Asian Journal of International Peace and Security (AJIPS)

ISSN-e: 2707-8809

Vol. 10, No. 2, (2026, Summer), 1-16

The Second-Generation Threat: AI-Assisted Decentralized Militancy and Counter-Extremism Governance in South Asia

Kalim Ullah,[1] & Muhammad Umar Abbasi[2]

Abstract:

Existing counter-extremism scholarship and governance practice remain largely anchored to the Islamic State threat model of 2014–2018, which assumed centralized production, stylistically recognizable output, and predominantly Arabic and English distribution. Whether that framework is still adequate for the South Asian threat environment is the question this article addresses. Generative AI tools have, since approximately 2019, enabled decentralized Militant Islamist networks to produce vernacular-language propaganda at near-zero marginal cost, a qualitative shift that the post-2014 counter-extremism consensus did not anticipate. Drawing on a structured comparative case study of three national sub-systems, Pakistan's Tehreek-e-Taliban Pakistan and its Umar Media wing, Bangladesh's Ansar al-Islam ecosystem, and India's WhatsApp distribution environment, and applying Marc Sageman's network theory of leaderless jihad alongside Tarleton Gillespie's platform governance framework, with J. M. Berger's account of algorithmic amplification as a bridging device, the analysis identifies three compounding structural deficits in the current governance architecture: a South Asian vernacular language moderation gap, an absence of definitional categories for AI-generated content in national regulatory frameworks, and inter-state coordination failures shaped by competing geopolitical interests.

Keywords: South Asia, Tehreek-e-Taliban Pakistan, Ansar al-Islam, artificial intelligence, militant communication; governance gap, counter-extremism governance

INTRODUCTION

The qualitative nature of militant Islamist communication in South Asia has undergone a shift which the post-2014 counter-extremism consensus did not foresee. From about 2014 to 2018, the most common analytical benchmark was the Islamic State's high production media output, such as the magazines Dabiq and Rumiyah, the videos compiled by Al-Hayat Media Center, and the Arabic and English Twitter spheres identified by Berger and Morgan (2015). That model was centralized, recognizable in style and largely Arabic-language, and English was the main language used for recruiting the diaspora. It in turn produced a counter-extremism architecture that was commensurate with its outward manifestations: keyword filters, image hash databases, takedown

[1] Assistant Professor, Department of International Relations, National Defence University, Islamabad, Pakistan. Email: kalimullah@ndu.edu.pk

[2] Assistant Professor, Department of International Relations, National Defence University, Islamabad. Email: umarabbasi@ndu.edu.pk

protocols negotiated through the Global Internet Forum to Counter Terrorism, and bilateral intelligence-sharing arrangements that assumed propaganda would come out of nodes that were identifiable as being produced in Arabic or English.

The world it was designed for is no longer. The South Asian militant Islamist communication environment has changed since 2019, and especially since the Taliban's return to power in Afghanistan in August 2021, towards decentralized, vernacular, AI-assisted production. Now, Umar Media, the broadcasting arm of the Tehreek-e-Taliban Pakistan (TTP), is producing what Basit (2025) calls "AI-generated news bulletins, which include translations and voice-overs in Sindhi, Punjabi, and Balochi languages, apart from Urdu, Pashto, and English." Generative tools have been used by pro-Islamic State networks to create visual propaganda, which Criezis (2024) catalogued across hundreds of items on Instagram, Telegram, and related platforms. Asif Adnan is one of the urban operatives in Bangladesh who, as Bashar (2025) notes, "with mainstream academic backgrounds, can reach more than three hundred thousand followers on Facebook, TikTok and YouTube through linguistically vernacular content." In India, the encrypted distribution via WhatsApp's user base of around 532 million creates a downstream environment where AI-generated communal incitement meets with militant Islamist mobilization (Banaji & Bhat, 2019).

This article contends that the threat model that underpinned the development of the counter-extremism governance architecture across South Asia, including Pakistan's Prevention of Electronic Crimes Act 2016 (PECA), Bangladesh's Cyber Security Act 2023 (CSA), India's Information Technology Act 2000 with the Intermediary Guidelines 2021, and the content moderation systems of Meta, Google, ByteDance, and Telegram, was based on the threat model of 2014–2018 and is structurally ill-equipped to address the current threat model. The argument is in three steps. The first is the empirical change in the threat itself: production has been decentralized, costs have fallen, algorithmic dispersion of stylistic recognisability has occurred, and vernacular languages have taken the place of Arabic as the main medium. The second relates to the design assumptions of the governance architecture: moderation algorithms are still disproportionately geared towards Arabic and English, regulatory definitions do not consider AI-generated content as a legal category, and bilateral intelligence arrangements are systematically undermined by competing strategic interests. The third relates to the cumulative nature of these mismatches.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

The argument that the threat model that underpinned the development of the counter-extremism governance architecture across South Asia, including Pakistan's Prevention of Electronic Crimes Act 2016 (PECA), Bangladesh's Cyber Security Act 2023 (CSA), India's Information Technology Act 2000 with the Intermediary Guidelines 2021, and the content moderation systems of Meta, Google, ByteDance, and Telegram, was based on the threat model of 2014–2018 and is structurally ill-equipped to address the current threat model is motivated by two theoretical resources.

