends with a celebration during the college festival with the tagline, *Be like Rajat: Share joy, not rumors*.

The third ad shows a young woman, Geeta who is a cooking enthusiast and shares recipes with her friends and colleagues through a WhatsApp group. She receives compliments from everyone. Then one day, she receives a forwarded message that asks her to spread a rumor. When she tells her mother, she asks her to forward the message but Geeta simply responds that such messages can be really dangerous. She urges the audience to delete such messages and blocks the sender. The same tagline follows.

The three-ad series has been airing on mainstream TV channels in India. In trying to accomplish the goal of getting WhatsApp users to not share rumors, the campaign’s protagonists are primarily urban individuals who are portrayed as rational, media-savvy, and generally aware. They take on a pedagogical role in relation to their peers or family members who are portrayed as gullible users of technology who forward messages indiscriminately. There are three central points of critique that demonstrate how WhatsApp media awareness ad campaigns fall short of addressing the issue in a consequential manner:

1. **Context:** The context of the ad storyline is typically urban settings, a setting that precludes the predominantly rural setting of most mob lynchings provoked by rumors around child kidnappers. It is likely that the ads may be viewed by those living in remoter parts of the country but we bracket that question for further research. The context in all three ads is set up along the lines of harmonious and jovial exchange of WhatsApp messages through group dynamics. These are then interrupted by a vague message from an unknown source that leads to the
pedagogical staging of the receiver explicating the dangers of such indiscriminate forwarding to the known sender. The situation is quickly resolved by a celebration in the climax as the tagline appears.

2. Content: When the suspicious message is received by the protagonist, the content of the message remains nonspecific—dangerous times ahead in the first instance, shocking: spread the news in the second, and an un-downloaded image in the third. While references to rumor and fake news are made in all three stories, what the actual content of the disinformation that was circulated is not made explicit. This is in stark contrast to the incidents of mob lynchings where the rumor pertains specifically to stranger-danger style alerting of a threatening presence in the local context. The content of the brief pedagogical lesson does not delve into any specific reason other than the warning that spreading such rumors/fake news can lead to violence, tensions among groups, or an undesirable outcome. The overall message of the ad campaign reiterates the linkage between rumor and adverse consequence while underplaying the role of technology that is primarily meant for the sharing of joy.

3. Concern: Given the primarily urban context of the ad campaigns and the lack of specificity in content that only refers to fake news and rumors without specific examples, the performance of concern over the adverse consequences of spreading misinformation does not adequately address the gravity of actual incidents that inspired WhatsApp to release such ads in public interest. The concern about fake news references no particular incident nor does the pedagogical performance illuminate any specific behavior modification save for a request to not engage in spreading fake news. The nationwide concern generated by WhatsApp rumors linked to mob lynchings is reduced to a friendly reprimand in the overall message of the ad.

While the ad campaign is certainly a commendable effort in the right direction, it barely skims the surface of the complexity that possibly drives mob behavior and does not account for local sensitivities at play in each individual instance of mob lynchings. The campaign broadly reinforces the link between rumors circulating on WhatsApp and untoward incidents while the actual problematic content is never specifically identified in the narrative. The ads’ focus on simple behavior modification dichotomizes users as either media literate, savvy, and aware of their civil duties or gullible users who indiscriminately forward messages without regard for their veracity.
In light of the research cited earlier, the following examples represent just a few ways in which new public service campaigns and media literacy initiatives undertaken by WhatsApp could be modeled on existing or past endeavors that have been more rooted in local information contexts. In the most basic sense, public service campaigns aimed at increasing media literacy must be tailored to addressing the grounding of viral rumors in specific local myths and tensions rather than the current campaigns’ vague warnings about “danger ahead.” First, they need to identify the key gaps in the cycle of news dissemination, consumption, and circulation among WhatsApp users in certain regional contexts. These contexts may be more prone to incidents of violence owing to a historical prevalence of cultural differences among communities combined with a lax law and order situation. For instance, they could be formulated according to ongoing empirical research that seeks to discover different media consumption attitudes and behaviors across India. The Delhi-based Digital Empowerment Foundation’s (DEF), for instance, conducted awareness workshops across 11 states in 2019 titled “Fighting Fake New” in Tier II and Tier III cities. Based on the results of a survey given to around 4,000 respondents, the DEF recommends that more organizations need to work with WhatsApp, the police, local administrations, teachers, and students to educate users via awareness campaigns with a focus on news verification technique.64

