



Jus Corpus Law Journal

Open Access Law Journal – Copyright © 2026 – ISSN 2582-7820

Editor-in-Chief – Prof. (Dr.) Rhishikesh Dave; Publisher – Ayush Pandey

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike 4.0 International (CC-BY-NC-SA 4.0) License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium provided the original work is properly cited.

Digital Lynch Mobs & Platform Impunity: The Battle for Free Speech in India

Balpreet Kaur Bhatti^a

^aKES' Shri Jayantilal H. Patel Law College, Mumbai, India

Received 28 July 2026; *Accepted* 29 August 2026; *Published* 03 September 2026

This article examines the widening fault line between free expression and content moderation through the lens of Indian law and lived experience. It traces the architecture of platform liability from America's Section 230 to India's Section 79 safe harbour and the 2021–2026 Intermediary Guidelines and asks a harder question underneath the statutes: who is actually accountable when a rumour becomes a riot, or when a government quietly outsources censorship to a company's compliance department? Using the state action doctrine as a diagnostic tool, and grounding the analysis in real events — the WhatsApp lynchings of 2017–2019, the year-long information blackout in Manipur, the Facebook–BJP hate speech controversy, the farmers' protest takedowns, and the landmark judgments in Shreya Singhal, Anuradha Bhasin, Kunal Kamra, and X Corp v Union of India (2025). This article argues that India's regulatory drift toward executive-controlled takedown portals is quietly eroding the judicial safeguards the Supreme Court built in 2015. It closes with a proposed framework for distinguishing speech that democracies must protect from speech that gets people killed, and for putting a human being, not an algorithm or a bureaucrat, back at the centre of that decision.

Keywords: *free speech, intermediary liability, state action doctrine, hate speech.*

INTRODUCTION

At no point in human history has so much speech been generated, amplified, and mediated by a handful of corporate-owned algorithms. For India's over 800 million smartphone users, platforms like WhatsApp, YouTube, X, and Meta have transformed into the public square. This digital transformation has unlocked unprecedented democratic participation, giving voice to marginalised communities, enabling grassroots mobilisation, and decentralising information access. However, this vast digital architecture operates without a neutral safety net. When a moderation failure occurs in Silicon Valley, it is typically measured in engagement metrics or ad revenue; when it occurs in rural India, it can manifest as a body in a police mortuary.

This article examines the widening fault line between free expression and content moderation through the lens of Indian law and lived experience. It traces the architecture of platform liability from America's Section 230 to India's Section 79 safe harbour, exploring the steady evolution of the 2021–2026 Intermediary Guidelines. Beneath the statutory texts lies a harder, more urgent question: who is actually accountable when a viral rumour becomes a fatal riot, or when a government quietly outsources censorship to a company's compliance department under the guise of due diligence?

To diagnose this shift, the analysis deploys the state action doctrine alongside real-world events. It confronts the tragic reality of the WhatsApp lynchings of 2017–2019, where end-to-end encryption rendered viral child-kidnapping rumours structurally invisible to platform moderation systems until lives were lost. It examines the year-long information blackout in Manipur, the longest state-wide internet suspension in modern Indian history outside Jammu and Kashmir, where preemptive communication silences obscured severe human rights violations under the pretext of maintaining public order. It reviews the high-stakes controversies surrounding political hate speech, content takedowns during farmers' protests, and landmark Supreme Court and High Court jurisprudence including *Shreya Singhal*, *Anuradha Bhasin*, *Kunal Kamra*, and the 2025 *X Corp v Union of India* ruling.

The central thesis of this article is that India's regulatory drift toward executive-controlled takedown portals, exemplified by the Sahyog portal and frequent emergency rule amendments, is quietly eroding the robust judicial safeguards established by the Supreme

Court in its landmark 2015 rulings. By conditioning legal safe harbour on hyper-compressed compliance windows and opaque administrative takedown mechanisms, the state has created an architecture of collateral censorship. In this system, private intermediaries, anxious to protect their business models, naturally err toward over-removal, bypassing the constitutional checks and balances that would otherwise govern direct state suppression.

Ultimately, the debate over content moderation is not a binary contest between ‘free speech’ and ‘safety’ as abstract principles. It is a structural struggle over power, transparency, and due process: who gets to decide what can be said, on what evidence, with what oversight, and how quickly an erroneous decision can be corrected. This article closes with a proposed normative framework designed to distinguish speech democracies must protect from speech that inflicts catastrophic harm, placing human judgment, rigorous procedural oversight, and constitutional accountability back at the centre of digital governance.¹

A VILLAGE, A RUMOUR, AND A VERDICT SIX YEARS LATE

On 22 June 2019, in a small town called Kharsawan in Jharkhand, a group of men tied a young Muslim trader named Tabrez Ansari to a lamppost and beat him through the night, forcing him at intervals to chant religious slogans. He died in police custody four days later. It took the Indian legal system exactly six years to convict ten of his attackers of culpable homicide. In the interim, a hashtag-length rumour the same rumour, more or less, that had already killed strangers in Jharkhand, Maharashtra, Assam, and Tripura since 2017 kept circulating: warnings about child-kidnapping gangs ‘roaming the area,’ complete with invented physical descriptions, forwarded from phone to phone until a stranger's face matched somebody's fear. In a market town in Uttar Pradesh's Hapur district that same year, a goat trader named Mohammed Qasim was beaten to death by a mob that had convinced itself, on the strength of a forwarded message, that he was a cattle smuggler. Villagers in the crowd later told investigators they did not know Qasim personally. They knew the rumour.

¹ Pradyumn Sharma and Tanisha Khanna, ‘Bombay High Court brings Fact Check Unit saga to a close: Delivers tie breaker judgment striking down Fact Check Unit set up under IT Rules as unconstitutional’ (*Nishith Desai Associates*, 11 October 2024) <<https://nishithdesai.com/research-and-articles/hotline/technology-law-analysis/bombay-high-court-brings-fact-check-unit-saga-to-a-close-delivers-tie-breaker-judgment-striking-down-fact-check-unit-set-up-under-it-rules-as-unconstitutional-15155>> accessed 23 July 2026

By the time WhatsApp took out full-page newspaper advertisements in 2018 asking Indians to stop and think before forwarding, at least twenty people had already been killed and hundreds injured in what journalists came to call, starkly, the 'WhatsApp lynchings.' The company added a forward limit and a small grey label 'Forwarded many times' to viral messages. It was a modest, almost apologetic intervention, arriving after the deaths rather than before them.