The analytical anchor is provided by Sageman's (2004; 2008) network theory of militant Islamist mobilization, which suggests that the "leaderless jihad" logic that he identified is network-resilient when production costs approach zero and stylistic recognisability is diffused through algorithmic variation. The South Asian language deficit exacerbates rather than softens the hard limits that Gillespie's (2018) account of platform governance provides for understanding why content

moderation alone cannot close the gap: commercial incentives, jurisdictional ambiguities, and the technical architecture of moderation systems set hard limits on what platforms can achieve. The connection is provided by Berger's (2018) research on algorithmic amplification: recommendation systems amplify engagement, irrespective of political content, and do the distribution work that decentralized cells could not organize for themselves.

Section I presents the structural change from the 2014-2018 model to the post-2019 AI-assisted environment, and builds the theoretical framework based on Sageman. Section II maps the governance architecture in the three cases in light of Gillespie and the design assumption mismatch. The Pakistan, Bangladesh and India cases are discussed in Section III, IV and V respectively. Section VI brings together the three structural elements of the governance gap and speculates on the nature of realistic, narrowly targeted policy responses. The conclusion summarizes the theoretical contribution, notes limitations, and suggests a research agenda. Throughout the empirical claims made are separated from the normative claims, and when evidence is incomplete or disputed, this is noted rather than obscured.

RESEARCH METHODOLOGY

This analysis proceeds by means of a structured comparative case study across three national sub-systems within the South Asian digital governance space: Pakistan, Bangladesh, and India. The three cases are selected on three criteria. First, each hosts a documented Militant Islamist communication ecosystem alongside a distinct national regulatory framework, allowing cross-case comparison of governance design assumptions. Second, the cases vary on the independent variable of interest – the degree to which AI-assisted production is empirically documented – from Pakistan, where evidential density is highest, through Bangladesh, where documentation gaps exist alongside structural warning indicators, to India, where the threat operates primarily through enabling conditions rather than directly documented AI-generated extremist content. This range provides variation rather than selection on the dependent variable. Third, together they encompass the principal language communities and the three principal regulatory architectures of South Asian counter-extremism governance. The evidence base comprises government statutes, platform policy statements, and legal filings; secondary scholarship in peer-reviewed journals; and reports from established research institutions including GNET, RSIS, the Jamestown Foundation, the International Crisis Group, and the Soufan Center. The analysis is qualitative throughout: theoretical frameworks are applied to explain cross-case variation rather than to generate causal inference from Large-N data. Claims resting on incomplete or inference-dependent evidence are flagged as such in the body text.

AI-ASSISTED DECENTRALISATION: THE STRUCTURAL SHIFT IN SOUTH ASIAN MILITANT COMMUNICATION

For much of the period from 2014 to 2019, the analytical vocabulary of online jihadism in South Asia tracked developments elsewhere. Understanding why it no longer does requires tracing the confluence of three structural trends that have reshaped how Militant Islamist content is produced, transmitted, and received in the region.

For most of five years, the analytical lexicon of online jihadism in South Asia was borrowed from another theatre. Berger and Morgan (2015) found that there is “a relatively small but hyperactive

core of between 500 and 2,000 accounts" that account for the majority of high-volume amplification (p. 5). In his reflection on the Quilliam study, Winter listed the Islamic State's output for the Islamic month of Shawwal in 2015: 1,146 separate items in videos, photos, news articles, magazines, radio bulletins and songs, which he reduced to 892 discrete items through his analysis (Winter, 2015b). Farwell (2014) highlighted the importance of the group's media strategy, which relied on "deft" cross-platform amplification (p. 49). The picture of militant Islamist communication that was created was centralized, branded, stylistically overloaded, and predominantly Arabic in its primary distribution and English in its diaspora extension. The analytical literature on South Asian Militant Islamist media during this period tended to view it as derivative, with TTP content being seen as adapted from al-Qaeda's As-Sahab template, and with little vernacular adaptation (Mir, 2024).

The structural change since 2019 is not just one technological development, but rather the confluence of three trends explains its contours. The first is the fall of production costs. Generative models for text, voice, and image are now available to actors who do not have specialized studios, and Criezis (2024) reports on pro-ISIS accounts sharing AI-generated content on Instagram, Meta-owned platforms, and Telegram. The second is the replacement of Arabic and English with vernacular languages of South Asia. As mentioned by Umar Media in Basit (2025), the recent production of Umar Media has been translated into Urdu, Pashto, Sindhi, Punjabi, Balochi, and Brahui with the help of artificial intelligence. The third is the shift from open social media to encrypted messaging and algorithmically-driven short-video ecosystems, which Conway (2017) predicted would be a future trend in terrorism studies, which should move beyond Twitter and Facebook.

