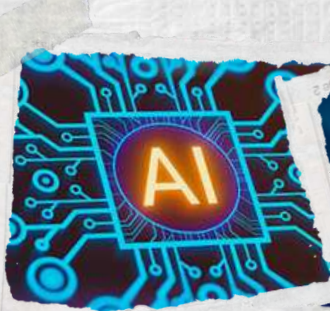


# CONFLICT IN THE AGE OF ALGORITHMS:

*DISINFORMATION, PLATFORM  
POWER AND THE INDIA-PAKISTAN  
CRISIS OF 2025*

# About

Digital Rights Foundation (DRF) is a women-led, not-for-profit organization working since 2013 to advance digital rights and online freedoms across South Asia and the broader Global Majority. We advocate for inclusive, rights-respecting digital spaces by driving policy change through strategic engagement with relevant stakeholders. Our work focuses on making digital platforms and emerging technologies more equitable, accessible, and accountable. At the grassroots level, we empower individuals particularly women, marginalized communities, and human rights defenders with the tools and knowledge to navigate the Internet safely and assert their rights online.

## Contact Information:

info@digitalrightsfoundation.pk

www.digitalrightsfoundation.pk

# Acknowledgements

This report comes at a time when the world is seeing an unprecedented rise in cyber warfare tactics that leverage AI-enabled technologies to spread hate, confusion, panic and to mislead individuals in the midst of political and social unrest. DRF's findings highlight how during the India-Pakistan conflict of May 2025, there was a steady increase in disinformation across social media platforms which amplified hardships faced by communities and citizens in both countries. The report also goes into detail on the growing role of AI in spreading false narratives, hate speech and inciting violence and the impact this has on destabilizing already fragile environments.

This report would not have been possible without the contributions of esteemed partners and collaborators. DRF would like to thank the Center for the Study of Organized Hate (CSOH) for providing us with data from India, and Brandon Silverman and David Thiel for giving us access to data from BlueSky.

**Author(s):** Seerat Khan (Research Lead), Hassan Khan (Research Associate), Sara Imran (Research Associate), and Maria Nazar (Research Associate)

**Edited by:** Seerat Khan

**Design and Layout by:** Ahsan Zahid and Talha Umar

# Table of Contents

|                                                                                  |           |
|----------------------------------------------------------------------------------|-----------|
| <b>Executive Summary</b>                                                         | <b>01</b> |
| <b>Introduction</b>                                                              | <b>02</b> |
| Conflict Context .....                                                           | 02        |
| India-Pak Tensions Before May 2025 .....                                         | 03        |
| Why Disinformation Thrives During Geopolitical Crises .....                      | 05        |
| <b>Literature Review</b>                                                         | <b>07</b> |
| Global Case Studies on Information Warfare in Conflict Settings .....            | 07        |
| Russia-Ukraine War (2022-) .....                                                 | 08        |
| Israel-Hamas War (2023-) .....                                                   | 09        |
| Israel-Iran War (2025) .....                                                     | 10        |
| Iran-US-Israel War (2025–26) .....                                               | 11        |
| Pakistan-Afghanistan Border Tensions (2025) .....                                | 11        |
| Gaps in Existing Scholarship .....                                               | 11        |
| Platform Accountability: How Social Media Platforms Address Disinformation ..... | 12        |
| <b>Methodology</b>                                                               | <b>14</b> |
| Research Objectives .....                                                        | 14        |
| Data Sources and Collection .....                                                | 15        |
| Analytical Framework .....                                                       | 15        |
| Limitations .....                                                                | 16        |
| <b>Findings</b>                                                                  | <b>17</b> |
| Platform-Specific Analysis .....                                                 | 17        |
| X (Formerly Twitter) .....                                                       | 17        |
| TikTok .....                                                                     | 20        |
| YouTube .....                                                                    | 22        |
| Facebook & Instagram .....                                                       | 29        |
| Bluesky .....                                                                    | 32        |
| Cross-Platform Comparison .....                                                  | 33        |

## **Discussion 35**

### **Platform Accountability & Policy Gaps 39**

|                                                                                   |    |
|-----------------------------------------------------------------------------------|----|
| Challenges in regulating cross-border disinformation .....                        | 40 |
| Pakistan's disinformation clauses in PECA are leading to further censorship ..... | 41 |
| The need for regional and global cooperation .....                                | 42 |

### **Recommendations 43**

|                             |    |
|-----------------------------|----|
| Platforms .....             | 43 |
| Governments .....           | 44 |
| Civil Society & Media ..... | 45 |

## **Conclusion 46**

# Executive Summary

The May 2025 India-Pakistan crisis demonstrated that military escalation and information disorder now unfold simultaneously. After the 22 April attack in Pahalgam that killed 26 civilians<sup>[1]</sup>, India downgraded ties with Pakistan, suspended the Indus Waters Treaty<sup>[2]</sup>, and launched strikes under Operation Sindoor on 7 May<sup>[3]</sup>; Pakistan condemned the strikes, launched its own military response<sup>[4]</sup>, and both sides moved toward a ceasefire understanding on 10 May that the United Nations publicly welcomed.<sup>[5]</sup> The confrontation was short in duration but intense in tempo, producing exactly the kind of fast-moving, high-uncertainty environment in which false claims, manipulated media, rumor cascades and strategic propaganda can flourish.

Amidst this war it was seen that disinformation is not incidental to crises of this kind, it is structurally enabled by them. Research shows that rumors spread fastest when authoritative information is scarce and contested in an emotionally charged and identity-laden environment. Throughout the period of escalations and thereafter online content over social media platforms farther drifted from neutral information. It was seen that engagement-optimized algorithmic systems amplified low-credibility material and states, partisans and commercial actors all have incentives to shape narratives in real time.

This study maps the patterns of disinformation that emerged during the May tensions and compares how those patterns traveled across X, Facebook, BlueSky, YouTube and TikTok. The study identifies the principal actors and narratives involved, assesses platform moderation and state responses and the accountability gaps left by current platform governance arrangements.

<sup>[1]</sup> <https://www.aljazeera.com/news/2025/5/2/pahalgam-attack-a-simple-guide-to-the-kashmir-conflict>

<sup>[2]</sup> <https://www.bbc.com/news/articles/cd7vjyeyypqo>

<sup>[3]</sup> <https://www.aljazeera.com/news/2025/5/6/india-launches-attacks-on-several-sites-in-pakistan>

<sup>[4]</sup> <https://www.aljazeera.com/news/2025/5/10/pakistan-launches-operation-bunyan-marsoos-what-we-know-so-far>

<sup>[5]</sup> <https://www.reuters.com/world/asia-pacific/world-reacts-after-india-pakistan-agree-stop-firing-2025-05-10/>

# Introduction

## Conflict Context

### Timeline of key events

| Date               | Event                                                                                                                     |
|--------------------|---------------------------------------------------------------------------------------------------------------------------|
| Apr 22, 2025       | Attack in Pahalgam kills 26 civilians                                                                                     |
| 23 - 24 April 2025 | India suspends Indus Waters Treaty, downgrades ties. Pakistan's NSC rejects Indian measures and signals reciprocal steps. |
| May 7, 2025        | India announces Operation Sindoor against nine sites in Pakistan and Azad Kashmir                                         |
| 8 - 10 May 2025    | Missile, drone, artillery and air-base claims escalate on both sides.                                                     |
| May 10, 2025       | Pakistan launches Operation Bunyan Marsoos; hours later a ceasefire is announced by US President Donald Trump             |

The immediate escalations for the crisis was the 22 April 2025 attack in Pahalgam, in Indian-administered Kashmir, in which 26 civilians were killed. India blamed Pakistan for these attacks, while Pakistan denied involvement and called for an international inquiry. India then suspended the Indus Waters Treaty, closed the Attari land crossing<sup>[6]</sup>, downgraded diplomatic ties and cancelled visas for Pakistani nationals under a regional scheme<sup>[7]</sup>. Pakistan responded with reciprocal measures, including the closure of its airspace to Indian airlines, which has remained in effect to date.<sup>[8]</sup> On 7 May, India announced that it had struck nine sites in Pakistan and Pakistan-administered Kashmir under “Operation Sindoor,” describing the action as focused, measured and non-escalatory. Pakistan condemned the strikes as unlawful aggression and subsequently announced retaliatory action. After several days of drone, missile, artillery and information escalation, both sides announced a halt to military action on 10 May, though accusations of violations followed almost immediately.<sup>[9]</sup>

## India-Pak Tensions Before May 2025

Interminable tensions between the two South Asian neighbors have characterized the region's political landscape for over 70 years. The political consolidation of right-wing nationalism in Indian democracy, as evident by the rise of Modi in the 2010s, opened up a new phase in regional hostilities between the nuclear-armed countries. Discourses around religious extremism, Islamophobia, terrorism, and hyper nationalism were regularly drawn upon in mainstream forums to discredit and disavow the territorial sovereignty of the other country. It is in this context of rising tensions that the Indo-Pakistan conflict came to be played out in online forums and digital spaces. Long before the Pahalgam attacks that opened the door for a full-blown conflict between India and Pakistan, a coordinated proxy war was being carried out by the two countries seeking to leverage the digital space towards shaping global public opinion<sup>[10]</sup>.

The avalanche of disinformation examined over the course of the May conflict cannot be viewed in isolation from pre-existing efforts by both countries to capture public opinion in the international community. Disinformation narratives are wedded to powerful institutional structures in the media landscape that have, for years, been allowed to construct hateful narratives against foreign adversaries. The EU Disinfo Lab in 2021 released a report claiming that it had uncovered a web of disinformation networks that are amplifying and spreading hateful narratives against Pakistan. This network consisted of social media accounts, private media platforms (including Indian diaspora-owned websites), and NGOs<sup>[11]</sup>. The coordinated campaign involved multiple UN-accredited Indian NGOs (e.g Baluchistan House and South

<sup>[6]</sup> <https://www.thehindu.com/news/national/attari-wagah-border-between-india-and-pakistan-completely-shut/article69521483.ece>

<sup>[7]</sup> [https://www.lemonde.fr/en/asia-and-pacific/article/2025/04/24/india-orders-pakistani-citizens-to-leave-by-april-29\\_6740585\\_153.html](https://www.lemonde.fr/en/asia-and-pacific/article/2025/04/24/india-orders-pakistani-citizens-to-leave-by-april-29_6740585_153.html)

<sup>[8]</sup> <https://tribune.com.pk/story/2608999/pakistan-extends-ban-on-indian-aircraft-in-airspace-until-june-24>

<sup>[9]</sup> <https://www.bbc.com/news/articles/cvg9d913v20o>

<sup>[10]</sup> <https://www.techpolicy.press/the-information-crisis-that-brought-india-and-pakistan-to-the-brink/>

<sup>[11]</sup> <https://www.aljazeera.com/opinions/2021/1/4/are-india-and-pakistan-in-a-fifth-generation-war>

Asian Democratic Forum) and over 265 fake media outlets propagating explicitly anti-Pakistan narratives, promoting the narrative of Pakistan being a sponsor of global terror. Simultaneously, separatist causes in Balochistan were actively lauded in a way that sought to co-opt existing Baloch grievances for geopolitical objectives. These narratives were propagated most extensively by Asian News International (ANI), a private media company that amplified news coverage from these outlets<sup>[12]</sup>.

On the other hand, state institutions in Pakistan have regularly leveraged the national security threats emanating from their Eastern border to implicate ethnic movements and dissidents. On digital spaces, institutions such as the Pakistan Telecommunications Authority (PTA) have engaged in arbitrary bans on social media websites over allegations that such spaces are hotbeds of terrorist activity and recruitment<sup>[13]</sup>. Digital authoritarianism through arbitrary bans and deliberate internet slowdowns significantly hamper the democratic space available for dissidents and citizens to participate in political discourse, even if that discourse involves refuting propaganda narratives emanating from “enemy states.” On May 2025, for instance, after almost a year-long ban on X, the state arbitrarily restored social media access and praised internet users within Pakistan for their efforts in countering Indian propaganda narratives<sup>[14]</sup>.

Outside of internet restrictions under the guise of national security imperatives, the Pakistani state has also repeatedly disseminated problematic discourses surrounding the alleged links between dissident movements and Indian expansionism. This is evident in consistent discrediting of nationalist movements as “Fitna-al-Hindustan”, implicating them directly as proxies weaponized by India to threaten Pakistan’s sovereignty<sup>[15]</sup>. While these labels are applied explicitly only to armed insurgents operating in Balochistan, they have also been deployed indirectly against non-violent movements such as the Baloch Yakhjehti Committee and individuals like Mahrang Baloch, who has been incarcerated and recently been given a life sentence due to this<sup>[16]</sup>.

In conclusion, this section establishes the pre-existing state strategies, social media discourses, and the broader political landscape, which was amplified during the May 2025 conflict between India and Pakistan. Both countries perceive digital spaces as critical to shaping public opinions and discrediting their adversaries. While Indian media has leveraged its vast connections in state institutions and diaspora communities to construct coordinated campaigns against Pakistan, the latter has primarily viewed these spaces with suspicion-engaging in arbitrary bans over websites considered a serious threat to national security.

<sup>[12]</sup> <https://www.disinfo.eu/publications/indian-chronicles-deep-dive-into-a-15-year-operation-targeting-the-eu-and-un-to-serve-indian-interests/>

<sup>[13]</sup> <https://link.springer.com/article/10.1007/s10767-024-09493-2>

<sup>[14]</sup> <https://dailytimes.com.pk/1318674/students-are-pakistans-first-line-of-defense-in-information-war-says-dg-ispr/>

<sup>[15]</sup> <https://tribune.com.pk/story/2548800/govt-declares-baloch-terror-groups-organizations-as-fitna-al-hindustan>

<sup>[16]</sup> <https://www.bbc.com/news/articles/cpweywygylo>

# Why Disinformation Thrives During Geopolitical Crises

During this crisis, it was seen that disinformation thrives in geopolitical crises because crises create acute informational vacuums. In breaking-news environments, information arrives incrementally, eyewitness accounts become hard to verify, official accounts share incomplete or contradictory information and audiences are forced to interpret events before official confirmations are available.