This article is, at one level, about statutes: Section 79 of India's Information Technology Act, the 2021 Intermediary Guidelines and their four subsequent amendments, the state action doctrine borrowed and reworked from American constitutional law, and the judgments *Shreya Singhal*, *Anuradha Bhasin*, *Puttaswamy*, *Kunal Kamra*, *X Corp* that have tried to referee the fight between speech and safety. But underneath every one of those doctrines sits a person: a trader tied to a lamppost, a mother in Manipur who could not call for help because the internet had been switched off for the two-hundredth consecutive day, a comedian who spent a year in court defending his right to call a government claim false, a 22-year-old climate activist arrested for editing a Google document. The law is not an abstraction to any of them. It is the difference between being protected and being left alone with a mob, a blackout, or a summons.²

Scale is part of why this matters more in India than almost anywhere else. India is home to the largest national user base for WhatsApp, Facebook, and YouTube in the world, and among the largest for X and Instagram; well over 800 million Indians now carry a smartphone connection into a public square that did not exist for their parents' generation, most of them communicating not in English but in Hindi, Bengali, Tamil, Marathi, and a dozen other languages that the platforms themselves, as later sections will show, have historically been slower to understand. A moderation failure in Silicon Valley is an abstraction measured in engagement metrics; a moderation failure in Seraikela-Kharsawan is a body in a police mortuary. That asymmetry between where these companies are governed and where their decisions land hardest runs through every case study that follows.

² Rebant Juyal, 'Constitutional Liberties and Cyberspace: Analysing the *Anuradha Bhasin v Union of India* Case and its Impact on Fundamental Rights' (2023) 23(3) *Legal Information Management* <<https://doi.org/10.1017/s1472669623000622>> accessed 23 July 2026

That is the argument this article makes: that the debate over content moderation in India and everywhere is not really a contest between ‘free speech’ and ‘safety’ as competing abstractions. It is a contest over who gets to decide, on what evidence, with what oversight, and how quickly that decision can be reversed when it is wrong. Get that architecture right, and both democratic discourse and vulnerable communities are better protected. Get it wrong, and you can lose both at once, which, this article suggests, is roughly where India stands today.³

WHAT FREE SPEECH WAS BUILT TO PROTECT?

Before turning to statutes and portals, it is worth remembering why free expression is treated as a foundational right rather than a mere convenience. Three justifications recur across constitutional traditions, and all three appear, in different words, in the debates of India's own Constituent Assembly.

The first is epistemic: no single authority, however well-intentioned, reliably knows the truth in advance, so the best available method for sorting good ideas from bad ones is open contest, what American jurisprudence calls the ‘marketplace of ideas.’ The second is participatory: self-government requires that citizens be able to criticise, mock, and organise against those who hold power over them, without asking permission first. The third is personal: expression is bound up with dignity and self-realisation; the ability to say who you are and what you believe is part of what makes a person a citizen rather than a subject.

India's framers folded all three into Article 19(1)(a), the guarantee of freedom of speech and expression, but they paired it deliberately, after considerable argument, with Article 19(2), which permits ‘reasonable restrictions’ in the interests of the sovereignty and integrity of India, state security, public order, decency, morality, contempt of court, defamation, and incitement to an offence. Unlike the First Amendment's near-absolute text, the Indian Constitution builds the balancing act into the right itself. This is not, contrary to a common misreading, an invitation to censor freely. The Supreme Court has repeatedly insisted that ‘reasonable’ is a genuine constraint, not a rubber stamp; restrictions must be proportionate,

³ Abhinav Ravi and Aravind Sundar, ‘The Constitutional Case Against State-Controlled Fact-Checking: A Case Comment on Kunal Kamra v Union of India’ (2025) 36(2) National Law School of India Review <<https://repository.nls.ac.in/nlsir/vol36/iss2/1/>> accessed 23 July 2026

narrowly tailored, and connected by a close causal link to the harm they claim to prevent, not merely a vague discomfort with the speech in question.

That distinction between speech that is merely offensive, annoying, or critical of power, and speech that is causally and proximately linked to violence, defamation, or the collapse of public order is the fulcrum on which nearly every dispute in this article turns. A tweet mocking a politician is protected even if it wounds the politician's pride; a WhatsApp forward falsely identifying a named individual as a child-kidnapper, in a village already primed by three prior lynchings, is a different animal entirely, because it does not just offend; it gets someone killed. The hard cases, as we will see, live in the vast grey zone between those two poles, and the question of who gets to classify speech into one bucket or the other a judge, a bureaucrat, or an algorithm answering to a company's shareholders turns out to matter enormously.⁴

THE SAFE HARBOUR: HOW PLATFORMS ESCAPED RESPONSIBILITY, AND HOW THAT BARGAIN IS NOW BEING REWRITTEN

The American Origin Story: Modern platform regulation begins with a piece of 1996 American legislation that nobody, least of all its authors, expected to become the constitutional backbone of the internet economy. Section 230 of the Communications Decency Act states, in twenty-six words, that no provider of an 'interactive computer service' shall be treated as the publisher or speaker of information provided by someone else. The provision was written to overturn a specific court decision that had held an early online bulletin board liable for a user's defamatory post because it had moderated some content, the reasoning being that any editorial hand at all made a platform a publisher. Congress's fix was to say: moderate as much or as little as you like, in good faith, and you will not be treated as the author of what your users post.

That single sentence is arguably the single most consequential piece of internet legislation ever written. It is the reason a comment section, a product review, a Wikipedia page, or a tweet can exist at all without every platform being sued into bankruptcy by the first defamatory post a user uploads. It is also, three decades later, the subject of a fierce and

⁴ Aditya Tripathi, 'Case Summary: Shreya Singhal v Union of India' (*Law Lex*, 17 April 2020) <<https://lawlex.org/lex-bulletin/case-summary-shreya-singhal-vs-union-of-india/18989>> accessed 23 July 2026

increasingly bipartisan American backlash. As Section 230 turned thirty in February 2026, it no longer looks like the untouchable shield it once was. Multiple bills before the 119th Congress the Sunset Section 230 Act, the Sunset to Reform Section 230 Act, and several narrower proposals conditioning immunity on duties of care around child safety, cyberstalking, and algorithmic amplification have real, if uncertain, momentum, with hearings held and sponsors drawn from both parties. President Trump has separately called for repeal and issued an executive order targeting platforms he accuses of politically motivated moderation. Meanwhile, courts have begun carving their own exceptions, holding that immunity protects hosting third-party content but does not necessarily extend to a platform's own algorithmic design choices about what to recommend and to whom a distinction that, if it holds, would strip away a great deal of the shield's practical value for large recommendation-driven platforms.