The term "AI-assisted content" needs to be carefully defined in this context. As used here, the term applies to content where generative models contribute in a non-trivial way to the creation of the content: text written or significantly edited by large language models; voice synthesized using neural networks trained on archival speech samples; images generated from textual prompts using diffusion models; and video edited or composited using automated tools. Two observations follow. Today, the predominant mode of communication for South Asian Militant Islamists is hybrid, with AI-generated content being voiced over archival video footage, or AI-translated subtitles accompanying human-generated content. The problem with AI-assisted content is that it's hard to fingerprint: while the IS visual brand was consistent from 2014 to 2018, AI-generated content is different in style from one generation to the next, making it impossible to fingerprint.

For this article, the term "second-generation threat" denotes a qualitatively distinct configuration of Militant Islamist communication, distinguishable from the 2014–2018 Islamic State model on four criteria. First, production is decentralized: content is generated by dispersed cells rather than institutional media commissions. Second, stylistic recognisability is algorithmically dispersed – generative variation makes signature-based detection technically unreliable across production runs. Third, vernacular South Asian languages replace Arabic and English as the primary distribution media. Fourth, distribution operates through encrypted messaging and algorithmically driven short-video ecosystems rather than open social media platforms subject to conventional takedown protocols. A first-generation threat retained at least two of these four features in inverted form: it tended towards centralized production, stylistic consistency, Arabic or major-language primary distribution, and open-platform channels. The distinction matters analytically because

each feature of the second-generation configuration directly undermines a corresponding design assumption of the counter-extremism architecture examined in Section II.

The theoretical underpinning for understanding the importance of this shift is Sageman's (2004, 2008) network theory. Sageman (2004) claimed that the Militant Islamist movement after 9/11 was not hierarchical but rather a collection of "spontaneous, self-organized bunches of guys" (p. 137). The threat had now become "a multitude of informal local groups" inspired by, but not directed by, al-Qaeda central; Leaderless Jihad (Sageman, 2008) continued the argument. At the time, Sageman's argument was challenged, most notably by Hoffman (2008), who argued that organized structures were still important. What is important for the present argument is that Sageman's framework identified a generative mechanism: mobilization can take place without central coordination when production and communication costs are low enough. The 2008 formulation was not able to predict the size of the cost collapse that generative AI would create. The leaderless logic is non-linear, and when the marginal cost of producing a propaganda video in Pashto with synthetic voice-over approaches zero, the logic becomes even more so.

The transformation is an observable change. According to RSIS analysis by Sayed (2024), TTP released around sixty-five videos between 2007 and 2013, and in the last four years, TTP has released over sixty-four videos. Hegghammer (2017) has written about jihadi culture, and his work highlights the analytical significance of aesthetic dispersion: while previous scholarship on jihadist culture has focused on its ornamental aspects, the post-2019 South Asian context demonstrates that aesthetic specificity, which uses Pashtun tribal imagery instead of generic Arabic calligraphy, serves the recruitment function that Hegghammer identified, but now at zero marginal cost. Weimann's (2006) previous mapping of terrorist use of the Internet, based on data mining and psychological warfare, created a typology of eight functional categories, which was still useful, but the assumption that each function would need its own infrastructure has been superseded by generative consolidation.

Berger's (2018) description of algorithmic amplification rounds out the picture. Recommendation systems, overall, do not measure political valence; they measure engagement. Emotionally engaging, visually striking, and linguistically relatable AI-generated content is effective in these areas. The algorithmic systems of platforms like TikTok, YouTube Shorts, and Instagram Reels are, therefore, actively engaged in the distribution function that decentralized cells cannot otherwise organize. Cronin's (2020) analysis of "open" technological diffusion reflects the political economy behind it: capabilities once monopolized by states or by well-resourced organizations have been diffused to small networks of users. All these analytical tools point to the fact that the vocabulary of the Caliphate era is no longer a good description of the South Asian threat environment. In practice, the policy instruments that have been developed as a result of the Caliphate-era response are geared towards features that the current threat no longer has.

Platform Governance and Regulatory Architecture: Design Assumptions and Structural Constraints

The mismatch between the current threat environment and the governance architecture designed to address it is not incidental – it follows from the design assumptions baked into both platform moderation systems and sovereign regulatory frameworks during the Islamic State era. Three

registers of that mismatch are examined here: commercial platform governance, national legislative frameworks across the three cases, and the bilateral and multilateral coordination layer.