Research in the past has repeatedly shown that this kind of uncertainty is a primary condition for rumor formation and circulation, especially during fast-moving emergencies when users turn to platforms for immediate updates rather than verified reporting.<sup>[17]</sup>

During a time of crisis, false rumors invoke fear, anger, humiliation, revenge, pride or moral outrage, which tends to change the preferences of audiences from neutral content to more emotional content. In politically sensitive settings, identity threat and group grievances tend to be high and as seen during the cross-border tensions with India-Pakistan, this dynamic was intensified by pre-existing nationalist narratives, sectarian tensions and symbolic issues such as military honor, sovereignty and patriotism.<sup>[18]</sup>

Moreover, disinformation thriving during this time is due to platforms' engagement laden design. Major platforms are not neutral and they intend to rank and show algorithmic recommendations systems based on engagements, predicted relevance, watch time, reposts and social endorsement. Recent research shows that algorithmic recommender systems can amplify low-credibility or ideologically aligned content.<sup>[19]</sup> Both the United Nations<sup>[20]</sup> and Organisation for Economic Co-operation and Development (OECD)<sup>[21]</sup> have warned that today's digital information environment is shaped by opaque algorithms and advertising-driven business models that can reward polarizing, manipulative or hateful content. During crises, when user attention spikes, these incentives become even stronger.

Additionally, actor intentions during a crisis are heightened to showcase strength. Hostile social manipulation during a crisis is purposeful and linked to systematic production and dissemination of harmful information to shape beliefs and behavior.<sup>[22]</sup> The Oxford Internet Institute's research on computational propaganda shows that bots and coordinated manipulation are especially likely to appear during international crises.<sup>[23]</sup>

<sup>[17]</sup> <https://arxiv.org/abs/>

<sup>[18]</sup> [https://www.researchgate.net/publication/400092824\\_Moral\\_Outrage\\_Networks\\_The\\_Sociology\\_of\\_Digital\\_Anger](https://www.researchgate.net/publication/400092824_Moral_Outrage_Networks_The_Sociology_of_Digital_Anger)

<sup>[19]</sup> <https://www.nytimes.com/2022/10/13/technology/misinformation-integrity-institute-report.html>

<sup>[20]</sup> <https://news.un.org/en/story/2026/04/1167405>

<sup>[21]</sup> [https://www.oecd.org/en/publications/facts-not-fakes-tackling-disinformation-strengthening-information-integrity\\_d909ff7a-en.html](https://www.oecd.org/en/publications/facts-not-fakes-tackling-disinformation-strengthening-information-integrity_d909ff7a-en.html)

<sup>[22]</sup> [https://www.rand.org/content/dam/rand/pubs/research\\_reports/RR2700/RR2714/RAND\\_RR2714.pdf](https://www.rand.org/content/dam/rand/pubs/research_reports/RR2700/RR2714/RAND_RR2714.pdf)

<sup>[23]</sup> <https://www.oii.ox.ac.uk/research/projects/computational-propaganda/>

In practical terms, crises give state actors reasons to project strength, deny losses, claim victories, rally domestic publics, shape foreign audiences and obscure civilian harm. They also give partisan influencers, sensationalist media outlets and engagement-seeking creators reasons to monetize attention through spectacle, exclusives and misinformation framed as patriotic urgency.<sup>[24]</sup>

During times of crisis, narratives do not stay on one platform or within one jurisdiction. Research on encrypted messaging apps shows that misinformation often moves between Telegram, WhatsApp and YouTube, with closed or semi-private spaces functioning as early distribution layers before narratives move to larger public platforms.<sup>[25]</sup> In South Asia, the relevance of this is magnified by long-standing cross-border influence efforts and information networks documented by EU DisinfoLab, which found an extensive anti-Pakistan influence operation spanning fake media outlets, NGOs and republishing chains.<sup>[26]</sup>

The May 2025 crisis provided an interesting case because the information war was not confined to isolated false posts. Major media analyses and fact-checks documented how recycled footage from Gaza, Ukraine and video games, unverifiable battlefield claims, manipulated images and AI-generated visuals circulated across both Indian and Pakistani social media platforms. The core problem, therefore, is not simply whether false content appeared, but how it moved across platforms, who amplified it, which narratives gained traction and why existing moderation and response systems did not prevent it from shaping public perception in real time.

<sup>[24]</sup> <https://misinforeview.hks.harvard.edu/article/taking-the-power-back-how-diaspora-community-organizations-are-fighting-misinformation-spread-on-encrypted-messaging-apps/>

<sup>[25]</sup> [https://issi.org.pk/wp-content/uploads/2025/05/IB\\_Muskan\\_Moazzam\\_May\\_10\\_2025.pdf](https://issi.org.pk/wp-content/uploads/2025/05/IB_Muskan_Moazzam_May_10_2025.pdf)

<sup>[26]</sup> [https://www.rand.org/content/dam/rand/pubs/research\\_reports/RR2700/RR2713/RAND\\_RR2713.pdf](https://www.rand.org/content/dam/rand/pubs/research_reports/RR2700/RR2713/RAND_RR2713.pdf)

# Literature Review

Digital spaces have received significant attention in recent literature on the surge of disinformation within conflict-settings. The rapid consolidation of new media outlets and social media platforms has transformed how state and non-state actors engage with, and deploy, manipulative information as a weapon of war<sup>[27]</sup>. While International Humanitarian Law (IHL) has typically remained permissive on the use of information manipulation as long as it satisfies principles of necessity and proportionality, serious questions have been raised regarding the rights implications of new forms of disinformation that proliferate on social media outlets. New literature has underscored the dangers present in the extensive use of Generative AI, mass disinformation campaigns, and hate speech that targets minority citizens located in active conflict zones<sup>[28]</sup>. In many cases, such disinformation is diffused and proliferated by civilian and non-state actors through coordinated troll networks<sup>[29]</sup> or the propagation of uncritically embraced propaganda narratives<sup>[30]</sup>. This blurs the distinction between civilian and state uses of information warfare, and often deliberately distorts situational awareness to incite panic decisions and hate speech.

This literature review will focus on the spread of online disinformation on new social media platforms and identify the specific gaps in existing literature that we intend to intervene on through our analysis of the 2025 Indo-Pak Conflict. The study's focus will be on research on the use of information warfare in contemporary conflict settings, the pre-conflict context of disinformation in South Asia, and the shifts in platform policies towards content moderation as a response to the intensified geopolitical fragmentation characterizing at present.

## Global Case Studies on Information Warfare in Conflict Settings

In this section, we examine literature on emerging tactics and forms of online disinformation across three major conflicts that garnered significant media attention over the last five years. The case studies consist of literature on the ongoing Russia-Ukraine War, the Israeli-led genocide against Palestinians in Gaza, the Twelve-Day War between Iran and Israel in June 2025, as well as more recent events including the Iran-US-Israel war of 2026 and ongoing tensions between Pakistan and Afghanistan.

<sup>[27]</sup> <https://geneva-academy.ch/wp-content/uploads/2025/09/Digital-disinformation-operations-in-Armed-Conflict-1.pdf>

<sup>[28]</sup> <https://www.cigionline.org/static/documents/Pauwels-Nov2024.pdf>

<sup>[29]</sup> <https://www.disinfo.eu/publications/indian-chronicles-deep-dive-into-a-15-year-operation-targeting-the-eu-and-un-to-serve-indian-interests/>

<sup>[30]</sup> Analyzing the State of Digital Information Warfare Between India and Pakistan on Twittersphere (<https://journals.sagepub.com/doi/10.1177/21582440211031905>)

# Russia-Ukraine War (2022-)

The Russian invasion of Ukraine saw the digital space transforming into a distinct battleground within which the two states pursued distinct military objectives. This involved shaping public and international perception to their benefit, deliberately proliferating fake news meant to incite panic within civilian populations close to active battlefields, and undermining support for local state institutions through fabricated propaganda that included short-form videos and AI-generated content.

Existing research has almost exclusively focused on Russian-led campaigns of disinformation, underscoring collaborative efforts by state and non-state actors in producing propaganda efforts during the Russian invasion<sup>[31]</sup>. Campaigns promoting pro-Russian sentiments involved the use of deputized bot armies and non-state networks that used terms like “de-nazification” and “Ukrainian corruption” that served to delegitimize the Ukrainian government in international opinion. Pro-Russian disinformation has involved demonizing the European Union by using selective clips of Ukrainian refugees being abused by their European patrons<sup>[32]</sup>, promoting fake stories meant to discredit refugees as terrorists<sup>[33]</sup>, and focusing on the Western-led instigation that prompted Russia’s invasion<sup>[34]</sup>.

In the early phases of the war, Russian forces made extensive use of popular social media platforms like TikTok and Telegram to promote war spectacles and fake stories that served to undermine civilian trust in the Ukrainian government<sup>[35]</sup>. Short-form videos on TikTok showed Russian soldiers executing successful operations in Eastern Ukraine. These were sensationalized to distort the truth regarding Russian gains and severely impacted civilians looking to evacuate from neighboring war zones.<sup>[36]</sup> On Telegram, which is the most popular application in Ukraine for news consumption, Russian forces utilized propaganda discouraging civilians from evacuating conquered areas while spreading incendiary news regarding future conflicts.<sup>[37]</sup>

<sup>[31]</sup> [https://civiliansinconflict.org/wp-content/uploads/2023/11/CthatlVIC\\_Disinformation\\_Report.pdf](https://civiliansinconflict.org/wp-content/uploads/2023/11/CthatlVIC_Disinformation_Report.pdf)

<sup>[32]</sup> <https://www.ifes.org/pub-rccurhc-ch-3>

<sup>[33]</sup> [https://www.isdgglobal.org/digital\\_dispatches/russia-aligned-campaigns-amplify-negative-sentiment-towards-ukrainians-in-poland-ahead-of-a-decisive-presidential-vote/](https://www.isdgglobal.org/digital_dispatches/russia-aligned-campaigns-amplify-negative-sentiment-towards-ukrainians-in-poland-ahead-of-a-decisive-presidential-vote/)

<sup>[34]</sup> <https://euvsdisinfo.eu/report/monstrous-western-propaganda-has-led-to-ukraine-conflict/>

<sup>[35]</sup> <https://www.reuters.com/world/europe/russia-vs-ukraine-biggest-war-fake-news-era-2024-07-31/>

<sup>[36]</sup> [https://real.mtak.hu/182442/1/02\\_baptista-et.al\\_27-48\\_KOME\\_2023\\_02.pdf](https://real.mtak.hu/182442/1/02_baptista-et.al_27-48_KOME_2023_02.pdf)

<sup>[37]</sup> [https://civiliansinconflict.org/wp-content/uploads/2023/11/CIVIC\\_Disinformation\\_Report.pdf](https://civiliansinconflict.org/wp-content/uploads/2023/11/CIVIC_Disinformation_Report.pdf)

# Israel-Hamas War (2023-)<sup>[38]</sup>

Following the October 7 attacks by Hamas, the Israeli state launched a campaign of collective punishment in the Gaza Strip in what has been termed a “genocide of the Palestinian people<sup>[39]</sup>.” In order to shore up support for its campaign, the Israeli state made use of an extensive and organized campaign of online disinformation that lent legitimacy to its aggression. The Israeli state and intelligence have made active and organized efforts to shape global public opinion, particularly in Western countries, in their favor. This has involved deliberate use of AI-based fake news, hate speech, and online disinformation promoted through paid advertisements, traditional media outlets, and social media bot armies<sup>[40]</sup>. The proliferation of disinformation in a conflict defined by asymmetric capabilities and mass aggression has led some to term Israel’s use of information warfare as a form of “infocide”- a sustained policy of distorting and denying objective reality and facts.

Israel’s disinformation strategy was multifaceted, involving paid advertisements that were meant to vilify U.N relief agencies operating in Palestinian territories<sup>[41]</sup>, propagating myths about the extent of violence committed by Hamas in the October 7 attacks - including the infamous “Hamas killed 40 babies” narrative that was regurgitated by then U.S President Joe Biden - and by consistently downplaying the level of civilian casualties in Gaza<sup>[42]</sup>. When dealing with civilian casualties, disinformation involved first denying any responsibility for a particular incident. An example of this was the Al-Shifa Hospital bombings, an incident that generated significant outcry and stood as evidence of the state’s willingness to commit war crimes as part of a doctrine of collective punishment. Social media disinformation, fuelled by AI-based footage and the dispersion of state narratives, first alleged that these bombings were a consequence of a Hamas missile misfiring. When this was debunked, new narratives propped up alleging that Hamas and other Palestinian militant organizations use public infrastructure as bases of operation- thereby rendering them active military targets for the Israeli military.<sup>[43]</sup> Israel’s sophisticated techniques of information warfare were countered effectively by on-ground journalists in Gaza, hundreds of whom were killed in Israeli airstrikes.

<sup>[38]</sup> While a declaration of ceasefire was signed in October 2025, Israel continues to violate the agreement with almost daily attacks against the region.

<sup>[39]</sup> <https://www.ohchr.org/en/press-releases/2025/09/israel-has-committed-genocide-gaza-strip-un-commission-finds>

<sup>[40]</sup> <https://www.iemed.org/publication/gaza-infocide-disinformation-and-new-narratives-wars/>

<sup>[41]</sup> <https://www.unrwa.org/newsroom/official-statements/state-israel-continues-dis-information-campaign-against-unrwa>

<sup>[42]</sup> <https://www.iemed.org/publication/gaza-infocide-disinformation-and-new-narratives-wars/>

<sup>[43]</sup> <https://journals.akademicka.pl/politeja/article/view/6387/5944>

# Israel-Iran War (2025)

As mentioned above, Israel's central role in conflict across the Middle East has taken center stage in social media discourse. The Israel-Hamas war was an instance of "infocide", in which asymmetric military capabilities were reflected in the sphere of "information warfare," where Israeli narratives were explicitly manufactured to legitimize its military operation amidst a climate of global scrutiny. The "Twelve Day" war between Israel and Iran in June 2025 played out rather differently, with both states deploying sophisticated tactics of disinformation and hate speech designed to shape public opinion and claim military victories. In many ways, this conflict, which occurred just over a month after the skirmishes between India and Pakistan, reflects a new form of political and information warfare where new technologies are leveraged to garner public sympathy within an extremely polarized and hyper-partisan political environment.