Layered onto this is the Supreme Court's 2024 decision in *Moody v NetChoice*, which arose because Florida and Texas, both controlled by Republican legislatures convinced that platforms were systematically censoring conservative voices, passed laws restricting how large platforms could moderate content and requiring individualised explanations for takedowns. The Court did not resolve the ultimate constitutional question; it sent the case back down for a fuller factual record, but Justice Kagan's opinion for the majority planted a flag that will matter for years: content moderation, the Court said, is itself an exercise of a platform's own First Amendment-protected editorial judgment, no different in kind from a newspaper deciding what to print, even when that judgment is exercised by an algorithm at a scale a human editor could never match. That holding cuts hard against laws in Florida, Texas, or anywhere else that try to compel platforms to carry speech they would rather remove. It is the American system, in miniature: platforms are simultaneously shielded from liability for what users say (Section 230) and independently protected in their choices about what to remove (the First Amendment), a double insulation that critics call an accountability vacuum and defenders call the only workable way to run a service hosting billions of daily posts.⁵

The Indian Variation: Safe Harbour with Strings Attached: India borrowed the basic safe-harbour architecture but never adopted America's hands-off philosophy toward the

⁵ 47 U S Code 1934, s 230

conditions attached to it. Section 79 of the Information Technology Act 2000 grants intermediaries a term broad enough to cover everything from an internet service provider to a social media platform to a cloud host immunity from liability for third-party content, but only if the intermediary observes 'due diligence' and, crucially, acts to disable access to unlawful content upon receiving 'actual knowledge' of it, whether through a court order or a notification from an appropriate government agency. That conditionality is the seed from which everything that follows in Indian platform regulation grows: safe harbour in India was never unconditional. It was always a bargain, and the government has spent the years since 2021 steadily renegotiating the terms in its own favour.

The turning point was 2021's Intermediary Guidelines and Digital Media Ethics Code, popularly the 'IT Rules 2021,' which converted a thin, largely unenforced due-diligence requirement into a dense compliance regime: significant social media intermediaries (those crossing a user threshold, currently five million) must appoint a resident grievance officer, a chief compliance officer, and a nodal contact person, all physically located in India; they must publish monthly transparency reports; messaging services must build the capacity to trace the 'first originator' of a message when ordered to under narrow categories of serious offence; and platforms must take down flagged content within tight timeframes or forfeit their safe harbour entirely.

Each subsequent year brought another layer. The 2022 amendment created Grievance Appellate Committees, government-appointed panels with the power to overturn a platform's own content moderation decisions on a user's appeal, giving an ordinary citizen a route to challenge a private company's takedown (or refusal to take down) without going to court. Between March 2023 and June 2024 alone, these committees registered 1,065 appeals and disposed of 937 of them, a meaningful, if imperfect, safety valve for people who feel a platform got it wrong, though critics note the panels are appointed and funded entirely by the government whose own removal requests they may one day have to review. The 2023 amendment went further and more controversially still, empowering a government-run Fact Check Unit to identify 'fake, false or misleading' content about the government's own business and effectively compel its removal a proposal so nakedly self-interested (the government policing claims about the government) that it triggered one of the most closely watched litigation battles in recent Indian internet law, discussed below. Then, in February

2026, the government notified yet another amendment: platforms now face a compressed three-hour window to act on certain takedown notices before risking loss of their safe harbour altogether, alongside new obligations to label AI-generated and synthetic ‘deepfake’ content, a response, MeitY officials have said, to the proliferation of non-consensual synthetic imagery and AI-generated political disinformation ahead of state elections.

Running quietly alongside all of this rule-making has been a second, less visible track: the ‘Sahyog’ portal, launched to let government ministries, state police departments, and other authorised agencies send content-takedown requests to intermediaries directly, framed by the government as a facilitation tool and by its critics as a parallel censorship pipeline that bypasses the due-process safeguards the Supreme Court built into the formal blocking power under Section 69A back in 2015. Between October 2024 and April 2025 alone, roughly 130 takedown notices went out through Sahyog to major platforms including Google, YouTube, Amazon, and Microsoft. X Corp challenged the portal's constitutionality outright, a case examined in detail in Part VI, and lost, in a ruling that has reshaped, for now, the balance of power between platforms and the Indian state.

For all the criticism the Indian regime attracts, and there is much to criticise, it is worth being honest about what it gets right that the American model does not. Section 230's near-total immunity leaves victims of coordinated harassment, non-consensual imagery, or incitement with genuinely nowhere to turn if a platform simply declines to act, short of the slow and unreliable route of suing the anonymous poster directly. India's conditional safe harbour, whatever its faults, at least in theory gives an ordinary citizen, the parent of a bullied teenager, the target of a deepfake, the family of someone killed by a WhatsApp rumour an administrative lever to pull. The trouble, as the next two Parts will show, is who holds the other end of that lever.⁶

A Brief Comparative Note: The European Middle Path: A third model deserves mention, if only in passing, because it is instructive by contrast. The European Union's Digital Services Act, fully applicable since 2024, keeps a Section 230-style liability shield for hosting third-party content but layers onto it detailed, court-reviewable due-diligence obligations calibrated to platform size: risk assessments, algorithmic transparency toward regulators and

⁶ Tripathi (n 4)

researchers, independent audits, and a 'trusted flagger' system for expedited but still judicially reviewable notice-and-action, with the heaviest obligations reserved for 'Very Large Online Platforms' crossing 45 million EU users. Notably, the DSA keeps takedown power largely outside government hands for ordinary content, reserving direct government-ordered removal for narrower categories such as terrorist content, and routes most disputes through independent out-of-court dispute settlement bodies rather than a ministry-run appellate committee. It is neither America's near-total deference to platforms nor India's increasingly executive-centred model, and as India's own Digital India Act inches, glacially, toward an eventual draft, EU-style structured transparency is one of the more frequently cited templates by Indian civil society groups pushing back against a Sahyog-style architecture.⁷

WHO GUARDS THE GUARDIANS? THE STATE ACTION PUZZLE

The American Doctrine, and its Limits: Constitutional free-speech guarantees, in almost every democracy that has one, restrain the state, not private actors. A newspaper can refuse to print your op-ed; a shopping mall can eject you for handing out leaflets its owners dislike; a private social media company can, in principle, ban you from its service for any reason or none because the First Amendment, like Article 19, was written as a shield against government power, not a sword compelling private entities to platform speech they find objectionable. This is the state action doctrine, and it is one of the most consequential and most contested thresholds in constitutional law, because everything downstream depends on which side of the public/private line an actor falls.