Gillespie (2018) views content moderation as not an ancillary function, but as an integral part of what platforms are. Moderation is not an ancillary aspect of what platforms do, he writes; it is essential, constitutional, definitional. Not only can platforms not survive without moderation, but they are not platforms without it (p. 21). This formulation has a number of analytical implications for the South Asian situation. It suggests, first, that the content that platforms moderate is driven by commercial incentive structures, not by public-interest reasoning; second, that there are hard limits to moderation, given the jurisdictional and legal-architectural constraints of platforms; and third, that the technical architecture of moderation, including reliance on training data and language-specific models, structurally privileges some categories of content over others.

In all three main cases studied here, the current governance structure is characterized by instruments used for a previous threat model that are not well adapted to the current threat and are used for political control. Under the Prevention of Electronic Crimes Act 2016, which was passed as part of the National Action Plan response to the December 2014 Army Public School attack in Peshawar, Pakistan, a wide range of behaviour is criminalized, such as glorification of an offence, cyber terrorism, and dissemination of hate speech and extremist content (Government of Pakistan, 2016, 9-11). The law gives the Pakistan Telecommunication Authority the power to issue blocking orders and the Federal Investigation Agency the power to enforce. In reality, the academic literature shows that PECA has been used more often against journalists and political dissidents than against the militant content it was nominally created for. The statute does not envision AI-generated content as a separate legal category, and the blocking provisions assume the existence of identifiable hosts and locatable distribution chains.

The regulatory path of Bangladesh is a case in point with a stronger manifestation. Digital Security Act 2018 (subsequently replaced by the Cyber Security Act 2023) was explicitly described as an anti-extremism measure. Amnesty International's (2024) comprehensive analysis of the CSA concludes that the new law is “a categorical statement: the new statute is essentially a replication of the DSA. It is merely an attempt to douse growing international pressure through performative reforms which repackage authoritarian provisions and practices into an ostensibly new law” (p. 5). Similar continuity is found in the comparative analysis by Transparency International Bangladesh (2023). The moral objection to the statute is not the issue here, but the fact that its use against opposition voices has taken up enforcement resources that could have been used against the second generation threat, and that its use has created a regulatory credibility deficit that undermines the legitimacy of any counter-extremism enforcement action taken under its authority.

India has the most institutionally developed regulatory system of the three. The Information Technology Act 2000 was originally conceived as an e-commerce tool, and the counter-content provisions, especially Section 69A and Section 79, were introduced by way of amendment to deal with blocking and intermediary safe harbour. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 impose several new duties, including the much-debated Rule 4(2) traceability requirement, which is the subject of WhatsApp's litigation in the Delhi High Court. In its filing in 2021, WhatsApp contended that Rule 4(2) would necessitate the company to “build the ability to identify the first originator for every message sent in India” and

"break end-to-end encryption" (Thapliyal, 2021). The Internet Society brief concludes that traceability "would involve an intermediary possessing the key to decrypt the originator's information in escrow, which would be a fundamental change to the encryption model" ("Traceability in end-to-end," 2024).

The commonality of these three regulatory frameworks is that they are based on a design assumption that is inconsistent with the post-2019 threat. Both assume that there are identifiable content originators, that the content is hosted somewhere, and that the content is signature recognizable. All were written before the advent of generative AI as a production technology. None is a legal term that does not distinguish between human and AI-generated content. None offers methods to check the origin in a world where stylistic fingerprints are no longer valid. As a consequence, the regulatory instruments are in practice blocking rather than predictive tools for disruption, and their reactive nature is exacerbated by a lack of resourcing.

There are similar restrictions in platform-side governance. Meta, Google, ByteDance and Telegram are the major platforms operating in South Asia, all of which have content moderation systems that are trained on English and other major European languages, with Arabic receiving significant investment since 2014, directly in response to the Islamic State challenge. Moderation capacity in Urdu, Bengali, Pashto and Tamil is still weak. In 2021, an investigation revealed that Meta has only implemented AI content moderation algorithms in forty languages, while claiming to support over one hundred (Duffy, 2021). In 2024, the Terrorist Content Analytics Platform (TCAP) entry in the Tech Against Terrorism inclusion policy for TTP states that it alerts on "official content" but excludes "supporter-generated content, accounts, and channels" from automated scope ("TCAP Expansion Profile," 2024) . This boundary implies that most of the decentralized amplification is outside of the scope of platform alerting systems. The structural mismatch is exacerbated by the shift of activity to encrypted and algorithmic spaces where moderation is technically impossible or politically disputed.

One could say that platform moderation, sovereign regulation, and bilateral coordination are a layered defense, each of which makes up for the other's shortcomings. This is not the case in the empirical record. The design assumptions of each layer are the same, and the capacity constraint of each layer shares the same characteristics: linguistic asymmetry, definitional incompleteness, and an enforcement architecture tailored to identifiable hosts and identifiable producers. In practice, the three layers don't compensate; they compound: when platforms fail, states point to platforms; when states fail, platforms point to the lack of jurisdiction; when coordination fails, both point to geopolitical conditions.