During this conflict, disinformation drew upon existing repertoires of resentment and political stereotypes. Israel-linked accounts presented the country as a bastion of Western civilization protecting Europe and America from the "barbaric horde" of Islamists in the Middle East<sup>[44]</sup>. The Israeli state made extensive use of paid advertisements on social media to promote AI-generated content<sup>[45]</sup> presenting dystopian scenarios of Europe drowning into a wasteland because of a "rogue" Iranian government. Drawing upon global contentions over Iran's alleged development of nuclear weapons, social media disinformation fuelled anxieties that served to legitimize Israeli-American intervention in Iran. Government-affiliated accounts and journalists linked to Israel and Western mainstream outlets also portrayed false videos and images of "mass protests" in Iran against the clerical government. Provided video evidence was eventually linked to older videos of traffic jams, mass gatherings, and in many instances, AI was used to generate false clips of Iranians celebrating Israeli strikes<sup>[46]</sup>.

The scale at which AI-generated images and clips proliferated multiple social media platforms over the course of this conflict made it exceptionally difficult for journalists and fact-checkers to keep up. Ineffective platform policies with regard to content moderation, as discussed in the final section of this literature review, made it even more difficult to undermine the adverse effects of this disinformation war. This involved both sides showing AI-generated footage of successful air and missile strikes in the opposing country. One such fabricated clip, that showed Iranian air defence systems taking down an Israeli jet, was re-uploaded by a government-linked Iranian account as well<sup>[47]</sup>. The Iranian government and its affiliated supporters on social media drew upon cultural and religious symbols, positioning the state as a victim of imperialist and genocidal aggression perpetrated by Israel and the United States, with the intent of projecting Iran as a pariah state in public consciousness<sup>[48]</sup>.

<sup>[44]</sup> <https://www.dohainstitute.org/en/Lists/ACRPS-PDFDocumentLibrary/iran-vs-israel-how-deception-and-the-information-war-shaped-public-perceptions.pdf>

<sup>[45]</sup> <https://ideas.repec.org/a/edt/jsserr/v12y2025i1p395-399.html>

<sup>[46]</sup> <https://www.bbc.com/news/articles/c0k78715enxo>

<sup>[47]</sup> <https://ideas.repec.org/a/edt/jsserr/v12y2025i1p395-399.html>

<sup>[48]</sup> <https://www.fairobserver.com/world-news/us-news/the-lies-the-united-states-and-media-told-about-iran/>

## Iran-US-Israel War (2025–26)

The recent escalation involving Iran, the United States and Israel further illustrates how contemporary conflicts are increasingly fought across both military and digital domains. Similar to the India-Pakistan crisis, competing state narratives, unverified battlefield claims and AI-generated or manipulated media circulated rapidly across social media platforms, making independent verification increasingly difficult. Researchers during this time have documented the widespread use of synthetic media, recycled footage and contradictory official statements to shape public perception and claim battlefield successes. These developments demonstrate how generative AI, engagement-driven platforms and information vacuums have become central features of modern conflict, allowing disinformation to spread faster than traditional verification mechanisms.<sup>[49]</sup>

## Pakistan-Afghanistan Border Tensions (2025)

Recent Pakistan-Afghanistan border tensions similarly demonstrate that information warfare has become a recurring feature of regional conflicts in South Asia. Alongside military exchanges, social media platforms were flooded with misleading claims, recycled videos and fabricated footage purporting to show attacks by both sides. Fact-checking organizations have repeatedly identified viral content that reuses footage from unrelated conflicts or misrepresents older videos as current military operations. These repeated disinformation tactics have continued in 2026, demonstrating how rapidly false narratives can circulate during periods of heightened political tension, reinforcing the need to examine disinformation as an integral component of contemporary interstate conflict, rather than an isolated phenomenon confined to the India-Pakistan context.<sup>[50]</sup>

## Gaps in Existing Scholarship

Existing scholarship and policy debates on digital regulations have only recently caught up with the sophisticated techniques and discourses deployed in these spaces during conflict settings. The studies used in this literature review provide us with critical insights into these evolving techniques and how they problematize existing notions of information warfare as a legitimate weapon of war. In contrast to this perceived legitimacy, we find overwhelming evidence of how existing forms of disinformation propagate hostile narratives, incite panic in civilian populations, and distort crimes of mass aggression to mitigate global criticism. The brief conflict between India and Pakistan in May 2025 provides us with a unique case of how digital warfare operates in parallel to structurally determined tensions and hostilities. Our analysis focuses on how online disinformation during this conflict played out differently across social media platforms, how the nature and source of disinformation implicated multiple actors

<sup>[49]</sup> <https://digitalrightsfoundation.pk/wp-content/uploads/2026/05/Disinformation-in-Warfare-in-the-age-of-AI-and-Synthetic-Media.pdf>

<sup>[50]</sup> <https://digitalrightsfoundation.pk/wp-content/uploads/2026/05/Disinformation-in-Warfare-in-the-age-of-AI-and-Synthetic-Media.pdf>

ranging from media conglomerates to private users, and on the pace at which disinformation-fuelled discourses were consolidated in the aftermath of the conflict, leading to future conflicts following the same rules.

## **Platform Accountability: How Social Media Platforms Address Disinformation**

Amidst the proliferation of disinformation in conflict settings, social media platforms are compelled to tread a tightrope, managing freedom of expression with various forms of regulation against hate speech and misinformation. Existing literature has highlighted their failure to balance these commitments properly.

During the so-called Israel-Hamas conflict, platforms such as Meta removed hundreds of thousands of pro-Palestine posts under the pretence that such content was “glorifying terrorist organizations” designated as such by Meta’s community guidelines<sup>[51]</sup>. The Human Rights Watch reported on how Meta has made broad use of its DOI (Dangerous Organizations and Individuals Policy) – a policy which is not made public – to apply sweeping restrictions on over a thousand social media accounts providing authentic updates on the unfolding humanitarian catastrophe in Gaza. Previously, Meta recognized its failure to uphold human rights due diligence when regulating conflict-specific content on their applications, especially when concerning advocacy for Palestine<sup>[52]</sup>. However, despite the numerous commitments made, it has failed to fully align policies with its human rights responsibilities.

In early 2025, Meta revamped its content moderation policy in line with X’s policy that allows hate speech disguised as disinformation to remain on social media platforms to encourage “free speech”.<sup>[53]</sup> X’s approach, which prioritizes community fact-checking and personalized content, has been criticized severely due to the platform’s key role as an arena of political discourse. The ineffectiveness of this policy was made clear during the Indo-Pak conflict, where the removal of administrative measures regulating hate speech led to an intense proliferation of hate speech against Indian Muslims, Kashmiris, and minorities in Pakistan<sup>[54]</sup>. Amnesty has criticized Meta’s new raft of changes in content moderation as evidence of partiality to U.S political imperatives, a criticism that was levelled against the platform for its role in denigrating Rohingya Muslims in Myanmar and promoting paid political ads by Indian parties on Facebook<sup>[55]</sup>.

During the Indo-Pak Conflict, accusations of state collusion and platform incentives became even more evident - especially in the case of X. It is widely believed that content moderation policies, whether it be Meta’s DOI which recently introduced the term “shahada” (Islamic term for martyr) in its definition of dangerous speech or TikTok’s automated removal mechanisms

<sup>[51]</sup> <https://www.hrw.org/news/2023/12/20/meta-systemic-censorship-palestine-content>

<sup>[52]</sup> <https://www.bsr.org/en/reports/meta-human-rights-israel-palestine>

<sup>[53]</sup> <https://dig.watch/updates/metass-eased-content-moderation-and-governance>

<sup>[54]</sup> [https://www.instagram.com/p/DI-B\\_yQo66i/?img\\_index=1](https://www.instagram.com/p/DI-B_yQo66i/?img_index=1)

<sup>[55]</sup> <https://www.amnesty.org/en/latest/news/2025/02/meta-new-policy-changes/>

that have disproportionately targeted authentic, if graphic, portrayals of the horrors of Israeli atrocities in Gaza<sup>[56]</sup>, are aligned with foreign policy imperatives of powerful countries. These imperatives, driven by financial incentives that concern retaining their share in large populations, were made evident in X's decision to geoblock thousands of Pakistan-based accounts in India<sup>[57]</sup>. These claims were framed as state compulsion, but the large domestic market provided by Indian users was also used as a determining factor in complying with the Indian government dictates.

In sum, it is clear that existing platform regulations are both insufficient in form and lack enforcement in content. Policies have largely pivoted towards creating an enabling environment for hate speech and misinformation, and this proves immensely problematic in conflict situations characterized by asymmetric capabilities and a crackdown on minority voices. Secondly, enforcement of policies remains largely lackluster, and policy space is regularly ceded to political actors that leverage their large domestic markets to force compliance.

<sup>[56]</sup> <https://researchcentre.trtworld.com/wp-content/uploads/2023/10/Israel-War-on-GazaV2.pdf>

<sup>[57]</sup> <https://asia.nikkei.com/politics/x-accounts-news-sites-blocked-in-india-following-pakistan-clash>

# Methodology

This study employs a mixed-methods digital research approach to examine the production, circulation, amplification and moderation of disinformation during the May 2025 India-Pakistan crisis. The research combines quantitative analysis of platform activity with qualitative content analysis to understand how narratives evolved across different digital environments and how platform architectures shaped their visibility and reach.

## Research Objectives

This study seeks to examine how disinformation emerged, evolved and circulated across digital platforms during the May 2025 India-Pakistan crisis. Rather than focusing solely on individual instances of false information, the research investigates the broader information ecosystem that enabled misleading narratives, manipulated media, rumors and synthetic content to gain visibility during a period of heightened geopolitical tensions.

The objective of the study is to map the volume, format and timing of disinformation throughout the crisis, identifying key moments when misleading content surged and assessing how these spikes corresponded with major military, diplomatic and political developments. Particular attention has been paid to the circulation of recycled footage, manipulated images, AI-generated content, unverifiable battlefield claims and emotionally charged narratives. Moreover, the study aims to compare the dynamics of disinformation across major digital platforms, including X, Facebook, BlueSky, YouTube and TikTok. The study recognizes that platforms differ significantly in their technical architecture, recommendation systems, visibility mechanisms and moderation practices, all of which shape how information spreads and gains legitimacy. Understanding these differences is critical to assessing platform-specific vulnerabilities during periods of crisis.

Additionally, the research identifies the key actors and narratives involved in the production and amplification of disinformation. This includes examining the role of state-linked actors, political influencers, media organizations, anonymous networks, coordinated amplification accounts and ordinary users.

The study intends to see the effectiveness of platform moderation and state responses during the crisis. This includes examining content spreading disinformation that is still present over platforms, account restrictions, fact-checking interventions, labeling mechanisms and other measures implemented by platforms, alongside official rebuttals, legal restrictions, content-blocking measures and public communication strategies deployed by state authorities. The research identifies accountability gaps within the information ecosystem. By tracing how misleading content moved from creation to amplification and public visibility, the research assesses where existing governance mechanisms failed and develop recommendations for platforms, policymakers, media organizations, fact-checkers and civil society actors seeking to strengthen information integrity during future crises.

# Data Sources and Collection

Data for this study was collected from five major social media platforms, which were X, Facebook, Instagram, YouTube, TikTok, and Bluesky. The collection period spans from May 2025, beginning with the escalation following the Pahalgam attack and subsequent military exchanges, through December 2025, to capture both the immediate and the aftermath of disinformation narratives that persisted after the ceasefire.

Data collection was conducted through a combination of platform-specific scraping and archiving methods. DRF's research team utilized APIs and other digital collection tools to extract publicly available posts, videos, comments, hashtags, engagement metrics, account information and related metadata. The collected data was archived and organized using Google Sheets and Excel datasets to facilitate collaborative coding and analysis. Given the cross-border nature of the crisis, the project was conducted in collaboration with researchers across the globe who contributed to data extraction.

Following the collection, the dataset underwent a structured coding process. The research team annotated content according to predefined categories, including misinformation, disinformation, manipulated media, AI-generated content, hate speech, nationalist narratives, military claims, casualty-related claims, and state-linked messaging. The coding also captured platform-specific variables such as content format, engagement levels and account type. This annotation process enabled the identification of recurring themes, emerging narratives and patterns of amplification across platforms.

## Analytical Framework

The analysis is structured around two interconnected levels.

First, a temporal analysis was conducted to examine fluctuations in the volume of content over time. Using time-series analysis, the study mapped spikes in disinformation activity against key offline events, including military operations, official statements, diplomatic developments and ceasefire announcements. This approach enabled the identification of moments when misinformation and manipulated content circulated most intensely.

Second, the study conducted keyword and hashtag analysis to identify dominant narratives and track how specific themes evolved throughout the crisis. Frequently used keywords, slogans, hashtags and framing devices were analyzed to understand how users, influencers, media organizations and political actors constructed and amplified particular interpretations of events.

Additionally, a cross-platform comparative analysis was conducted to assess how platform design, recommendation systems, moderation mechanisms and content governance practices influenced the spread and persistence of disinformation. This framework enabled

the study to identify platform-specific vulnerabilities, moderation gaps and patterns of narrative migration across X, Facebook, Instagram, YouTube, TikTok, and Bluesky. Data analysis and visualization were conducted using Google Colab, spreadsheet-based coding systems and network visualization techniques to identify trends, relationships and patterns within the dataset.