American courts carved out a narrow exception in 1946's *Marsh v Alabama*, holding that a company-owned town that functioned exactly like a municipality with streets, a business district, and a resident population could not evict a Jehovah's Witness for leafleting, because it had voluntarily taken on all the trappings of a public space and could not then claim private-property immunity from the Constitution. For decades, civil liberties advocates hoped this 'public function' exception might extend to the internet's new company towns: the social media platforms that had, by the 2010s, become the primary venue for political speech in a way no shopping mall or company town ever was. The Supreme Court itself

⁷ 'Digital Services Act: Questions and Answers' (European Commission) <<https://digital-strategy.ec.europa.eu/en/faqs/digital-services-act-questions-and-answers>> accessed 23 July 2026

seemed to gesture in that direction in 2017's *Packingham v North Carolina*, describing social media as among the most important places, in a spatial sense, for the exchange of views in the modern age.

But the exception never actually extended that far. In 2019's *Manhattan Community Access Corp v Halleck*, the Supreme Court held that even a private nonprofit operating a public-access television channel created by New York City precisely to host public speech was not a state actor merely because it hosted a forum for public expression; hosting speech, even speech of a fundamentally public character, does not itself convert a private operator into 'the state.' Lower courts have since applied that logic squarely to social media companies, rejecting arguments (including from Prager University against YouTube) that platforms hosting political speech at civic scale become state actors merely by virtue of that scale and function. The upshot, confirmed and extended by *Moody v NetChoice* in 2024, is that American platforms remain constitutionally private actors, free indeed, per *Moody*, constitutionally entitled to moderate as they see fit, however central their role in democratic discourse has become. Where the doctrine gets more interesting is at the point of contact between platform and government: in *Knight First Amendment Institute v Trump*, the Second Circuit held that a government official's own Twitter account, used to conduct official business, became a public forum in which he could not constitutionally block critics a ruling about a state actor's use of a private platform, not about the platform itself, but one that illustrates how state action analysis has to track the actor, not merely the medium.⁸

The Indian Variant and the Deeper Problem it Obscures: India's equivalent doctrine flows from Article 12, which defines 'the State' for Part III fundamental rights to include not just government departments but any 'other authority' that is, in substance, an instrumentality of the state a test the Supreme Court has fleshed out over decades through cases like *Ajay Hasia* and *Pradeep Kumar Biswas*, looking at factors such as government funding, government control over management, and whether the entity performs functions of a fundamentally public character. On a straightforward application of that test, a privately owned, foreign-headquartered platform like X, Meta, or YouTube is not an instrumentality of the Indian state; however central it has become to Indian public discourse, the Karnataka

⁸ Erika L Andersen, 'Social Media, the Modern Public Forum: The State Action Doctrine and Resurrection of Marsh' (2024) 50(1) Mitchell Hamline Law Review <<https://open.mitchellhamline.edu/mhlr/vol50/iss1/2/>> accessed 23 July 2026

High Court said as much, bluntly, in the 2025 X Corp judgment, holding that as a foreign company X cannot even invoke Article 19 rights in the first place, since those rights, per longstanding precedent, belong to citizens.

But framing the question as ‘is the platform the state?’ is, this article suggests, the wrong question or at least an incomplete one because it obscures a subtler and arguably more urgent problem: what happens when a plainly private platform is induced, through conditional legal liability, to do the state's censorship work for it, without the constitutional scrutiny that would apply if the state did the same job directly? American scholars call this ‘collateral censorship’ when a law makes an intermediary liable for a third party's speech unless it removes that speech; the intermediary rationally, and often over-cautiously, removes far more than the law actually requires, because a private company weighing legal risk against a stranger's expression will almost always err toward deletion. The state never has to issue a formal, appealable order; it simply threatens the intermediary's business model and lets the intermediary do the deleting. No court ever reviews the individual decision, because no state action, in the constitutional sense, technically occurred even though the state's fingerprints are all over the outcome.

This is precisely the architecture that the Sahyog portal, and Section 79(3)(b) more broadly, enables, and it is why the X Corp litigation was never really about whether X Corp, a for-profit American company, deserved sympathy. It was about whether the government could route what is functionally a blocking power through the conditional-safe-harbour provision (79(3)(b)) rather than the dedicated blocking provision (Section 69A), which the Supreme Court in 2015 had subjected to strict procedural safeguards: a written order stating reasons, notice to the originator or intermediary wherever practicable, and review by a government review committee. X Corp's argument, distilled, was that Sahyog let a police sub-inspector in any state do administratively, in an afternoon, what the Constitution required a Joint Secretary-level review committee to do with reasons on record. The Karnataka High Court, in its 351-page judgment of 24 September 2025, disagreed, characterising the portal not as a censorship apparatus but as, in the Court's own words, ‘an instrument of public good’, a facilitative channel for lawful, authenticated communication between agencies and intermediaries, resting on Rule 3(1)(d)'s due-diligence obligation rather than a freestanding blocking power. The Court held that online expression, like all expression, remains ‘hedged

by restrictions,’ and that a citizen-country’s law ‘cannot be honoured in one jurisdiction and flouted in another’ by a company that operates globally but wants to pick and choose which nation’s rules apply to it.

Whatever one makes of the outcome, and X Corp has said it intends to appeal, the deeper point for this article’s purposes is structural, not personal to any one litigant: when a takedown channel exists that requires no written reasons visible to the public, no independent review committee, and no judicial oversight before it takes effect, the state action doctrine’s neat public/private line stops doing useful work. The platform is private. The decision-maker sending the notice is unambiguously the state. And the person whose post disappears experiences something functionally identical to state censorship, dressed in the procedural clothing of a private company’s ‘due diligence.’ A constitutional framework built entirely around the question ‘is this actor the state?’ will keep answering ‘no’ to that question for the platform even as the substance of state power keeps flowing, unreviewed, through exactly this kind of channel. The right question, this article submits, is not whether the platform is the state, but whether the process by which the state induces removal replicates the safeguards the Constitution would otherwise require if the state acted directly and on the Sahyog model, as currently structured, it largely does not.⁹

THE FACES BEHIND THE FEED

Doctrine explains the architecture. It does not, by itself, convey what is actually at stake when the architecture fails in either direction, over-removal or under-removal. This Part tells four stories that, between them, capture nearly the entire spectrum of harm this article is concerned with.