Pakistan: AI-Assisted TTP Communication and Afghan Dimension

Pakistan offers the most empirically dense case of AI-assisted decentralized Militant Islamist communication in South Asia, making it the appropriate starting point for case analysis. Three qualitative changes in TTP's information operations since 2021 are directly relevant to the governance gap identified in Section II.

Pakistan is the most empirically developed case of AI-assisted decentralized Militant Islamist communication in South Asia. The Tehreek-e-Taliban Pakistan, which was established in December 2007 as an umbrella organization for the militant networks of the tribal belt, has an information

operation that Basit (2025) calls “irregular” video testimonies under the early Umar Studio brand to a “professional and centralized” media commission that oversees audio, video, magazines, FM radio, and social media. The change has gained momentum since August 2021 when the Afghan Taliban consolidated their control over the state, providing a conducive environment for the production of TTP media content from Afghan soil.

There are three interwoven aspects of the qualitative nature of TTP communication that have changed. The first is the diversification of language. Basit (2025) reports that “Umar Media is now regularly producing news bulletins in Urdu, Pashto, English, Sindhi, Punjabi and Balochi languages, which are translated and voiced by AI.” The use of regional languages, such as Brahui, enables the group to reach audiences in Sindh, Punjab and Balochistan that the previous Pashto-centric communication approach failed to reach. The second is related to the incorporation of vernacular cultural symbolism, including Pashtun nationalist imagery, references to the former Federally Administered Tribal Areas, and explicit framing of the Pakistani military as an “occupier” instead of an apostate force, which is a shift from the takfiri vocabulary of previous eras to anti-colonial discourse (Basit, 2025). The third relates to integration of personnel: the appointment in 2023 of Muneeb Jatt, who was previously linked to al-Qaeda's As-Sahab Urdu operations, as Umar Media director, which integrated the technical and editorial capability of the operation.

As Basit (2025) himself notes, the ramifications for recruitment are far-reaching: “TTP's use of AI tools to generate propaganda will draw educated youth from the urban middle class, particularly those with information technology, computer science, and software engineering backgrounds.” The most analytically arresting aspect of this claim is that it is an inversion of the previous sociology of South Asian militant recruitment. While the dominant pattern of the late 1990s to the late 2010s was based on the mobilization of madrasa networks and tribal-belt, the AI-assisted turn opens the door to recruitment among urban, technically educated demographics that were not targeted by previous propaganda formats. This is the Sagemanian logic pushed to its empirical limit: leaderless cells generate vernacular content; which algorithmic dissemination brings to audiences whose social networks were previously protected from militant messages.

The Prevention of Electronic Crimes Act 2016 framework in Pakistan is structurally not equipped to this development. Under the National Action Plan, the statute criminalizes glorification of an offence, cyber terrorism and dissemination of unlawful content, and the Pakistan Telecommunication Authority (PTA) has the authority to issue blocking orders (Government of Pakistan, 2016). PECA has three structural features that constrain its ability to address decentralized content with AI assistance. First, the definitional categories in the statute require an identifiable information system host and an identifiable service provider, both of which are not easily identifiable when content is created via distributed cells using consumer-grade generative tools and routed via encrypted channels. Second, the statute does not distinguish between AI-generated content and human-generated content, and prosecutions would be brought under the statute as written for human-generated content, which raises evidentiary issues of intent, authorship, and provenance. Third, the enforcement apparatus is chronically under-resourced: the FIA's Cyber Crime Wing has been overburdened with cases of online harassment, defamation and political offences, as opposed to counter-militancy.

These restrictions are exacerbated by the Afghanistan dimension. Since the reconsolidation of the Taliban in August 2021, the space for TTP media production in Afghanistan has grown considerably. According to The Soufan Center (2025), "Pakistan is pressuring the Afghan Taliban to crack down on the TTP, which the former finds unpalatable for ideological reasons", and that "kinetic actions against the Pakistani Taliban could lead to splintering, with some of the more hardcore fighters migrating to ISK and reinforcing its ranks". The collapse of the capacity of the Afghan state to govern digitally has left a regulatory vacuum: there is no functional Afghan equivalent of PECA, no functional content-blocking authority, and no bilateral mutual-legal-assistance mechanism by which Pakistan could seek enforcement action against TTP media infrastructure in Afghanistan.

According to Sayed (2024), TTP produced approximately sixty-five videos from 2007 to 2013, and has produced over sixty-four since 2021. This compression is partly due to the operational flexibility of the Afghan environment, partly to the use of AI tools that enable lower-cost translation, voice-over, and editing, and partly to the personnel rationalization that was part of the 2023 reorganization. The downstream effect is that the information ecology surrounding TTP attacks has expanded: every major operational event is now followed by professionally produced multilingual video and audio output, which is disseminated via Telegram channels, supporter networks on X and Facebook, and an official Umar Media website that, according to Tech Against Terrorism's TCAP profile, receives an average of around eleven thousand monthly visitors, with sixty-five percent of traffic coming from Pakistan ("TCAP Expansion Profile," 2024).