## Limitations

This study seeks to provide a comprehensive analysis of disinformation during the May 2025 India-Pakistan crisis; however, several limitations should be acknowledged. First, access to platform data remains uneven across social media platforms. While public content on platforms such as X, YouTube, TikTok, Facebook, Instagram and Bluesky can be collected and analyzed to varying degrees, access to platform APIs, recommendation systems, moderation decision and internal engagement metrics remains limited. As a result, the study cannot fully assess the reach or algorithmic amplification of all content examined.

Second, the research is constrained by the visibility of publicly accessible information. Disinformation frequently circulates through private or encrypted channels, particularly on platforms such as WhatsApp and Telegram. Although the study incorporates publicly shared messages, screenshots and content from accessible platforms, it cannot comprehensively capture information flows occurring within private conversations or closed networks. Consequently, some narratives may have achieved wider circulation than reflected in the dataset.

Third, the dynamic nature of social media presents challenges for data preservation and verification. Content may be deleted, accounts suspended, visibility reduced or platform policies altered after publication. Although efforts were made to archive content systematically during the collection period, some material may no longer be accessible, limiting verification. Fourth, identifying coordinated behavior, state-linked influence operations, bot activity and organized amplification networks presents methodological challenges. While network analysis and qualitative assessment can reveal patterns of coordination, publicly available data does not always permit definitive attribution of intent, organizational affiliation or state involvement. Findings relating to coordinated activity should therefore be interpreted as indicative rather than conclusive.

Fifth, the study focuses primarily on English and Urdu-language content and may not fully capture narratives circulating in regional languages or within localized digital communities. While collaboration with regional researchers helped contextualize emerging narratives and cross-border information flows, linguistic and cultural variations may still influence the visibility and interpretation of certain forms of disinformation.

Finally, this study does not seek to establish the prevalence of disinformation across all digital platforms or measure its direct impact on public opinion. Rather, it aims to identify patterns,

actors, narratives and governance challenges that emerged during the crisis. The findings should therefore be understood as an analysis of observable information dynamics and platform responses rather than a comprehensive account of all disinformation activity related to the conflict.

Despite these limitations, the study provides a valuable cross-platform examination of how disinformation evolved during a period of geopolitical crisis and offers important insights into the strengths and weaknesses of current platform governance and accountability mechanisms.

# Findings

## Platform-Specific Analysis

### X (Formerly Twitter)

X saw some of the largest volumes of disinformation throughout and after the conflict, both through text posts, and pictures/videos, many of which were AI-generated disinformation. Widespread celebration of Operation Bunyaan ul Marsoos was observed among content on X from Pakistani sources, with the conflict decisively framed as an overwhelming victory for Pakistan, which was presented as a victim of Indian aggression. Numerous clips of “Yom e Tashakkur<sup>[58]</sup>” celebrations circulated, along with footage of the Pakistani army visiting universities across the country, particularly in Khyber Pakhtunkhwa and Balochistan, speaking on Pakistan’s military prowess.



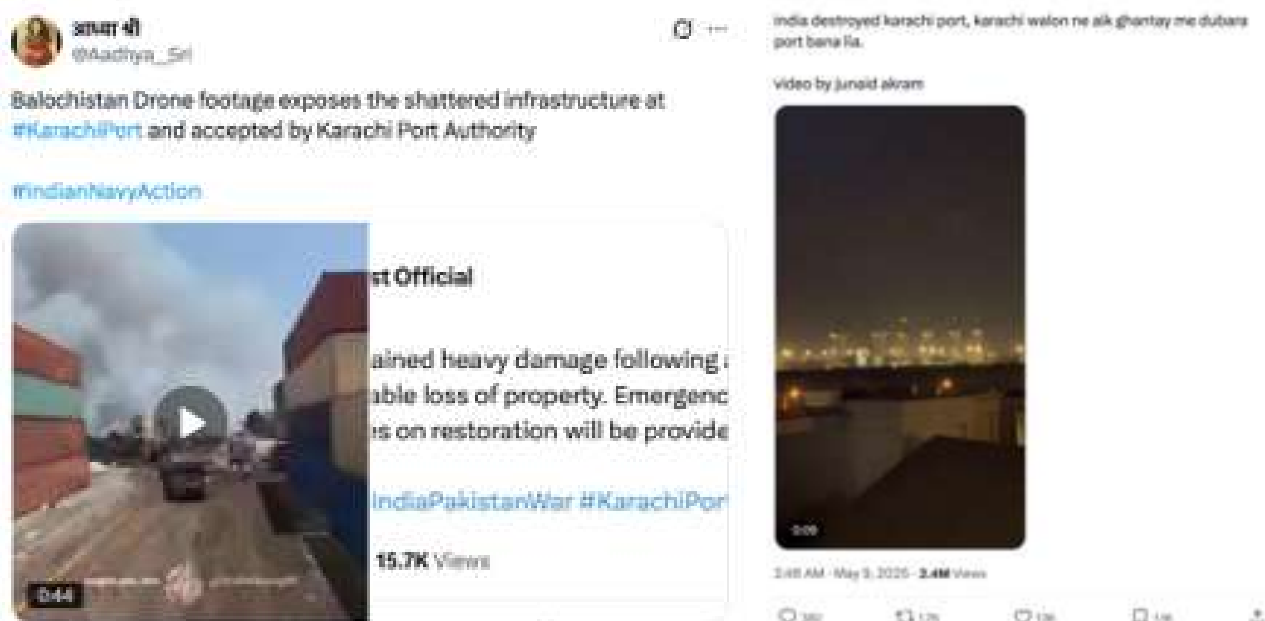
There was misinformation alleging no loss of civilian lives in Pakistan’s attack and blanket glorification of the Pakistani military. Pakistan’s military response was framed as a religious battle against infidels. While there was no overt Hinduphobia, the narrative around the conflict was heavily inspired by religious symbolism, and the military was portrayed as a guardian of Islam. Instances of hate speech against Indians were embellished using AI. There was also a great deal of gender-based misinformation and hate speech, including the use of phrases

<sup>[58]</sup> Day of Gratitude observed annually on 16 May to commemorate Pakistan’s victory in the 2025 India-Pakistan conflict

like “maal e ghanimat<sup>[59]</sup>” directed at Indian women, as well as posts showing foreign female soldiers in similar terms. Acting on widespread rumours of a captured Indian female pilot, there was a significant amount of fake AI-generated content sexualising Indian women.



Among Indian misinformation and malinformation, fake sensationalised news circulated reporting the annihilation of numerous sites in Pakistan, including air defence systems and the Karachi Port, with such fabricated reports gaining traction as users relied on mainstream outlets.



<sup>[59]</sup> Spoils of war in Islamic jurisprudence.

Clips presented as evidence of this destruction were often taken from other conflicts or were completely fabricated. Hate speech included the celebration of the suspension of the Indus Waters Treaty and the anticipation of natural disasters in Pakistan resulting from this decision. Pakistani people were described using terms such as beggars and intolerant savages, alongside insults directed at Muslims and the religion of Islam, and fearmongering narratives about Pakistanis intending to take resources from India.

After getting bashed for 3-4 days

Pakistan launched Operation Buryan Al Marsoos today, and achieved absolutely nothing.

Yet, the beggars are calling it a "conclusive victory." How shameless. 🤡



11:35 PM · May 10, 2025 · 246.5K Views

548 60K 3.6K 54

I have always believed that every Pakistani is a terrorist.

Look at this video where Pakistani civilians are telling Pakistani soldiers that after war, under Maal-e-Ghanimat, all Indian girls will be ours, Taj Mahal is ours... everything is ours.

These people are animals, total savages. All Pakistanis are Jihadis and cannot be trusted and they deserve no mercy.



4:18 PM · May 27, 2025 · 911.1K Views

782 0.6K 15K 1.4K View quotes >

So there was a space where Pakistani men spoke about converting all Hindu women to Islam and will "accept" them in the most sexual and scary way.

Not a single 🇮🇳 girl on that space raised her voice and asked them shut their fucking mouth.

Rape culture starts from here.

4:56 PM · May 11, 2025 · 558.8K Views

700 0.2K 29K 1.4K View quotes >



Hyper militarism and triumphalism were prevalent, fuelled by fabricated reports such as Balochistan declaring independence, nuclear radiation leaks, threats to bomb Lahore, and claims about conquering Pakistan by force. Hate mongering was also directed at Indian Muslims, with those opposing Sindoor or questioning the state's position on Pehalgam labelled as enemies of the state and Muslim sympathisers. Individuals such as Jyoti Malhotra were accused of being ISI spies, alongside multiple fabricated claims of Indian Muslims acting on behalf of Pakistani intelligence.

Translated from Hindi. Show original

Name - Sadiq Khan

Work - Celebrating Terrorist Attacks

Dehradun resident Sadiq Khan was laughing at the deaths of Hindus who were killed after being asked their religion in Pahalagam

and was justifying the heinous acts committed by terrorists, after which Hindus gave him a taste of his own medicine and handed him over to the police station

It is reported that one person from the crowd behind Sadiq sprinkled red chili powder on him, after which Sadiq kept chanting "Pakistan Mundabad" for quite a while...

This idea is absolutely new

If someone around you does something like this too, you can change their mindset with 5 rupees worth of red chili powder 🌶️

Rate this translation: ⭐



**BIG NEWS** Man posing as PMO official held in Kerala for seeking details on INS Vikramt amidst Operation Sindoor

Mujeeb Rahman called Kochi Naval headquarters for the exact location of INS Vikramt

He introduced himself as Raghavan, a PMO official

Indian Navy denied information and alerted local Police

He was detained under new criminal laws. He is being questioned by the state police, Navy, and IB

Mujeeb's family has claimed that he has been undergoing treatment for mental health issues



Jyoti Mahotra:

- Class 12th dropped out.
- Visited Pahalagam 3 months back
- Was in touch after the attack as well
- Was in a romantic relationship (honey trapped) with Danish
- Visited Bali together (no comment on Danish's choosay hua nimbu looks)
- Was live with Danish during 'blackout' drill
- Investigators believe this was not a coincidence but to assess India's crisis readiness.

A small town YouTuber craving for early fame leaves no limits behind. Sold out to ISI for a few bucks maybe. The rest, she was dumb enough to be used by a Pakistani handler. Happens with the ppl who remain in their Instagram bubble and everything else is just "politics".



## TikTok

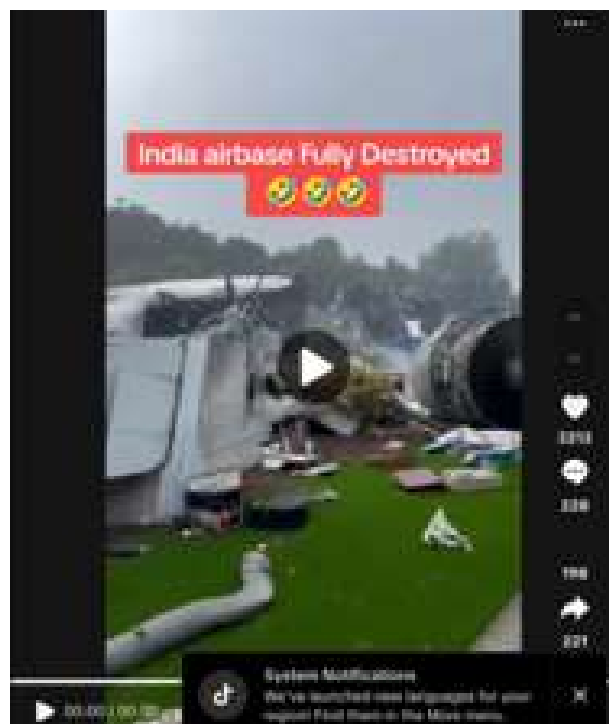
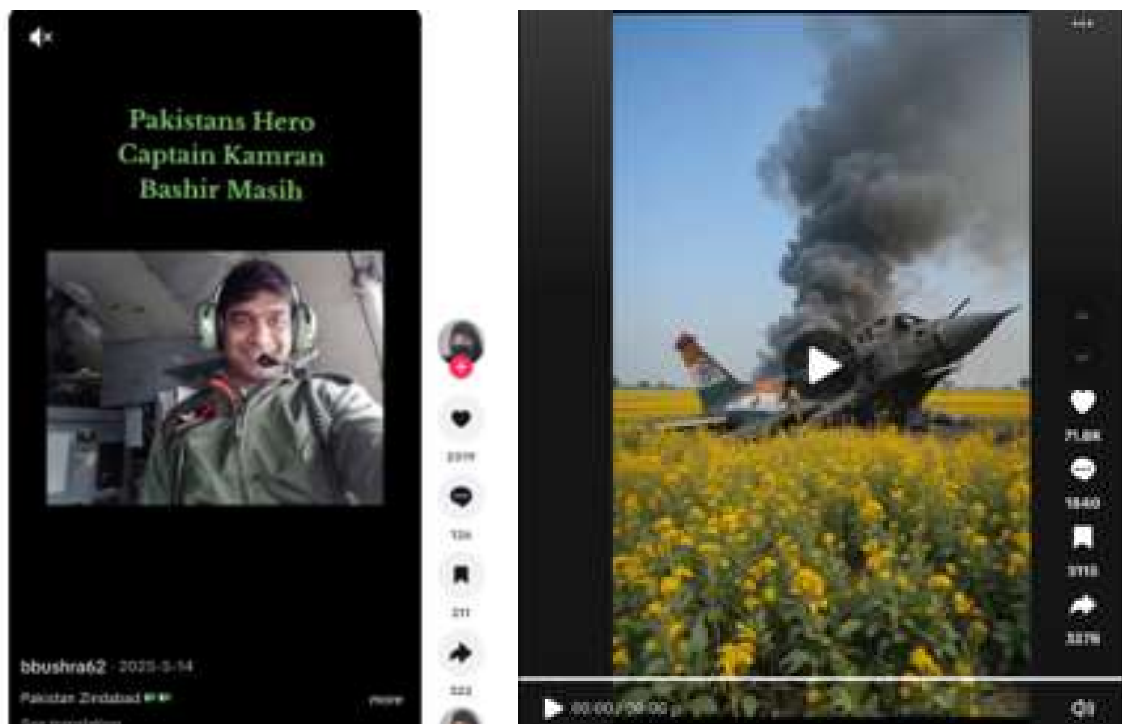
Similar to X, TikTok was one of the two platforms with the largest volume of harmful content, often surpassing the volume of X content.