The Rumour that would not Stop Killing: Between 2017 and 2019, at least twenty-nine people were killed and hundreds injured across roughly a dozen Indian states in incidents traceable, directly or substantially, to WhatsApp-forwarded rumours about child kidnapping gangs and cattle smuggling. The pattern repeated with grim consistency: a message, often lifted from an unrelated viral video shot in another country entirely, would describe strangers moving through the area with sedatives, injections, or cotton swabs, intent

⁹ Samantha Murray, ‘Beyond Coercion: Rethinking State Action Doctrine for Government-Platform Speech Regulation’ (2026) 3(1) Washington University Undergraduate Law Review
<<https://doi.org/10.7936/wuulr.9304>> accessed 23 July 2026

on abducting children; a stranger, a nomadic trader, a tourist, a person who simply looked unfamiliar or spoke a different language would be spotted; and a crowd, primed by weeks of repetition, would form within minutes and dispense its own justice before police could arrive. In Jharkhand's Seraikela-Kharsawan district in May 2017, seven men were beaten to death in a single week on the strength of exactly this rumour. Two years and dozens of deaths later, in the same district, Tabrez Ansari was tied to that lamppost.

What makes these deaths a content-moderation story, rather than simply a law-and-order one, is the specific mechanical property of the platform that carried the rumour: WhatsApp's end-to-end encryption, which rightly protects thousands of dissidents, journalists, and domestic-abuse survivors from surveillance, also made the rumour itself structurally invisible to the platform's own moderation systems, since the company could not read message content even if it wanted to. The only lever available was architectural throttling of how many people a message could be forwarded to at once, and labelling messages that had travelled unusually far. That lever arrived only after the deaths had already made international headlines, and it addressed the speed of the rumours' spread without doing anything to address their truth. This is the sharpest edge of the free-speech-versus-safety debate: no plausible content-moderation regime, however aggressive, could have 'moderated' a message it structurally could not read. The failure here was not a platform choosing profit over safety in a moderation queue; it was a genuine, hard technical limit that no simple appeal to 'just take it down faster' can resolve, and it is a useful corrective to any account of this debate that assumes better moderation is always merely a matter of will.¹⁰

A Blackout in the Hills: On 3 May 2023, ethnic violence broke out in Manipur between the majority Meitei community, concentrated in the Imphal valley, and the Kuki-Zo tribal communities of the surrounding hills, triggered by a dispute over Scheduled Tribe status. Within a day, the state government suspended mobile internet, then broadband, VSAT, and VPN services across the state a shutdown that, through repeated automatic extensions, ran for roughly seven months, becoming the longest state-wide internet blackout in India outside Jammu and Kashmir. Officially, the rationale was the one authorities invariably give: that unrestricted internet access would let 'rumours' and 'hate speech' spread and fuel further

¹⁰ Shakuntala Banaji et al, *WhatsApp vigilantes: an exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India* (London School of Economics and Political Science, 2019)

mob mobilisation, echoing almost exactly the WhatsApp-lynching logic described above, except deployed pre-emptively and at the scale of an entire state.

What the blackout could not do was stop the violence itself, which continued through the silence, eventually killing more than 250 people and displacing tens of thousands. What it did do, for more than two months, was prevent the outside world from finding out. On 4 May – the day the shutdown began a mob paraded two Kuki women naked through the streets of Kangpokpi district; one was later gang-raped. Video of the assault did not surface publicly until mid-July, more than two months later, when it finally leaked out through channels that circumvented the blackout, triggering national outrage, condemnation from the Prime Minister, and arrests that almost certainly would have come faster had the footage travelled at the speed information normally does. Digital rights groups pointed out, accurately, that the shutdown order violated the Supreme Court's own 2020 ruling in *Anuradha Bhasin v Union of India* discussed in Part VI, which had explicitly held that an indefinite suspension order, renewed automatically without fresh, periodic, and published justification, is impermissible. Manipur's order had no end date and was extended roughly twenty times by simple administrative repetition. By 2024, the state had recorded forty-seven separate shutdown episodes totalling more than five thousand hours, the second-longest cumulative shutdown record of any Indian state, and shutdowns have continued intermittently since, including a fresh suspension in April 2026 after a bomb attack killed two children.

The Manipur blackout is, in one sense, the mirror image of the WhatsApp lynchings: where encryption made moderation technically impossible, the shutdown made all speech true, false, urgent, mundane, life-saving impossible at once, on the theory that indiscriminate silence was safer than selective moderation. It is difficult to credit that theory when the single most consequential piece of information to emerge from the entire seven-month crisis, the video that finally forced a national reckoning, only escaped the blackout because someone, somewhere, found a way around it.¹¹

¹¹ Satish Irungbam, 'Repressive connectivity and digital resistance: internet shutdowns, speech and networked counterpublics in Manipur' (2026) 47(8) *Third World Quarterly* <https://econpapers.repec.org/article/tafctwqxx/v_3a47_3ay_3a2026_3ai_3a8_3ap_3a1519-1538.htm> accessed 23 July 2026

The Lobbyist and the Lamppost Report Nobody Wanted Written: In August 2020, the Wall Street Journal published an investigation, based on interviews with current and former employees, reporting that Ankhi Das, then Facebook India's top public policy executive, had intervened to stop the company's own internal hate-speech enforcement from applying to T. Raja Singh, a BJP legislator from Telangana who had publicly called for Rohingya Muslim refugees to be shot, along with at least three other Hindu nationalist figures and groups that Facebook's own content-review staff had flagged for promoting or participating in violence. According to the report, Das told colleagues that punishing violations by politicians from the ruling party would damage the company's business prospects in its largest market by user count. The story also surfaced Das's own social media posts expressing open political sympathy for the ruling party's ideological project; in one line, she reportedly described the effort to 'rid India of state socialism' as the culmination of thirty years of grassroots work.

The episode reads, on its face, like a story about corporate cowardice, and it is that. But it is more precisely a story about the structural incentive built into the entire safe-harbour bargain described in Part III: a platform that depends on continued government goodwill for market access, tax treatment, and regulatory forbearance has every commercial reason to bend its own hate-speech rules around the very political actors best positioned to make its life difficult the opposite failure mode from over-removal, and one that no amount of Sahyog-style government takedown power actually corrects, since a government uninterested in enforcing hate-speech norms against its own allies will simply decline to send the notice in the first place. Facing sustained public pressure and a parliamentary committee summons, Facebook did eventually ban Raja Singh under its 'dangerous individuals' policy; Ankhi Das resigned two months after the story broke, and the company later commissioned an independent human rights audit of its India operations. The episode remains one of the clearest public illustrations that content moderation failure is not always a story about too much censorship or too little; sometimes it is a story about a platform's incentives running in exactly the wrong direction on exactly the speech that matters most.¹²

¹² Dr Janjira Sombatpoonsiri and Dr Sangeeta Mahapatra, 'Regulation or Repression?: Government Influence on Political Content Moderation in India and Thailand' (GIGA, 2024) <<https://www.giga-hamburg.de/en/publications/contributions/regulation-or-repression-government-influence-on-political-content-moderation-india-and-thailand>> accessed 23 July 2026

A Toolkit, a Tractor Rally, and a 22-year-old in Tihar Jail: On 26 January 2021, India's Republic Day, a farmers' tractor rally protesting new agricultural laws turned violent in Delhi, with one protester killed and police officers injured. Days later, on 1 February, Twitter temporarily withheld more than 250 accounts inside India belonging to farmer unions, journalists, a magazine, and activists after a request from the Ministry of Electronics and Information Technology citing a hashtag it characterised as fake, provocative, and intimidatory. Public outrage over what looked, to many, like the silencing of an entire protest movement led Twitter to restore the accounts within hours; the government responded with a formal non-compliance notice and, according to contemporaneous reporting, warnings that Twitter's own India-based staff could face criminal liability if the company continued to resist. Twitter partially complied over the following weeks, permanently suspending some accounts while keeping others geo-blocked only within India, visible everywhere else in the world, invisible to Indians.