The bilateral dimension closes the analytical circle. The cooperation between Pakistan and Afghanistan in the field of countering extremism, if any, that existed in the post-2001 era has ended. On 9 October 2025, Pakistan launched Operation Khyber Storm on the cities of Kabul, Khost, Jalalabad and Paktika, to which the Afghan Taliban forces responded ("Pakistani, Afghan forces," 2025). Thus, the Pakistani state's response to Umar Media has shifted from cooperative disruption to kinetic targeting, which has proven ineffective against information infrastructure; as Cronin (2020) foresees, the productive capacity of decentralized cells is now largely independent of physical infrastructure. Structurally, kinetic action against suspected sites of TTP leadership does not impact the production of AI-assisted content, which can be shifted, duplicated, or rerouted at little or no capital cost.

Bangladesh: Urban Radicalization, Regulatory Credibility, and The AI Vernacular Turn

Bangladesh is in a unique position in the South Asian counter-extremism framework. The two main Militant Islamist groups of current interest are Ansar al-Islam, a longstanding affiliate of al-Qaeda in the Indian Subcontinent (AQIS) that has developed from previous Ansarullah Bangla Team networks, and a remnant Neo-JMB network that was once affiliated with the Islamic State. Urban, educated youth have been a key target group for recruitment in the Bangladeshi extremism scholarship. According to (Khan & Roul, 2023), recruitment is "primarily from the urban youths of middle and upper class economic status." The International Crisis Group (2018) report also noted that Ansar "recruited urban, educated youth, albeit in relatively small numbers, and perpetrated brutal attacks on secular activists and bloggers."

The nature of the digital propaganda of Ansar al-Islam in 2024–2025 is a typical example of the second generation. Asif Adnan's rise to political prominence after the fall of the Hasina government in 2024 is a case in point, as described by Bashar (2025): “new generation of urban and tech-savvy extremists with mainstream academic backgrounds” (para. 7). Adnan, who Bashar (2025) notes has historical ties to Ansarullah Bangla Team's spiritual leader, Jasimuddin Rahmani, and to other AQIS networks, “uses digital platforms such as Facebook, YouTube, Instagram and TikTok to disseminate his views” (para. 3). His Facebook page, which has “over 308,000 followers,” serves as the main channel for dissemination, which is then amplified by Bengali-language extremist publishers such as SEAN Publication and through “viral TikTok clips, often shared by Adnan’s followers” (Bashar, 2025, para. 12). A lecture, published in a November 2024 SEAN Publication, called for “revolutionary jihad and violent upheaval of the global order” (Bashar, 2025, para. 14).

What this evidence shows, with regard to the present argument, is that the Bangladeshi case has the platform-distribution and vernacular-language characteristics of the second-generation threat, but not yet the public evidence of AI-generated content production on the scale of the TTP case. Unlike Urdu militant content in Pakistan, Bengali-language content is much more difficult to differentiate from legitimate political-religious commentary as it flows through the educated urban social media spaces of Dhaka and Chittagong. The public sphere of religious commentary in Bangladesh, which has a long tradition of waaz mahfil oratory, a history of mobilization by the Hefazat-e-Islam and a history of accommodating conservative religious discourse in mainstream public forums, creates a context in which subtler extremist framing can slip through the cracks of moderation that more overt militant material would not.

This issue is exacerbated by the regulatory structure. The Cyber Security Act 2023, which supersedes the Digital Security Act 2018, largely carries over the same speech offences of the previous law (Transparency International Bangladesh, 2023). According to Amnesty International's (2024) review, the CSA “reiterates nearly all the repressive elements of the Digital Security Act 2018, which was repealed” (p. 3). The empirical trend of deployment, as reflected in various human rights reports, is highly concentrated on journalists, activists and opposition voices. There are two analytical implications. The first is regulatory credibility: If a counter-extremism law is used empirically against speech, not militant content, then it loses its credibility as a counter-extremism law with practical consequences for its operational use. The second is enforcement bandwidth: the state's resources for investigating political speech cases are limited; the more they spend on political speech, the less they have to devote to the more difficult analytical task of distinguishing extremist speech from legitimate religious speech in vernacular content.

The platform side issue is equally as bad. The Facebook Papers documentation indicates that Meta's moderation capacity in Bengali has been significantly lower than its English and major European-language capacities; Duffy (2021) reviewed leaked internal documents and found that the company only used AI moderation algorithms in forty languages, and Bengali was not included in the dedicated AI moderation tier. The presence of ByteDance in Bangladesh, and its size and expansion, is complicated by algorithmic distribution, as Berger (2018) highlights, which optimizes for engagement rather than evaluating content for political content.