The engagement numbers for May for content related to the conflict on Pakistani TikTok hit an all-time high. TikTok saw an uptick in meme-related video content spreading misinformation and disinformation.

Harmful content observed on TikTok was characterised by a mix of hate speech, misinformation, and gendered disinformation, often disguised as humour or nationalist expression. A significant portion of content relied on offensive humour stereotyping Indians, and particularly Hindus, through tropes such as drinking "cow urine" or falsely claiming Aryan identity.

There was also a proliferation of AI-generated content, particularly targeting Indian women, including public figures and military personnel such as Shivangi Singh. These videos are often deeply misogynistic and sexualised, portraying women as objects of conquest or humiliation in the context of geopolitical conflict. In some cases, such figures were depicted as "rewards" for military victory, or Muslim women as being "liberated" by opposing forces.

Misinformation was majorly seen on TikTok; viral narratives circulated about a Christian pilot in the Pakistan Air Force allegedly carrying out exaggerated or entirely fabricated military actions, such as downing advanced aircraft or infiltrating Indian airspace.



These stories appear designed to construct an image of national unity and inclusivity while simultaneously inflating military success. Similarly, unverified claims about India attacking its own territories, particularly in Punjab, were used to provoke internal divisions and incite Sikh communities.

Much of the content seen on the Pakistani side glorified the Pakistani military, often framing its actions as a moral struggle, and as a defender of Islam against perceived “infidels.” Military campaigns such as Operation Bunyan ul Marsoos were depicted as part of a larger religious crusade between good and evil. This narrative was reinforced through AI-generated videos and inflammatory language, including terms like “Operation Suhaag Raat,” which blended militarism with sexualised and derogatory undertones.



## YouTube

Another major platform that had a fairly high volume of disinformation and hate speech during the conflict was YouTube. Table 1 shows the most popular and common YouTube channels found to be taking part in posting conflict-specific content that was either categorised as false, technology-facilitated gender-based violence (TFGBV) in the form of generative-AI depictions of women from Pakistan and India, and hate speech especially against religious minorities.

*Table 1: Most popular YouTube channels posting content on India-Pakistan conflict*

| Username                           | Followers | Views    | Likes   | Engagement |
|------------------------------------|-----------|----------|---------|------------|
| WION                               | 10000000  | 45919    | 202     | 46121      |
| India Today                        | 10600000  | 116826   | 500     | 117326     |
| NewsX Live                         | 1400000   | 28258    | 155     | 28413      |
| Anurag Aggarwal:<br>Business Coach | 5340000   | 48546726 | 2800000 | 51346726   |
| Discover Pakistan                  | 898000    | 56820    | 806     | 57626      |
| TIMES NOW                          | 6740000   | 28659    | 111     | 28770      |
| Hindustan Times                    | 8400000   | 26048    | 68      | 26116      |
| Geo News                           | 19500000  | 20024    | 587     | 20611      |
| DNAIndiaNews                       | 1180000   | 7199     | 43      | 7242       |
| Business Today                     | 2900000   | 6597     | 46      | 6643       |
| Jaidev Jamwal                      | 870       | 5625     | 37      | 5662       |
| BOL News                           | 8050000   | 5336     | 199     | 5535       |
| Essay Store                        | 2670      | 1740     | 19      | 1759       |
| Syed Hasnat Ali                    | 222       | 509      | 0       | 509        |
| its umar reaction                  | 164       | 489      | 9       | 498        |
| Caca Muncher                       | 3         | 34       | 2       | 36         |
| Pakistan vs India                  | 83        | 28       | 0       | 28         |

Content on YouTube from Pakistani sources largely consisted of AI-generated clips that combined gendered disinformation with the glorification of the military. Female Indian soldiers were frequently depicted as crying victims being consoled by male pilots from the Pakistan Air Force.



operation suhag raat complete 🥰 #shortsvideo #unfrezzmyaccount #viralvideo #india #pakistan #duet

Narratives circulated about potential collaboration with China in future military incursions into India, alongside blatant misinformation about the nature of Pakistani retaliation and repeated, unverified state-linked claims regarding Indian attacks.

Generative AI content also promoted false stories about a captured Indian female pilot, often sexualising her and using tags such as “suhag raat<sup>[60]</sup>.” False quotes were attributed to Pakistan’s allies, particularly China, claiming joint military action against India.



China क्यों चाहता है India-Pakistan के बीच War? Retd. Major Gaurav Arya से सुनिए! TNNB #shorts

<sup>[60]</sup> Nuptial or wedding night

Religious framing was prominent, with references comparing the conflict to the Battle of Badr and portraying the Pakistani military as a defender of Islam against Hindu adversaries. The conflict was also used to advance reactionary narratives against domestic opposition groups, especially in Balochistan and Khyber Pakhtunkhwa.



A large volume of AI-generated battle scenarios depicted fictional engagements between Indian and Pakistani forces, including portrayals of female Pakistani pilots and PAF pilots offering a ‘captured’ Indian pilot a cup of tea.<sup>[61]</sup>

Content from Indian sources included widespread use of derogatory language such as references to Pakistan as a “beggar state,” characterisations of its population as “jahil awam<sup>[62]</sup>,” and depictions of Pakistani actors as terrorists, often amplified through AI-generated clips and mainstream media.

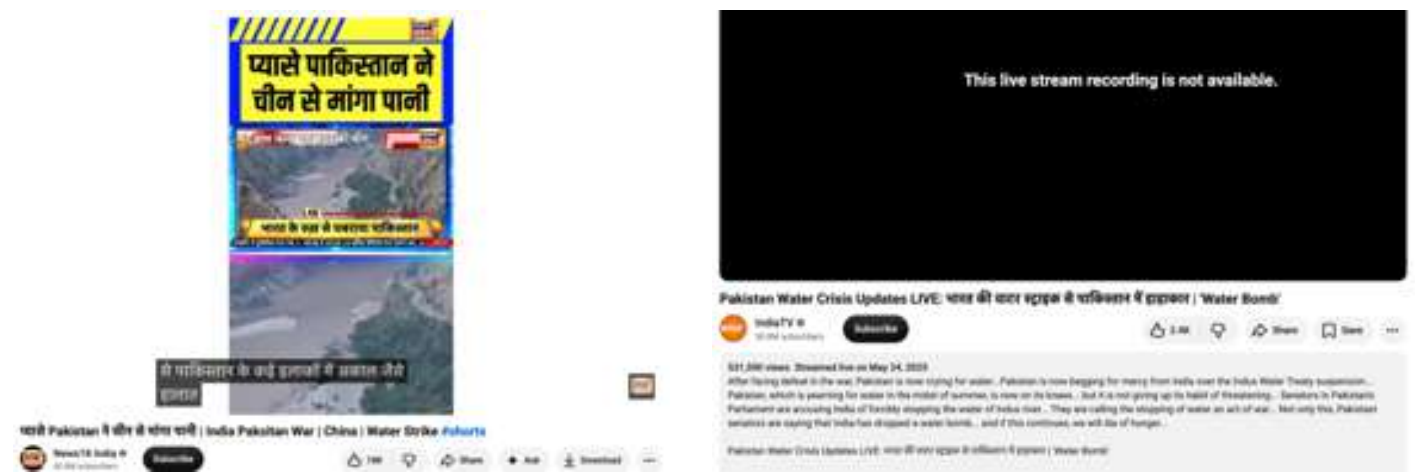


<sup>[61]</sup> This is in reference to a 2019 incident where an Indian pilot was captured by Pakistani forces and was offered a cup of tea <https://www.dawn.com/news/1466402>

<sup>[62]</sup> (Hindi/Urdu for illiterate or uncivilized population) [https://www.youtube.com/watch?v=YdVNOqtdg\\_4](https://www.youtube.com/watch?v=YdVNOqtdg_4)

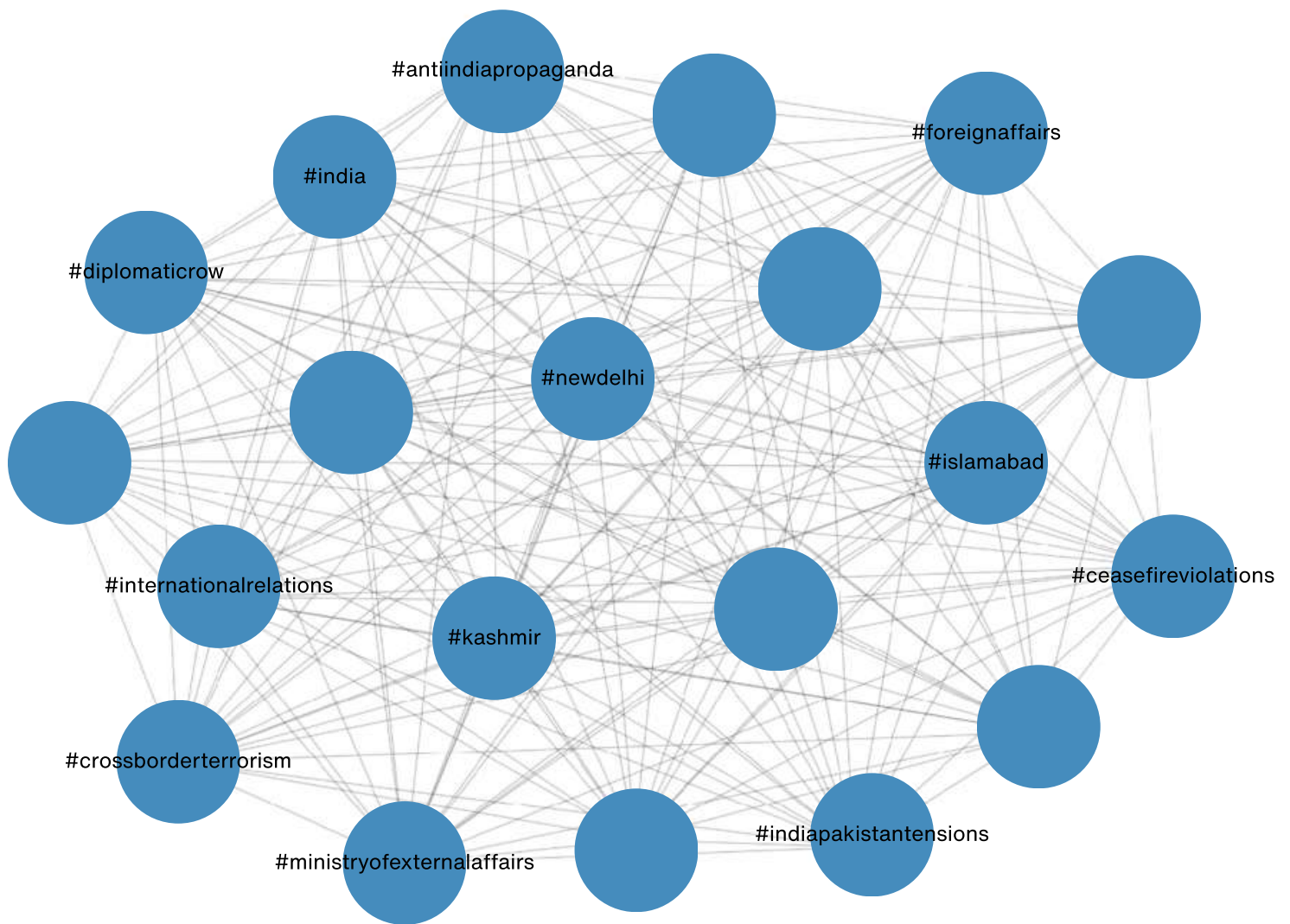
Prominent media figures and outlets such as Palki Sharma, Gaurav Arya, News18 India, Firstpost, and The Chanakya were associated with the spread of such narratives. Islamophobic sentiment was widely present across the content.

Misinformation and triumphalist narratives surrounded Indian military actions, including the glorification of anti-state militant actors within Pakistan and the portrayal of all casualties in Indian strikes as terrorists. Claims also circulated about Pakistan targeting children in India, alongside narratives conflating Pakistani state institutions with terrorism and linking them to Chinese influence. The Indus Waters Treaty crisis was frequently referenced in satirical AI-generated videos portraying Pakistani politicians as impoverished and begging for water. Additionally, a significant volume of AI-generated content presented Narendra Modi in an overtly masculine and heroic light, often framed in ways that invite adulation.



Following from this, Figure 1 presents a keyword co-occurrence network for YouTube videos published in May 2025. Nodes represent recurring keywords or hashtags used, while the edges indicate co-occurrence of multiple keywords within the same video. The network exhibits a highly dense structure, where central nodes such as ‘New Delhi’, and ‘Islamabad’ function as narrative anchors, consistently appearing across content. Ideological terms such as ‘anti-India propaganda’ are embedded within broader geopolitical and security frames, indicating intentional narrative positioning rather than ad-hoc commentary, further showing how content on the platform was consciously being made and posted with the intent to spread hate speech and disinformation during the conflict. A more comprehensive list of keywords identified from YouTube is given in Table 2.