The controversy widened further when the Swedish climate activist Greta Thunberg tweeted in solidarity with the farmers and shared a crowd-sourced 'toolkit' document, a fairly ordinary organising aid of the kind climate and protest movements share constantly, listing hashtags, contact points, and suggested actions for supporters. Delhi Police opened an investigation into the toolkit's authors, and on 13 February 2021, a 22-year-old Bengaluru climate activist named Disha Ravi, the local founder of Thunberg's youth movement, was arrested and charged with sedition and criminal conspiracy for allegedly helping to edit and circulate it. She spent ten days in custody, including time in Tihar Jail, before a Delhi court granted her bail. The additional sessions judge, Dharmender Rana, used his eighteen-page order to deliver an unusually pointed rebuke of the prosecution's case, finding the evidence 'scanty and sketchy' and, in a line that circulated as widely as the toolkit itself, observing that citizens function as the conscience-keepers of government in any democracy, and cannot be jailed simply for disagreeing with state policy.

Read together, the account-withholding episode and Disha Ravi's arrest illustrate a pattern distinct from either of the two failure modes above. This was not a platform declining to enforce its own rules against an ally (as with Facebook), nor a technical inability to moderate encrypted content (as with WhatsApp). It was the state itself treating ordinary advocacy speech, a hashtag, a shared document, as a security threat, and reaching for the two levers

described in Parts III and IV simultaneously: pressure on the platform to withhold content, and direct criminal process against the individual who produced it. Both levers were eventually rolled back; Twitter restored most accounts, and Ravi was freed on bail with the court's evident scepticism of the state's case on the record, but the rollback took public outrage, international attention, and, in Ravi's case, ten days in custody. A system that depends on that much friction to correct its own overreach is not, this article suggests, one that can be relied upon to get the harder, less internationally visible cases right.¹³

THE COURTS PUSH BACK, AND THE GOVERNMENT PUSHES FORWARD

If Part V showed what happens when the architecture fails, this Part turns to the institution that has, repeatedly if unevenly, tried to fix it: the Indian judiciary.

Shreya Singhal and the Moment the Internet was Taken Seriously: The foundational case is *Shreya Singhal v Union of India* (2015), triggered by the arrest of two young women in Maharashtra: one for a Facebook post questioning a city-wide shutdown after a politician's death, the other simply for 'liking' it under Section 66A of the IT Act, a provision that criminalised sending, via a computer, any information deemed 'grossly offensive' or having a 'menacing character,' without defining either term. The Supreme Court struck the entire provision down as unconstitutionally vague and overbroad, holding that its open-ended language swept up an enormous amount of protected speech satire, criticism, hyperbole, and art alongside any genuinely dangerous content, and that citizens could not be expected to guess, at the risk of arrest, where the line actually fell. In the same judgment, the Court did not strike down Section 69A, the government's dedicated content-blocking power, but it read the provision narrowly and insisted on the procedural safeguards described in Part IV: written, reasoned orders; notice wherever practicable; and review by a designated committee precisely the safeguards that the *X Corp* litigation, a decade later, would argue the Sahyog portal was circumventing. *Shreya Singhal* remains, ten years on, the single most important free-speech precedent in Indian internet law, and its central insight that vague, discretion-

¹³ Stein Monteiro, 'Farmer protests in India and the mobilization of the online diaspora on Twitter' (2021) SSRN <<https://doi.org/10.2139/ssrn.3849515>> accessed 23 July 2026

heavy speech restrictions chill far more expression than they legitimately prevent runs like a red thread through every case discussed below.¹⁴

Puttaswamy and the Right that made Encryption a Constitutional Question: Two years later, in Justice K.S. Puttaswamy v Union of India (2017), a nine-judge bench unanimously held that the right to privacy is intrinsic to the right to life and personal liberty under Article 21, encompassing informational privacy, communicational privacy, and decisional autonomy. The ruling did not arise from a content-moderation dispute at all; it arose from a challenge to India's biometric identity programme, but its reasoning has since become the load-bearing precedent in the WhatsApp traceability litigation still pending before the Delhi High Court, filed after the 2021 IT Rules required messaging platforms to identify the 'first originator' of certain flagged messages. WhatsApp's position, resting heavily on Puttaswamy, is that no technical method exists to enable message-tracing without either breaking end-to-end encryption for all 500-plus million Indian users or building a permanent identification capability that itself constitutes a standing surveillance infrastructure; the company has told the court it would rather stop operating in India than build it. The government maintains that the rule targets only the 'first Indian originator' of specific, narrowly defined categories of serious content and can, in principle, be engineered without breaking encryption, though technologists broadly dispute that claim as, at minimum, unproven, and the case remains unresolved more than five years after it was filed, a reminder that not every hard case in this space gets a clean, timely answer.¹⁵

Anuradha Bhasin and the Limits of the Switch: In August 2019, following the revocation of Jammu and Kashmir's special constitutional status, the government imposed a total communications blackout across the region, including internet, mobile, and landline services, that lasted for months. Anuradha Bhasin v Union of India (2020), brought by the executive editor of a Kashmiri newspaper unable to publish, produced a Supreme Court ruling that

¹⁴ Shaurya Shandilya, 'SHREYA SINGHAL v UNION OF INDIA (2015): FREEDOM OF SPEECH IN THE DIGITAL ERA' (2026) 3(9) The Indian Journal for Research in Law and Management <<https://ijrlm.com/wp-content/uploads/jms/ms/6a54e678dad46.-Union-of-India-2015-Freedom-of-Speech-in-the-Digital-.pdf>> accessed 23 July 2026