In part, the relative lack of public evidence of AI-generated Bengali extremist content may be due to a documentation gap rather than an empirical absence. The research infrastructure dedicated to

cataloguing TTP and Islamic State Khorasan content has put less investment in monitoring in Bengali language than in Urdu, Arabic or English. The cautious conclusion, distinguishing the empirical from the normative, is that the Bangladeshi environment shows all the ingredients of the second generation threat configuration: the urban educated population that AI-assisted content would most effectively target, the platform infrastructure through which it would flow, and the linguistic and moderation gaps that would allow it to slip past, while at present showing less direct evidence of AI-assisted production than the Pakistan case.

India's Encryption Problem: WhatsApp, Communal Incitement, And The Traceability Impasse

The second generation threat is in its most architecturally significant configuration in the Indian case. India has approximately 532 million monthly active users of WhatsApp, which is by far the largest in the country. With WhatsApp dominating digital communication in India, the technical landscape for any enforcement action against extremism is determined by the end-to-end encryption model of WhatsApp. The Indian Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules 2021 have introduced the concept of 'traceability' as the key regulatory requirement, and the current impasse has come to define the boundaries of the current governance framework (Government of India, 2021).

Under Rule 4(2) of the Intermediary Guidelines, significant social media intermediaries that offer messaging services must "enable the identification of the first originator of the information" on a court order (Government of India, 2021). The legal battle in the Delhi High Court has been based on WhatsApp's argument that complying would necessitate a fundamental change in the architecture, which is incompatible with the encryption model. The company's pleading stated that the requirement would make it "build the ability to identify the first originator for every message sent in India on its platform upon request by the government forever" (Thapliyal, 2021). The Internet Society's technical brief comes to a categorical conclusion: traceability, via escrow or fingerprinting, compromises end-to-end protection for all users ("Traceability in end-to-end," 2024).

This impasse is not directly relevant to the present argument, but for that reason more consequential. The social conditions for radicalization pathways are created by the dominant distribution vector for the broader category of communal-incitement content, such as AI-generated images and videos that inflame Hindu-Muslim tensions, which is WhatsApp. Banaji and Bhat (2019) have studied the mob violence on WhatsApp and found that "users shared mis/disinformation or acted to suppress its identification out of diverse motivations, including preexisting prejudice against specific communities, belief in a particular nationalist/religious ideology, and preexisting loyalty to political parties and ministers." Their finding that "functional literacy is of little help in preventing the spread of misinformation since even technically savvy users are willing to share disinformation and misinformation as long as it aligns with their values, beliefs, and ideological convictions" has important implications for the current argument: the distribution channel is structurally protected from technical solutions.

The relationship between communal-incitement content and Militant Islamist mobilization is indirect. Berger's (2018) analytical framework is helpful here: On his account, extremism is not primarily a matter of the content of ideology, but of the construction of in-group identity in relation to out-group enmity. Communal content generated by AI can exacerbate Hindu-Muslim tensions,

fostering the affective conditions, in-group solidarity and out-group hostility, that can be leveraged by Militant Islamist recruitment. The trend is clear: communal violence has been documented to be preceded by a spike in content inciting violence on WhatsApp, and analysis of the aftermath of such violence by Indian civil society groups has consistently found AI-generated or AI-edited visuals as part of the information that fed into the violence. Banaji and Bhat (2019) find that WhatsApp, TikTok, YouTube, ShareChat, Facebook and Instagram have been "heavily implicated in the circulation of what appears to be systematic disinformation by politically motivated groups and malicious persons wishing to destabilize communities."

The IT Rules 2021 impasse of traceability is analytically significant in this context. In part, the Indian state's strategic choice of traceability is a sensible enforcement concern: in a high-volume, encrypted distribution environment, the originators of mass-circulated incitement material are the most direct point of enforcement leverage. WhatsApp's reply, that traceability is technically impossible without breaking encryption for all users, is also reasonable on its own terms. The outcome is an unresolved structural standoff, the most impactful South Asian distribution platform in a regulatory grey zone, and the root issue – AI-driven communal incitement fostering radicalization-conducive environments – deepens. WhatsApp is used by over half a billion people in India, "not just as a chat app, but as a doctor's office, a campaigning tool, and the backbone of countless small businesses and service jobs," as Brandom (2024) stated.

There are two additional observations to complete the Indian analysis. The first relates to the liability of the platform under Section 79 of the IT Act 2000, which grants safe-harbour status to intermediaries' subject to certain conditions, including due diligence and redress of grievances (Government of India, 2000). The Indian intermediary regime is the most institutionally developed in South Asia, and its institutional development has failed to yield effective capacity to combat AI-generated content. The second relates to the cross-border aspect: AI-generated communal content shared in India is, based on the evidence available, often created by accounts and networks beyond the Indian jurisdiction. In these situations, the enforcement arm of the Indian Computer Emergency Response Team is structurally constrained.