*Figure 1: Keyword and hashtag co-occurrence network*  
*Keyword/Hashtag Co-occurrence Network (threshold  $\geq 1$ )*



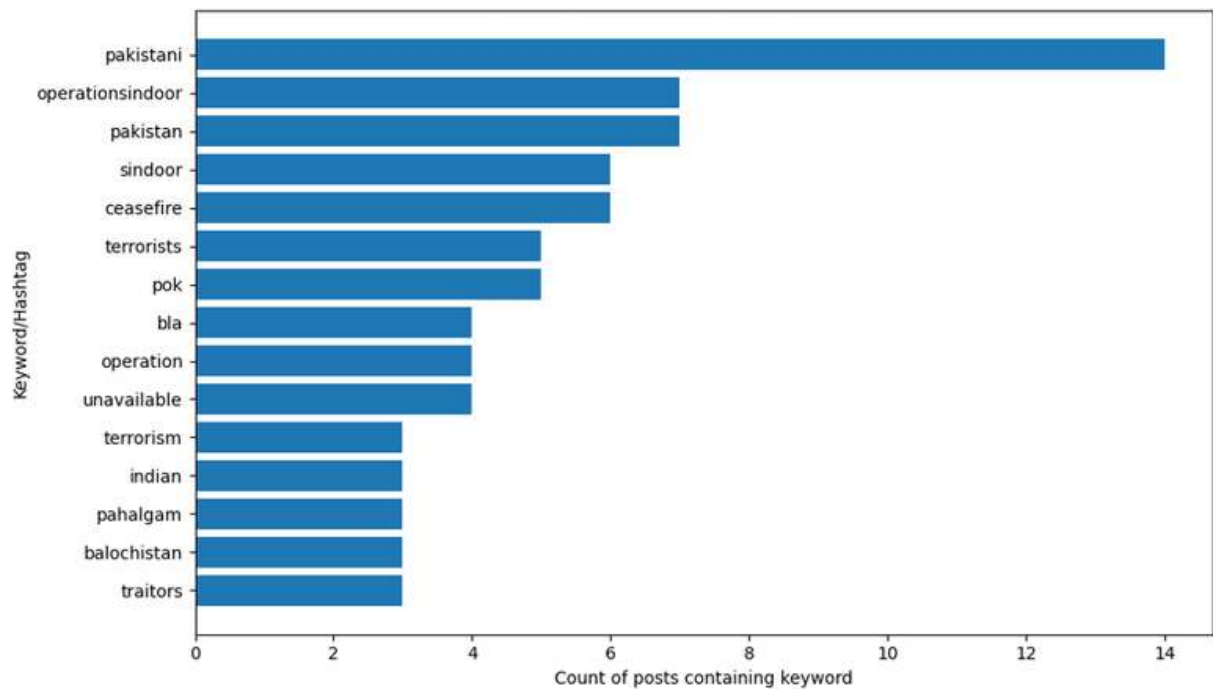
*Table 2: List of keywords from YouTube*

| List of Keywords        |
|-------------------------|
| #indiatukerytensions    |
| #boycottturkishproducts |
| #operationdost          |
| #operationsindoor       |
| #indiastrikesback       |
| #indiafirst             |
| #nurkhanairbase         |
| #pakindiaborder         |
| #indiapakistantensions  |
| #memeswar               |
| #amarnathyatra          |
| #warmovie2025           |
| #finalwarning           |
| #operationsuhagraat     |
| #nuclearalert           |
| #indiabeready           |
| #unlockthewar           |
| #realthreatfrompakistan |
| #modivspakistan         |
| #pakistanbehindpahalgam |
| #pahalgamterrorattack   |
| #crossborderterrorism   |
| #antiindiapropaganda    |

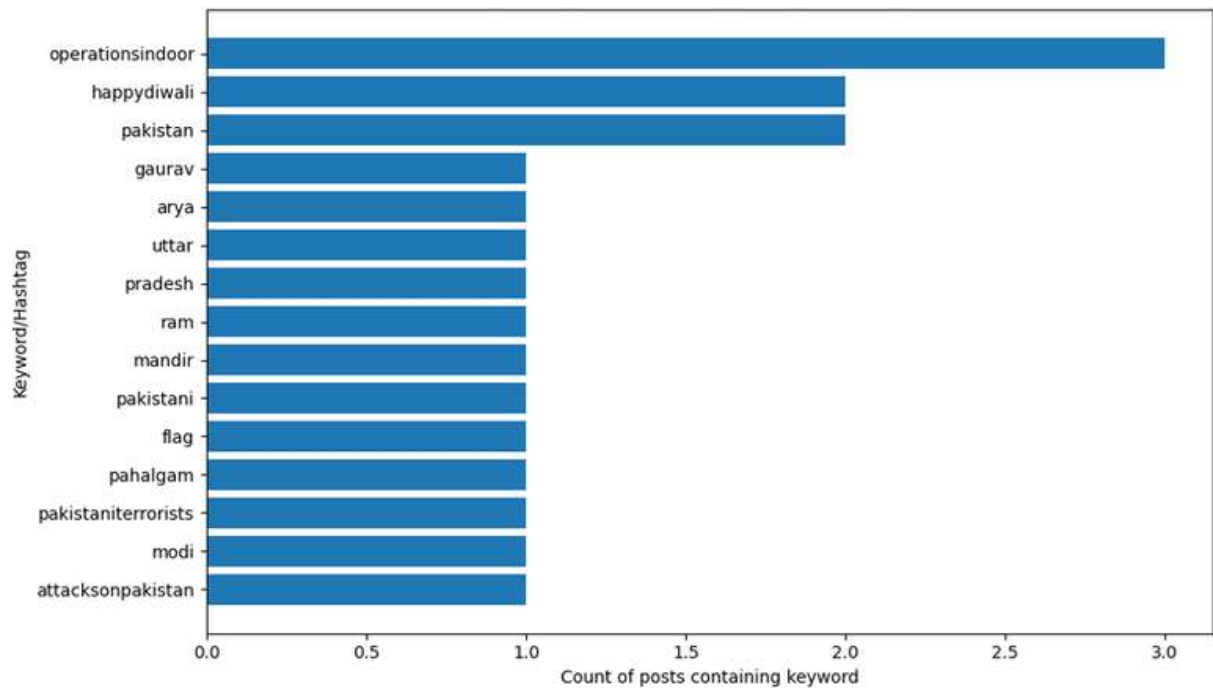
# Facebook & Instagram

Graphs 1 & 2 show the top 15 keywords and hashtags used on Meta platforms Facebook and Instagram during May 2025 at the height of the India-Pakistan conflict. Though the use of terms such as ‘Pakistani,’ Indian,’ or ‘Pahalgam’ is normal considering circumstances and current affairs at the time, terms such as ‘terrorists,’ ‘traitors,’ and ‘BLA’ show how hateful narratives were popular during this time period.

Graph 1: Top 15 keywords/hashtags (May) - Facebook



Graph 2: Top 15 keywords/hashtags (May) - Instagram



It was observed that among Pakistani content, misinformation frequently relied on video game footage from titles such as PUBG and World War 3 to depict Pakistani airstrikes and jet attacks in Indian locations such as Gujarat and Delhi that never occurred.



Hate speech romanticised the conflict as a religious struggle, with references to “ghazwa e hind<sup>[63]</sup>” and strong glorification of the Pakistani military, alongside hate speech directed at internal dissidents engaged in activist work. Gendered hate speech targeted Indian women using AI-generated content, employing tropes such as “maal e ghanimat” and mocking Indian Air Force women pilots who were falsely claimed to have been captured. This narrative was later proven false by officials on both sides. AI-generated footage also contained Hinduphobic and misogynistic elements, including clips depicting pilots from the Pakistan Air Force removing sindoor from Indian women.

A large number of entries originated directly from mainstream Indian news outlets, with Times Now having a particularly high volume of posts spreading disinformation and hate speech. These claims were often repeated by smaller accounts. Content included romanticising violence against Pakistan, using images from Gaza falsely labelled as “the situation in Pakistan,” calling for a new partition of Pakistan, invoking ideas of “Greater India,” and directing hate speech at Muslims by portraying them as inherently violent and savage and as the first colonisers of India.

<sup>[63]</sup> Battle of India



Warmongering narratives were common, with misinformation and sensationalised headlines claiming that Lahore had been destroyed, that Balochistan had declared independence, and that suspension of the Indus Waters Treaty was causing floods in Punjab. Additional misinformation included claims that Pakistan violated ceasefire agreements and was India's “nalaiq aulaad<sup>[64]</sup>,” along with false reports about Pakistan using explosive drones in India. Narratives also framed all Pakistanis as terrorists and repeated historical claims about Muslims as colonisers from figures such as Muhammad Ghorī to Aurangzeb.



<sup>[64]</sup> incompetent child

There was also vitriolic incitement to violence against Indian Muslims, including Bollywood celebrities, and Muslim women in India were targeted with claims that they were responsible for increasing the population of local terrorists.

## Bluesky

Although significantly smaller than platforms such as X, Facebook or TikTok, Bluesky emerged as an important platform within the crisis information ecosystem due to its growing adoption among journalists, researchers, policymakers and news organizations. Our observations found that while overall content volume remained comparatively low, misleading claims and unverified information were still able to circulate through niche networks and smaller news outlets seeking to establish visibility on the platform. In several instances, low-credibility accounts and lesser-known media sources shared unverified battlefield updates, casualty figures, and geopolitical claims that were subsequently amplified through reposts and quote-posts.

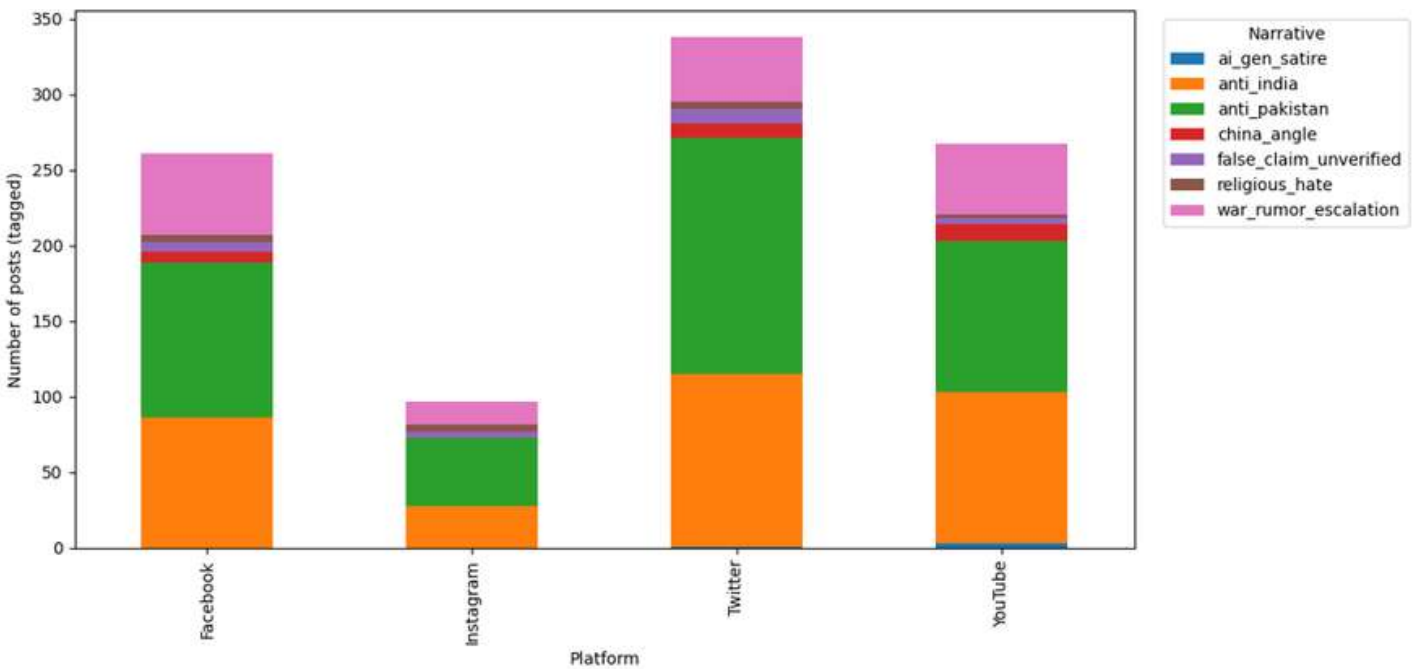


While Bluesky's decentralized architecture and community moderation features may reduce some forms of coordinated amplification seen on larger platforms, the platform nevertheless demonstrated vulnerabilities to the rapid spread of unverified information during periods of heightened uncertainty. The findings suggest that emerging platforms are not immune to crisis-related disinformation and should be included in future discussions on information integrity and platform accountability.