¹⁵ Moazzam Daksh Hada, 'PRIVACY ON INSTAGRAM VERSUS GOVERNMENT TRACKING: THE INDIAN LEGAL CONUNDRUM (A CRITICAL ANALYSIS OF END-TO-END ENCRYPTION ROLLBACK AND THE FIRST ORIGINATOR TRACEABILITY MANDATE)' (2026) 8(3) IJLLR Journal <<https://www.ijllr.com/post/privacy-on-instagram-versus-government-tracking-the-indian-legal-conundrum-a-critical-analysis-of>> accessed 23 July 2026

has become the constitutional benchmark for every internet shutdown since, including Manipur. The Court held that the freedom to access the internet is protected under Article 19(1)(a) and 19(1)(g) (speech and trade, respectively), that any restriction must satisfy the proportionality test: the least restrictive measure necessary to achieve a legitimate aim, for the shortest defensible duration and, critically, that indefinite suspension orders are impermissible: every order must be time-bound, published, and subject to periodic administrative review, with the underlying material made available to a review committee and, where feasible, to the courts. The ruling was, on paper, a significant win for digital rights. In practice, as the Manipur case study in Part V demonstrates, state governments have continued to issue open-ended shutdown orders and simply extend them by administrative rote; the letter of Anuradha Bhasin's procedural requirements was observed just enough to survive challenge, its underlying spirit largely evaded. It is a pattern that should give pause to anyone tempted to think that a strong judicial precedent, once written, enforces itself.¹⁶

Kunal Kamra and the Fact-Check Unit that Fact-Checked itself out of Existence: The most direct judicial confrontation with executive overreach in recent Indian content-moderation law came in *Kunal Kamra v Union of India*, a challenge brought by the stand-up comedian alongside the Editors Guild of India and the Association of Indian Magazines against the 2023 amendment empowering a central government Fact Check Unit to label online content about 'any business of the Central Government' as fake, false, or misleading with the practical effect, given the accompanying safe-harbour conditions, of forcing platforms to remove or demote whatever the FCU flagged. A two-judge Bombay High Court bench split in January 2024, with Justice G.S. Patel finding the rule unconstitutional and Justice Neela Gokhale upholding it, sending the matter to a tie-breaking third judge. On 20 September 2024, Justice A.S. Chandurkar sided with Patel, striking the amendment down as violating Articles 14, 19, and 21, holding that the phrase 'fake, false or misleading' was impermissibly vague, that the rule effectively made the government the sole and final arbiter of truth about its own conduct, and that it failed the proportionality standard the Court had set out five years earlier in *Anuradha Bhasin*. The Union government has appealed to the Supreme Court, which has issued notice but declined to stay the Bombay High Court's judgment,

¹⁶ Devdutta Mukhopadhyay and Apar Gupta, 'Jammu & Kashmir Internet Restrictions Cases: A Missed Opportunity to Redefine Fundamental Rights in the Digital Age' (2020) *Indian Journal of Constitutional Law* <<https://ssrn.com/abstract=4146203>> accessed 23 July 2026

meaning the Fact Check Unit remains, for now, dead in the water, a genuine victory for the argument that a government should not get to be judge, jury, and content moderator in disputes about its own record.¹⁷

X Corp v Union of India and the Road Not Fully Closed: Which brings the analysis back to where Part IV left it. In the X Corp litigation, decided in September 2025, the Karnataka High Court took the opposite path from Kunal Kamra on a structurally similar question: whether an executive-controlled mechanism could bypass the procedural safeguards Shreya Singhal had attached to Section 69A and upheld the government's position, finding the Sahyog portal a lawful exercise of the conditional safe-harbour power under Section 79(3)(b) rather than an unconstitutional end-run around Section 69A. The Court's reasoning leaned heavily on a long historical survey of communications regulation stretching back to eighteenth-century postal law, on the premise that every new communications technology in history has been subject to public regulation, and on the observation that a foreign corporation cannot invoke Article 19 protections designed for Indian citizens in the first place. X Corp has said it will appeal to the Supreme Court, and Digipub, a coalition of Indian digital news publishers that intervened in the case, has continued to press the due-process concerns even as the underlying constitutional challenge to the company's standing was rejected.

Read side by side, Kunal Kamra and X Corp reveal a judiciary genuinely divided not simply between judges, but within the broader doctrinal architecture itself about how much procedural insulation from executive control this domain actually requires. One court concluded that a government fact-checking its own conduct crosses a constitutional line no matter how it is dressed up; another concluded that a government-run takedown portal, so long as it is framed as facilitation rather than blocking, does not. Both cases, notably, invoke Shreya Singhal as authority for opposite conclusions a sign, if one were needed, that the 2015 judgment's safeguards were necessary but were never, by themselves, sufficient to settle where the line sits once new administrative technology finds new procedural workarounds.¹⁸

¹⁷ Ravi (n 3)

¹⁸ Sombatpoonsiri (n 12)

DRAWING THE LINE

Having traced the statutes, the doctrine, the human cost, and the case law, this Part attempts to state, plainly, where the line between protecting democratic discourse and preventing targeted hate speech should sit and what institutional design would actually hold it there.

Start with the distinction the law already half-recognises but rarely enforces consistently: speech that criticises, mocks, or embarrasses power is categorically different from speech that identifies a real person or group as a target and supplies a motive or method for harming them. A tweet accusing a Chief Minister of corruption, however intemperate, sits in the first category even if the government dislikes it intensely; a WhatsApp forward naming a specific community as child-kidnappers, circulating in a district that has already seen three lynchings, sits unambiguously in the second, regardless of who happens to be in power. The test the Indian Supreme Court has already articulated is proximate, not merely remote, connection to incitement; a close causal link, not a generalised discomfort, is the correct test. The problem, as Parts V and VI demonstrate, is not that India lacks the right legal standard. It is that the institutions applying it are increasingly executive rather than judicial, and increasingly opaque rather than reviewable.

From that diagnosis, several concrete design principles follow. First, restore the wall between the entity that wants content removed and the entity that decides whether it should be. Section 69A's safeguards written reasons, an independent review committee, and notice wherever practicable exist precisely because the Shreya Singhal Court understood that a government agency asking for its own critic's content to be removed is an inherently conflicted position. Whatever the Karnataka High Court's view of Sahyog's legality, the policy case for routing every takedown request, regardless of the statutory hook used to justify it, through a comparably independent and reasoned process remains strong; a three-hour compliance clock, of the kind introduced in February 2026, sharpens the platform's incentive to comply first and litigate later, precisely the collateral-censorship dynamic Part IV describes.

Second, make the Grievance Appellate Committees genuinely independent, not merely government-appointed. A committee with the power to overrule a platform's moderation call is a good idea in principle; it gives an ordinary user recourse without the expense of a

court case, and the volume of appeals it has already handled shows real demand. But a body wholly appointed and funded by the government cannot credibly review that same government's own takedown requests, and its legitimacy with users across the political spectrum depends on visible independence, ideally including judicial or cross-party appointees and a public record of its reasoning in each disposed appeal.