While WhatsApp’s public service ads may be insufficiently targeted and specific, the general way in which they position group leaders as key social influencers is potentially auspicious. There is precedent suggesting that a focus on deploying trusted community members to act as interactive information arbiters on the platform itself is a viable strategy for mitigating misinformation and counteracting inflammatory speech. Doing so can also conspicuously demonstrate the engagement of law enforcement, potentially minimizing the impression of its tacit approval of violence. In Pune in 2014, for instance, local officials helped to defuse threats of communal violence by “creating groups [on WhatsApp] of police stations, housing societies, social workers, and politicians, [which] helped them

---

64. Some of the major findings of the survey were: that 45 percent of the respondents did not believe in the information circulated via WhatsApp, while 24 percent were skeptical of it; only 9 percent reported to believing the information sent via WhatsApp forwards; 49 percent of the respondents understood what the “forward” label meant on a message, while only 22 percent were aware of the meaning of encryption. Finally, 60 percent of the respondents believed that WhatsApp did serve a positive function. “Fighting Fake News.”
have a substantial presence on WhatsApp communities to track images and utterances so as to respond quickly.”

Another potentially productive approach in this vein involves giving greater attention to the role that local citizen journalism can play in mitigating misinformation—specifically through the “WhatsApp broadcast” function that enables message transmission via user lists. An example of this comes from a small district called Pilibhit in Uttar Pradesh where one local reporter, Shivendra Gaur, launched a subscription-based WhatsApp news broadcast service called Rocket Post Live in 2016. With about 11, 400 subscribers as of 2017 who pay Rs. 100 (US$ 1.50), Rocket Post sends news developments to subscribers throughout the day with a Hindi news bulletin of three to five minutes around 8 p.m. everyday. Effectively circumventing the spread of misinformation through group messaging on WhatsApp, Gaur sends his subscribers a code to activate the news bulletin along with alerts through a request to be added to the Rocket Post Live network. Gaur works with five other reporters and has two part-time staffers who canvas for news subscribers for a 30 percent commission. The service is free for students.

Such broadcast features in the app can be further leveraged to debunk dubious information circulating within particular regional contexts. In contrast to the high production value PSAs aired on Indian news channels, WhatsApp, in partnership with policymakers, police, and local administration could enlist and incentivize enterprising local journalists to engage in educating subscribers via WhatsApp broadcast list. The role of the citizen journalist as a trusted agent can facilitate the verification of forwarded messages, familiarization with credible news sources, and classification of messages based on dubious or potentially provocative content.

Conclusion

Given the flaws detailed earlier regarding proposed changes to intermediary liability rules, compromising encryption, and current efforts to increase media literacy, we ultimately suggest that policymakers in India eschew sweeping changes to the operation of the platform itself in favor of locally tailored information literacy campaigns. First, in their eagerness to do something about situations that are admittedly dire, lawmakers must

65. Narrain, 397.
66. Sharma.
be sure not to overlook the negative consequences of incentivizing overzealous moderation practices and compromising dissident speech when they compel changes to platform architecture (specifically the removal of end-to-end encryption) and subsequently impose liability unless broad categories of putatively problematic content are removed proactively. These consequences have been documented in other national contexts when similar liability schemes have been advanced, and prominent scholars in law, communication studies, and international organizations have warned of the likely effects when AI screening is necessitated to comply with increased liability for third-party content.

Second, given that misinformation will likely continue circulating via WhatsApp, the platform may find it worth investing in locally grounded media literacy initiatives in India to explore grounds-up programs that involve diverse user groups. Taking a cue from existing endeavors that have been documented in previous studies and news reports, these could include tailoring messages to engage with particular regional tensions and demographics, as well as mobilizing trusted community voices and law enforcement to utilize WhatsApp’s group and broadcast functions in order to directly monitor and combat misinformation. Overall, focusing on these kinds of responses to the problem of misinformation on WhatsApp minimizes the threats to free speech associated with increased platform liability and decryption while taking into account the micropeculiarities of context to implement an educational agenda among the users, who remain the most affected of all by viral misinformation.

BIBLIOGRAPHY