## Cross-Platform Comparison

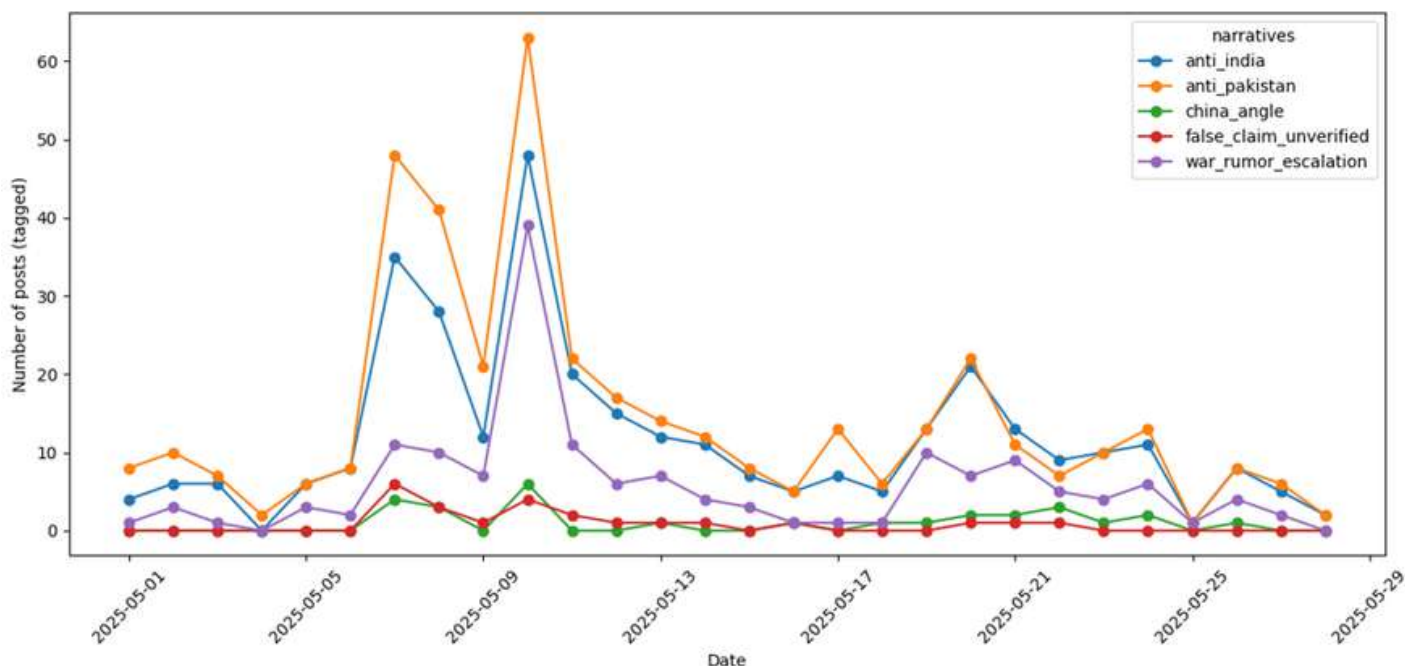
Consolidating these findings from each platform presents us with a comparison that shows how social media websites and applications today are widely being used in altogether similar ways to restate false narratives against Muslims and Hindus, exploit existing harms to vulnerable communities in society such as women and children, and fabricate false events, scenarios and evidence to instigate confusion and violence. Across all social media platforms, we witnessed a rise in hate speech, disinformation, deepfakes and gendered forms of harm – all of which were intensified with the use of generative AI tools. Graph 3 shows the dominant narratives found across social media platforms. Anti-Pakistan narratives were the most pervasive across YouTube, X, Instagram and Facebook in comparison to anti-India ones and there was a sizable amount of content on war rumors (which are counted as instances of mis/disinformation) alongside falsely claimed information. Though religious hate is shown as a mere sliver of content, this discrepancy may be due to content being generated in local languages and Roman Urdu/Hindi slang that platforms are currently unequipped to deal with and/or classify.

*Graph 3: Dominant narratives by platforms in May*



But while these commonalities in content are persistent and evident, a closer look at the data and trends shows variations in volume as well as platform moderation response. As graph 4 shows, while peaks in content were most prevalent during the conflict, significant rises were seen in content from YouTube after the ceasefire was signed. Interestingly, YouTube also showed the highest number of problematic content; a finding which was more prominent in India than in Pakistan. Indeed, as our analysis above shows, YouTube was a commonly used platform for hate speech and disinformation used by the Indian media and by self-proclaimed Indian journalists who routinely recreate long form content about current affairs and Indian politics.

*Graph 4: Timeline spikes: daily post volume by platforms in May*



Another major difference across platforms in the context of disinformation during the conflict (one that lingers even outside of significant events) is how responsive platforms are in moderating content whether that involves taking down problematic content or suspending entire accounts that persist in creating and uploading such content. At the time of this report’s publishing, all X posts documented by us from April - May 2025 are still up on the platform, whilst most YouTube videos propagating Islamophobia are still running alongside inappropriate generative AI images and videos of women from both sides of the border.

Though some isolated instances of disinformation and hate speech are still accessible on Meta platforms Facebook and Instagram, a significant chunk of videos and harmful content on the applications has been taken down. Similarly, most videos on TikTok have also been taken down and made unavailable on the platform. It’s important to note here that all content posted by legitimate Indian news channels on YouTube and Facebook, which violated platform guidelines on disinformation and hate speech, was not removed, whereas content posted by individuals was more likely to have been taken down once it was reported.

As we'll see further on this report, content moderation policies in their current state lack the nuance and geographical context needed to deal with violations that are steeped with information that is specific to the region, written in local languages and slang that is not supported by platforms and notions of gendered harm that breach South Asian societal norms and sensibilities.

## Discussion

While the staggering rise of disinformation on social media has been a serious concern to many politicians, journalists, and experts for years going as far back as Donald Trump's first election in 2016, it is evident from our findings that this is the inception of unprecedented times: disinformation is becoming more common than before, it is getting harder to detect and identify false content, and more importantly, its impact is drastically evolving into ever more treacherous and dangerous consequences for citizens, geopolitics, and how the world interacts with the news daily.

To reiterate, over the course of the India-Pakistan conflict which lasted from April to May in 2025, we recorded approximately 1,100 social media posts that were classified as disinformation, malinformation, hate speech, gendered disinformation, or generative AI content across various platforms. The posts found originated from Pakistan and India, with the majority of posts being from Indian accounts. Disinformation included spreading false information about ongoing attacks, fabricated lies about Muslims and Hindus from either side, gendered forms of online harm and hate speech, creating and posting AI-generated images and videos as well as isolated instances of falsely associating other states as allies.

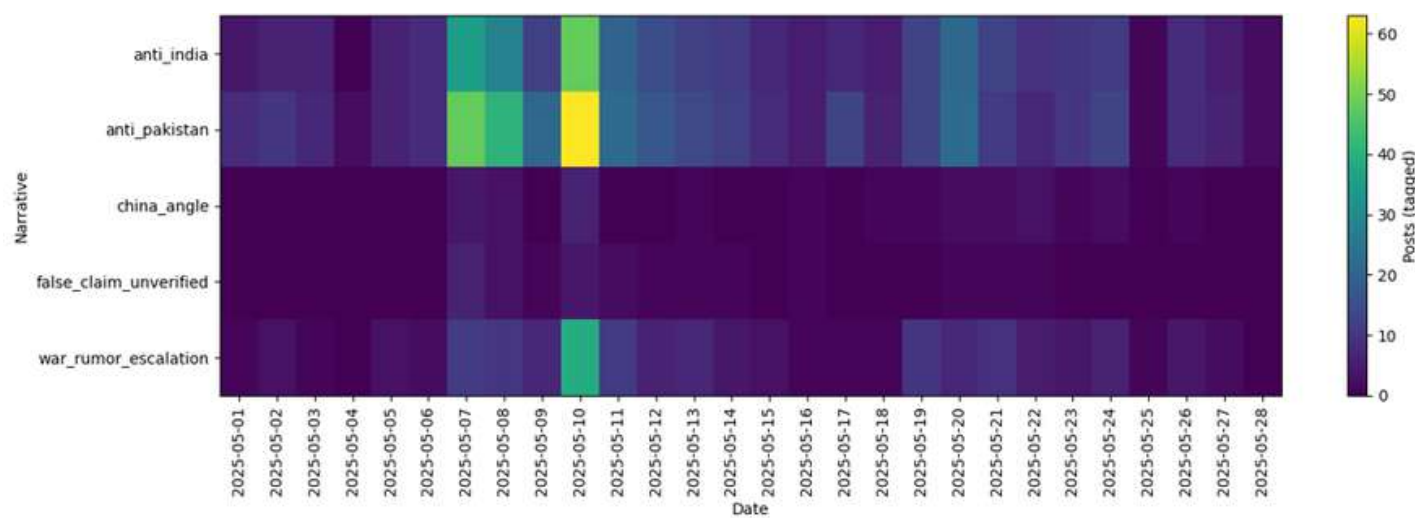
Though alarming statistics in and of themselves, the consequences of such blatant misuse of information for the exploitation of citizens and vulnerable populations can be felt immediately and even become more pronounced in the long run. Short-run tensions and long-term implications of disinformation in the wake of the conflict solidifies how in today's age, information is power and any control over facts and news, and its subsequent distortion, can mean control over regional responses, escalations and foreign policy. With an increase in cross-border tensions over the past 5 years, the world is seeing an exponential rise in disinformation, bringing to the forefront a global information crisis that raises multiple concerns, including political violence, lack of trust, heightened polarization in society, and serious human rights violations – a trend that is being increasingly persistent in conflicts that have cropped up ever since.

And while the immediate and instant reactions to information disorder during the India-Pakistan conflict were extremely harmful and a cause for concern, the impact does not stop there. Table 3 shows a list of narratives that persisted post-May 2025 just after a ceasefire was signed. Graph 5 similarly shows how persistent prominent specific keywords were, such as 'terrorists,' and 'indiapakistanwar,' even after hostilities were paused.

Table 3: Popular Post-May 2025 narratives on platforms

|                    |
|--------------------|
| war_rumor          |
| anti_pakistan      |
| false_claim_signal |
| anti_india         |

Graph 5: Heatmap of top 10 keywords by date



Post May 2025, the Indian subcontinent and global affairs felt the ripple effects of the disinformation campaigns surrounding the two countries, raising alarms for the long-term impression of the information landscape. Even at the time of this report, false narratives and animosities built a year ago have carried forward not only in this region but also to neighbouring countries.

EXPLAINER

Sport | ICC Men's T20 World Cup

### T20 World Cup: How India vs Pakistan relations went from bad to worse

From handshake row to match boycott – a look at the post-conflict controversial incidents in cricket's biggest rivalry.

Listen (7 mins)

Save

Share

Add AI Jazeera on Google

Inside Story

### Could water become a flashpoint between Islamabad and New Delhi?

Pakistan has warned India over the Indus Water Treaty. [Read more](#)

By Inside Story

5 Jul 2024

Research shows that spreading disinformation, either from state or non-state actors, has caused an erosion of public trust in the government and in all incoming news regardless of whether it's coming from a credible source. Belief in news outlets on television, the internet or social media is on a decline as a direct consequence of disinformation. According to the World Economic Forum (WEF), only 40% of people say that they consistently trust news<sup>[65]</sup>, with a growing number of people confirming a waning trust in journalism. With the media being an important component of democratic processes, acting as a watchdog, the public's distrust threatens journalism's central function of monitoring power and educating the public<sup>[66]</sup>. This has been found especially true in the aftermath of the India-Pakistan conflict, with many Indians, post-conflict, becoming wary of their local media outlets and information cycles.<sup>[67]</sup>

According to Al Jazeera, Indian media, which used to serve as a propaganda machine in the country, is now suffering a 'large-scale credibility crisis,' with many citizens responding to the spillover effects of its reputation in the global media landscape<sup>[68]</sup>. Alongside diminishing reliability amongst Indian citizens, the Indian government, in particular, has faced a largely public and massive backlash because of its blatant use of completely false, fabricated and propagandized information even by seemingly or previously trustworthy news channels, publications, and journalists<sup>[69]</sup>.

Similarly, information disorder in the form of hate speech, mis/disinformation, conspiracy theories, and sometimes hoaxes specifically aimed at shaping public opinion can impact citizens' trust in public institutions<sup>[70]</sup>. Considered an integral part of a successful democracy, a public's trust in their government and government-run offices allows for better political participation and general societal wellbeing<sup>[71]</sup>. In times of conflict or disaster, when civilians need access to real-time information to evacuation instructions, medical assistance, or entry/exit points, disinformation can have life-threatening impacts<sup>[72]</sup>. And instead of timely, accurate dissemination of information, the media begins inciting panic amongst communities, especially elder citizens, leading to heightening public anxiety and irrational behaviour<sup>[73]</sup>.

This mitigating factor of disinformation, i.e., its tendency to cause polarization in society, thus incites, motivates and rationalizes political violence. Disinformation that is driven by hate speech and propaganda campaigns against adversaries causes room for increased animosity

<sup>[65]</sup> <https://www.weforum.org/stories/2024/03/disinformation-trust-ecosystem-experts-curb-it/>

<sup>[66]</sup> <https://institute.aljazeera.net/en/ajr/article/2917>

<sup>[67]</sup> <https://www.newslaundry.com/2026/06/22/alarm-bells-ringing-why-indian-newsrooms-are-losing-public-trust>

<sup>[68]</sup> <https://www.youtube.com/watch?v=NLR9D4o2pqc>

<sup>[69]</sup> <https://www.theguardian.com/media/2025/may/28/how-social-media-lies-fuelled-a-rush-to-war-between-india-and-pakistan>

<https://www.nytimes.com/2025/05/17/world/asia/india-news-media-misinformation.html>

<sup>[70]</sup> <https://dl.acm.org/doi/epdf/10.1145/3603304.3604075>

[https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO\\_STU\(2021\)653635\\_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/653635/EXPO_STU(2021)653635_EN.pdf)

<https://thediplomat.com/2025/03/indias-growing-misinformation-crisis-a-threat-to-democracy/>

<sup>[71]</sup> <https://www.oecd.org/en/topics/trust-in-government.html>

<sup>[72]</sup> <https://civiliansinconflict.org/blog/disinformation-harms-civilians-in-conflict-in-more-ways-than-you-thought/>

<sup>[73]</sup> <https://institute.aljazeera.net/en/ajr/article/3188>

between countries or individuals divided by socio-economic groups, and more concerningly, can be exploited by states to capitalize on existing tensions to cause more harm and justify state action or military intervention against an aggressor<sup>[74]</sup>.

As our findings have shown, during the conflict, citizens from both sides of the border faced hate speech and trolling online, most of which derived from existing religious tensions between the two nations. Anti-Muslim hate surged online in India, not only in the form of discriminatory posts, comments and videos but also spilling out into physical forms of violence such as looting and vandalization of property<sup>[75]</sup>. Reports from The London Story detailed how social media platforms such as X and Facebook were becoming fertile grounds for the spread of “emotionally manipulative disinformation” that was turning into “drivers of nationalist incitement.”<sup>[76]</sup> This deliberate use of informational warfare can threaten public perceptions of security and power<sup>[77]</sup> alongside more extensive impacts to social and economic lives<sup>[78]</sup>. Moreover, it is the spread of such discriminatory and hate-driven speech that leads to a surge in human rights violations that run the risk of spilling outside borders.