Third, treat internet shutdowns as the constitutional emergency measure Anuradha Bhasin said they were, not a routine administrative reflex. Manipur's pattern of automatic extension, no fresh public justification, no meaningful periodic review is precisely the failure mode the 2020 judgment was written to prevent, and its recurrence suggests that procedural compliance without an enforcement mechanism is not compliance at all. A published, judicially reviewable shutdown register, with automatic sunset rather than automatic renewal, would convert the doctrine into a working constraint rather than a citation lawyers make in hindsight.¹⁹

Fourth, be honest about encryption's genuine trade-off rather than pretending a costless technical fix exists. The WhatsApp lynchings show what happens when moderation is structurally impossible; the pending traceability litigation shows a government reluctant to accept that limitation as final. There may be narrower, judicially supervised, case-specific mechanisms closer to a search warrant than a standing surveillance capability that thread this needle better than either extreme, but they require the same institutional honesty the rest of this framework depends on: independent authorisation, narrow scope, and a public accounting of how often the power is used and against whom.

Fifth, and perhaps most important, recognise that platforms themselves are not neutral referees, and regulation cannot assume they are. The Facebook-BJP episode shows a platform bending its own rules around a ruling party for commercial reasons; nothing in Indian law currently requires platforms to publish, in a standardised and independently auditable form, how consistently they apply their hate-speech policies across political affiliations precisely the kind of algorithmic and enforcement transparency the EU's Digital Services Act now mandates for its largest platforms, and precisely the gap India's still-unfinished Digital India

¹⁹ Ben Medeiros and Pawan Singh, 'Addressing Misinformation on WhatsApp in India Through Intermediary Liability Policy, Platform Design Modification, and Media Literacy' (2020) 10 Journal of Information Policy <<https://doi.org/10.5325/jinfopoli.10.2020.0276>> accessed 24 July 2026

Act has an opportunity to close, if it resists the temptation to fold transparency obligations into yet another executive-controlled compliance portal.

Sixth, close the language gap that sits underneath nearly every case study in this article. When the whistleblower Frances Haugen filed her 2021 disclosure with the American securities regulator, the internal Facebook research she submitted included an unusually candid admission from the company's own staff: automated hate-speech 'classifiers' existed for Hindi only from 2018 and for Bengali only from 2020, years after both languages had been vehicles for exactly the dehumanising, communally charged content this article has traced from Jharkhand to Manipur. One internal case study on networks linked to Hindu nationalist groups noted plainly that fear-mongering, anti-Muslim content was going unflagged for want of language coverage, and that companywide, automated systems were catching a vanishingly small fraction of reported hate speech. India has twenty-two scheduled languages and hundreds more in daily use; a moderation system built and tested primarily in English, with regional-language capability arriving years late and unevenly, will always lag furthest behind in precisely the places of rural Jharkhand, small-town Uttar Pradesh, and the hill districts of Manipur where the human cost of that lag has proven highest. No amount of takedown-portal speed or Grievance Appellate Committee independence compensates for a detection system that cannot read the language the harm is written in; closing that gap, through mandated language-coverage disclosure and independently verified regional-language moderation capacity, belongs alongside procedural safeguards as a first-order design requirement, not an afterthought.

Finally, none of these substitutes for the slower, less legislatively satisfying work of building resilience into the public itself: local-language media literacy in exactly the districts where the WhatsApp lynchings occurred, well-resourced independent fact-checking that does not depend on a government's own verdict about its own conduct, and civic education that treats 'conscience-keeping,' in Judge Rana's phrase, as a citizen's right rather than a threat to be managed. Takedown speed and shutdown breadth are the tools of a state that has already lost the argument for truth in the room where the rumour started. A robust democratic culture sceptical of forwards, literate in how misinformation spreads, and confident enough

in its own institutions not to need a Fact Check Unit adjudicating its own legitimacy is the tool of one that has not.²⁰

CONCLUSION

It took six years for the law to catch up to what happened to Tabrez Ansari in Kharsawan. In that time, India built an entire new architecture of digital governance: grievance officers, appellate committees, a fact-check unit that lived and died in the same eighteen months, a takedown portal upheld by one court and still being appealed to the next, a traceability case still unresolved half a decade after it was filed. Some of that architecture would, if it had existed and worked as intended in 2017, plausibly have done nothing at all to save him: no takedown notice, however fast, reaches an encrypted message the platform cannot read; no fact-check unit, however empowered, catches a rumour that never mentions the government at all. Other parts of it a functioning appeals process, transparent enforcement of hate-speech rules across party lines, a state government less willing to treat an entire population's internet access as expendable at the first sign of unrest might genuinely have helped, in Manipur if not in Jharkhand, in the farmers' protest if not in the lynchings.

The honest conclusion of this article is not that India has gotten the balance wrong in one uniform direction. It has, at different moments and through different mechanisms, gotten it wrong in both directions at once: too slow and too structurally blind to stop a rumour that killed dozens of its most vulnerable citizens, and simultaneously too quick, in other moments, to reach for the same blunt instruments: a blackout, a sedition charge, a portal with no public record of its own reasoning against speech that was inconvenient rather than dangerous. Those two failures are not opposites that cancel out. They are the same underlying defect, viewed from different sides: a system that still, a decade after Shreya Singhal, has not fully committed to putting a reasoned, reviewable, independent decision-maker between a claim of harm and the removal or the silence that follows it.

Judge Dharmender Rana, granting bail to a 22-year-old for editing a Google document, wrote that citizens are the conscience-keepers of their government. It is worth remembering that a

²⁰ Rama Adithya Varanasi et al., 'Accost, Accede, or Amplify: Attitudes towards COVID-19 Misinformation on WhatsApp in India' (Conference: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing System, New Orleans, LA, USA, 2022) <<https://doi.org/10.1145/3491102.3517588>> accessed 24 July 2026

government, in turn, is meant to be the keeper of something too not of the truth, which no single institution should be trusted to police alone, but of the conditions under which a rumour in a WhatsApp group and a fact in a courtroom can still, eventually, be told apart. Tabrez Ansari's family waited six years for that telling apart to happen in his case. The measure of whether India's digital governance experiment has actually worked will not be how fast the next takedown notice travels. It will be whether the next village, primed by the next rumour, gets the six minutes it needs instead of the six years his did not. The debate over content moderation in India and everywhere is not really a contest between 'free speech' and 'safety' as competing abstractions. It is a contest over who gets to decide, on what evidence, with what oversight, and how quickly that decision can be reversed when it is wrong.