Long-lasting, tangential impacts of the spread of disinformation during the India-Pakistan conflict have further shown how disinformation spills over into other aspects of bilateral relations even after escalations, and keeps fueling hostility. Within the region itself, this act of information disorder has created a pattern in reporting and discussions around major events where citizens (or governments themselves) resort to flooding social media platforms with a barrage of false information in an attempt to confuse and disorient the public. For example, early this year, India was found to be circulating disinformation ahead of Bangladesh’s elections<sup>[79]</sup>, in an attempt to dissuade voters. Pakistan likewise has been accused of taking part in disinformation campaigns against Afghans in the face of on-going tensions between the two countries<sup>[80]</sup> and citizens have also been found spreading false narratives that fuel Pakistan’s climate crisis<sup>[81]</sup>. More concerningly, the events and information landscape from May 2025 has set a dangerous precedent of governments assigning blame to their neighbours for anti-state narratives. At the onset of the war between Iran and the United States, Pakistani authorities claimed fake news was being spread by agents from Afghanistan and India<sup>[82]</sup>. While it is true that fabricated pieces of information about Pakistan hosting US military troops were widespread on the internet, their origin is still unclear. Further West, we see these trends multiplying in ever more harmful ways such as the use of AI to make hyper-realistic videos and images depicting missile attacks and bombs<sup>[83]</sup>.

<sup>[74]</sup> <https://hrcak.srce.hr/file/437200>

<sup>[75]</sup> <https://www.dw.com/en/india-pakistan-conflict-risks-deepening-religious-tensions/a-72529635>

<sup>[76]</sup> <https://thelondonstory.org/report/escalate-social-media-war-during-the-india-pakistan-conflict-april-may-2025/>

<sup>[77]</sup> <https://www.sciencedirect.com/science/chapter/edited-volume/abs/pii/B9780128162033000046?via%3Dihub>

<sup>[78]</sup> <https://hrcak.srce.hr/file/437200>

<sup>[79]</sup> <https://factcheck.afp.com/doc.afp.com.96N66KE>

<sup>[80]</sup> Ibid.

<sup>[81]</sup> <https://voicepk.net/2025/12/climate-disinformation-when-false-narratives-deepen-pakistans-climate-crisis/>

<sup>[82]</sup> <https://tribune.com.pk/story/2586751/info-ministry-calls-reports-of-pakistan-being-used-to-attack-iran-disinformation>

<sup>[83]</sup> <https://www.bbc.com/news/articles/ckg8wvz427vo>

# Platform Accountability & Policy Gaps

Across the platforms analysed in this study, the gap between stated policy and actual enforcement was consistent and conspicuous. X prohibits attacks on the basis of race, ethnicity, and religious affiliation, and bans the unwanted sexualisation of individuals - provisions violated wholesale across the 275 entries in our dataset. The Association for Progressive Communications (APC) found that X's community notes feature proved "ineffective and easily manipulated," with the platform taking no firm steps to curb hate speech during the conflict<sup>[84]</sup>. A USC Viterbi study<sup>[85]</sup> corroborates this, finding hate speech on X increased 50% between 2022 and 2023, with engagement on hateful posts growing at more than three times the rate of regular content, concluding that "the policies to reduce harmful content appear not to be sufficiently effective."

The picture across other platforms was no better. YouTube channels with millions of subscribers posted derogatory content targeting women, with no visible implementation of its strike system; the content remains live.



पाक कत्ताकार वेन, श्रद्धा लेगी मावरा की जगह? || #bollywood #indiapakistanwar



<sup>[84]</sup> <https://www.apc.org/en/pubs/apc-condemns-tech-platforms-inaction-disinformation-and-online-violence-escalate-during-india>

<sup>[85]</sup> <https://viterbischool.usc.edu/news/2025/02/a-platform-problem-hate-speech-and-bots-still-thriving-on-x/>



## Challenges in regulating cross-border disinformation

[86] <https://dc.law.utah.edu/ulr/vol2020/iss4/1/>

[87] <https://cyberpeace.org/resources/blogs/addressing-cross-border-misinformation-the-need-for-international-cooperation>

[88] <https://digitalrightsfoundation.pk/document/information-controls-in-india-and-pakistan/>

Some steps have, nevertheless, been taken to begin to curb cross-border disinformation. One example are the United Nations Global Principles for Information Integrity<sup>[89]</sup>, which lay down five principles for addressing misinformation, disinformation and hate speech. One of the principles lays out healthy incentives for tech companies to prioritise information integrity, as an alternative to current business models which enable “financial incentives and opportunities for purveyors of disinformation and hate who exploit the attention economy in which technology companies track user behaviour to collect data, feeding algorithms that prioritize engagement in a bid to maximize potential revenue for advertisers and creators”.

## **Pakistan’s disinformation clauses in PECA are leading to further censorship**

Pakistan’s legislative response to online disinformation has unfolded alongside and in many ways intensified the very crisis it claimed to address. The Prevention of Electronic Crimes (Amendment) Act 2025, passed in under fifteen minutes on 23 January 2025<sup>[90]</sup>, introduced a criminal offence for disseminating “false or fake information” carrying up to three years’ imprisonment, established the Social Media Protection and Regulatory Authority (SMPRA) to block content it deems unlawful, and extended the definition of defamable “persons” to include state institutions and the armed forces. Reporters Without Borders called it “a formidable legal weapon for blocking information<sup>[91]</sup>,” while Amnesty International warned the vague framing would “chill what little is left of the right to online expression in the country<sup>[92]</sup>.”

The central irony is that PECA was strengthened precisely as state institutions were mobilising disinformation as a tool of war. The DG ISPR toured universities across KP and Balochistan to shape narratives around Operation Bunyaan ul Marsoos, while state-affiliated accounts circulated unverified claims of military victory, none of which triggered the new provisions. What the law is structurally designed to target is the opposition that predates the conflict: Baloch activists, opposition journalists, and critics of the military who rely on digital spaces because mainstream media has already been tamed. The NCCIA, restored under Sections 29 and 30, now holds enhanced powers to investigate and arrest individuals for online speech violations, as seen most recently and publicly when human rights lawyers Imaan Mazari and Hadi Ali Chattha were arrested for making and posting “controversial tweets”.<sup>[93]</sup> Pakistan ranks 152 out of 180 on the RSF Press Freedom Index, and the Amendment does little to suggest that will change.

<sup>[89]</sup> <https://www.un.org/sites/un2.un.org/files/un-global-principles-for-information-integrity-en.pdf>

<sup>[90]</sup> <https://www.ifj.org/media-centre/news/detail/article/pakistan-peca-amendments-further-tighten-government-grip-on-digital-expression>

<sup>[91]</sup> <https://rsf.org/en/pakistan-instead-revising-legislation-censors-dissent-sharif-s-government-strengthened-it>

<sup>[92]</sup> <https://www.amnesty.org/en/latest/news/2025/01/pakistan-authorities-pass-bill-with-sweeping-controls-on-social-media/>

<sup>[93]</sup> <https://www.icj.org/pakistan-arbitrary-arrest-and-sentencing-of-human-rights-lawyers-imaan-zainab-mazari-hazir-and-hadi-ali-chattha/>

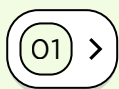
# The need for regional and global cooperation

The May 2025 conflict exposed a structural vacuum at the heart of digital governance in South Asia: no regional framework exists to coordinate a response to cross-border disinformation between adversarial states. Researchers at the Bulletin of the Atomic Scientists have warned that AI-generated synthetic content poses two specific dangers within a nuclear dyad: strategic confusion and the risk of misreading the adversary's intent. The conflict came close to demonstrating both: viral claims of radioactive leaks at Kirana Hills triggered widespread panic that no platform was equipped to contain, and no bilateral mechanism existed to rebut them. In a region with three wars, contested nuclear postures, and deep mutual distrust, the absence of any coordinated institutional response is not merely a policy failure — it is a security risk.

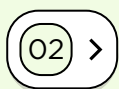
Meaningful cooperation would require concessions neither government appears willing to make unilaterally. For India, it would mean acknowledging that its media ecosystem, now in what Al Jazeera has termed a “large-scale credibility crisis”, was a primary vector of disinformation during the conflict. For Pakistan, it would mean subordinating PECA's surveillance logic to genuine press freedom commitments. The UN Global Principles for Information Integrity offer a partial foundation, and scholars have called for shared rapid-response fact-checking protocols, mandatory platform disclosure requirements during conflict periods, and bilateral channels to verify or rebut viral claims in real time. In a region where disinformation has already brought two nuclear-armed states to the edge of miscalculation, the cost of continuing without such frameworks is a risk the international community can no longer afford to treat as someone else's problem.

# Recommendations

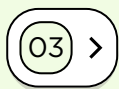
## Platforms



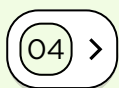
Strengthen crisis-response moderation mechanisms. Platforms should establish dedicated escalation protocols during periods of geopolitical conflict and military crises. Such mechanisms should prioritize the rapid review of content involving battlefield claims, casualty figures, manipulated media and incitement to violence. Given the speed at which misinformation circulated during the May 2025 crisis, delayed moderation often allowed false narratives to reach millions before intervention occurred.



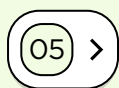
Improve detection and labeling of manipulated and AI-generated content. The widespread circulation of synthetic media, AI-generated visuals, recycled footage and fabricated battlefield imagery demonstrates the urgent need for stronger detection systems. Platforms should invest in multilingual AI-detection tools, visual provenance technologies and contextual labels that clearly identify manipulated or unverifiable content.



Address gendered disinformation and conflict-related hate speech. Platforms must recognize gendered disinformation as a distinct harm category during geopolitical crises. The study documented widespread misogynistic, sexualized and dehumanizing content targeting women, including female journalists, public figures, military personnels and ordinary users. Existing hate speech and harassment policies should be consistently enforced, particularly against content that weaponizes gender, religion or ethnicity.



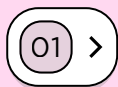
Increase transparency around moderation decisions. Platforms should publish crisis-specific transparency reports detailing content removals, labeling decisions, account restrictions, government requests and enforcement outcomes. Greater transparency would allow researchers, journalists and civil society organizations to independently assess platform responses during periods of heightened risk.



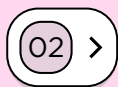
Improve researcher access to platform data. Independent researchers and civil society organizations continue to face significant barriers in studying disinformation. Platforms should provide secure and privacy-respecting access to public-interest datasets, enabling greater scrutiny of information flows, coordinated influence operations and platform governance practices.

# Recommendations

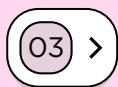
## Governments



Prioritize transparency over information restrictions. Governments should avoid relying solely on content blocking, internet restrictions or platform bans during crises. While combating harmful misinformation is a legitimate objective, broad restrictions often undermine access to verified information and may inadvertently increase reliance on rumors and unofficial sources.



Invest in crisis communication and public information systems. The study demonstrates that informational vacuums contribute significantly to the spread of misinformation. Governments should strengthen official communication channels that provide timely, evidence-based and verifiable updates during emergencies, reducing opportunities for false narratives to gain traction.



Develop safeguards against state-sponsored disinformation. Public institutions and state-affiliated actors should adhere to principles of accuracy, transparency and accountability when communicating during periods of conflict. Governments should commit to refraining from disseminating unverified claims, manipulated media or inflammatory rhetoric that may contribute to escalation.



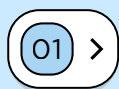
Support independent fact-checking ecosystems. Governments should create enabling environments for independent journalists, fact-checkers, researchers and media organizations to investigate and debunk false claims without fear of censorship, retaliation or political interference.



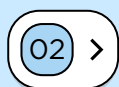
Promote regional cooperation on information integrity. Given the cross-border nature of disinformation during the India-Pakistan crisis, regional dialogue mechanisms should be developed to facilitate information sharing, crisis-response coordination and collaborative approaches to addressing harmful online content while respecting human rights standards.

# Recommendations

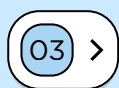
## Civil Society & Media



Strengthen rapid-response fact-checking networks. Civil society organizations, researchers and independent media should develop collaborative monitoring and verification mechanisms capable of responding quickly to emerging narratives during crises. Cross-border partnerships can play an important role in identifying and debunking misleading claims before they achieve widespread reach.



Invest in digital and media literacy initiatives. Public awareness campaigns should focus on helping users identify manipulated media, AI-generated content, misleading headlines and emotionally charged misinformation. Particular attention should be given to younger audiences who increasingly consume news through short-form video platforms.



Document and monitor gendered disinformation. Women's rights organizations, digital rights groups and media actors should continue documenting how gendered narratives, sexualized misinformation and online harassment are weaponized during political and military crises. Such monitoring can help inform platform accountability efforts and policy reform.



Develop ethical crisis-reporting standards. Media organizations should strengthen editorial safeguards for reporting during conflicts, including verification requirements for user-generated content, AI-generated imagery and battlefield claims. Newsrooms should prioritize accuracy over speed and avoid amplifying unverified information originating from social media.



Build long-term information integrity coalitions. Addressing disinformation requires sustained collaboration between researchers, journalists, civil society organizations, academic institutions and technology platforms. Long-term coalitions can support evidence-based advocacy, improve public understanding of information threats and strengthen accountability across the digital ecosystem.

# Conclusion

The May 2025 India-Pakistan crisis demonstrated that information disorder is no longer a secondary consequence of conflict but a central feature of contemporary geopolitical crises. Addressing the spread of misinformation, hate speech and AI-generated content requires coordinated action from platforms, governments, media organizations and civil society actors. Strengthening accountability, transparency and information integrity mechanisms will be essential to mitigating future harms and protecting democratic discourse during periods of heightened political tension.



DigitalRightsFoundation  
"KNOW YOUR RIGHTS"



@DigitalRightsFoundation



@digitalrightsfoundation



@digitalrightsfoundation



@digitalrightsfoundation



Digital Rights Foundation



@digitalrightspk.bsky.social



@DigitalRightsPK



@DigitalRightsPK

digitalrightsfoundation.pk