The Compounding Governance Gap: A Structural Analysis

The three cases, when considered as a whole, have a common structure that can be explicitly stated. Three elements of the governance gap can be identified, which mutually reinforce each other.

The first is a language moderation deficit. As Gillespie (2018) points out, platform moderation systems rely on training data and language-specific models that are shaped by the commercial geography of the platform's investment decisions. Moderation capacity has grown significantly since 2014 in Arabic, in response to the challenge of the Islamic State, and has always been the platform default in English. Urdu, Bengali, Pashto, Sindhi, Punjabi, Balochi, Tamil and the smaller South Asian languages have been much less invested in, and leaked documentation shows that Meta's AI moderation systems were only used in forty languages, out of over a hundred (Duffy, 2021). One operational expression of this lack is the explicit limitation of the Tech Against Terrorism TCAP framework to "official content" in named languages, excluding supporter-generated content from its scope ("TCAP Expansion Profile," 2024).). The result is that AI-generated vernacular content is widely under the radar of automated detection.

The second is absence by definition. All three national regulatory frameworks reviewed lack a definitional category for AI-generated content, or specific evidentiary processes for determining the origin of digital content in the AI-augmented world. Prosecutions must be based on provisions that are doctrinally geared towards human authorship, and there are attendant problems of intent, means rea, and identifiable defendants. In practice, orders to block AI-generated content are challenged by the fact that the generative capacity, which is in the hands of consumer models that are widely available, cannot be blocked by any sovereign authority.

The third is structural coordination failure. The second-generation threat is, by definition, decentralized and cross-border, and effective counter-response demands an equivalent degree of inter-state technical and informational exchange. What South Asia has instead is a regional architecture in which every significant bilateral relationship is encumbered by a distinct and persistent political constraint. Pakistan and India have had no functional bilateral counter-terrorism mechanism since the Composite Dialogue process collapsed in the early 2000s; the two states simultaneously accuse each other of harbouring non-state armed actors, a configuration that makes joint intelligence exchange on AI-generated militant content politically inconceivable in the near term. The Pakistan-Afghanistan relationship has deteriorated from uneasy coexistence to open military confrontation: Operation Khyber Storm on 9 October 2025 ("Pakistani, Afghan forces," 2025) marks the most direct kinetic exchange between Islamabad and Kabul on record, and any residual bilateral digital security cooperation has effectively collapsed alongside it.

The Afghan Taliban's reluctance to act against TTP media infrastructure stems not merely from incapacity but from a degree of ideological affinity that Soufan Center (2025) documents. India's engagement with Bangladesh has been operationalized at the intelligence level, but bilateral digital security cooperation remains modest in scope and is complicated by the political reconfiguration in Dhaka following the fall of the Hasina government in 2024, which brought to power a transitional administration whose relationship with Islamist civil society is still being defined. Multilateral frameworks – the SAARC Convention on Suppression of Terrorism and the UN counter-terrorism treaty architecture – have produced no operationally significant cooperation on AI-generated extremist content, and are themselves products of an era that predates generative AI as a governance challenge. The structural implication is not merely that cooperation is difficult, but that the political conditions enabling second-generation content production in Afghanistan and the political conditions preventing coordinated response to it are, in several cases, causally connected. Kinetic options, as Cronin (2020) anticipates, leave distributed information infrastructure intact.

These three structural features are not mutually exclusive, but rather work together. The language moderation gap is that content in vernacular South Asian languages is able to reach audiences before the platform moderation kicks in. The definitional absence results in sovereign enforcement being unable to prosecute content, even if it is identified, under coherent legal categories. The coordination failure implies that, even if sovereign enforcement is technically feasible, cross-border production environments are beyond the reach of sovereign enforcement. The end effect is a governance gap that is not just the sum of its parts but a multiplicative product: the weakness of each layer compounds the effects of the others.

What could realistic policy responses be? There are three directions that seem to be analytically defensible. The first is targeted investment in vernacular moderation capacity, such as creating

South Asian language models specifically tailored to detecting extremism and establishing dedicated regional reviewer teams within platforms that operate in the region. The second is definitional treatment of AI-generated content in sovereign regulatory frameworks, such as in provenance verification processes and evidentiary standards that reflect the technical realities of generative production. The third are technical, not political, cooperation mechanisms: even if state-state political cooperation is not possible, technical cooperation on shared moderation infrastructure, hash-sharing of AI-generated extremist content, and standardized provenance verification, may be possible at the platform-state technical exchange level.

CONCLUSION

The empirical record compiled in the foregoing sections leads to a provisional, but pointed, conclusion. The current state of the South Asian counter-extremism governance architecture lacks the structural ability to detect, disrupt and deter decentralized Militant Islamist communication with the help of AI. It is still based on the 2014-2018 Islamic State threat model, which assumed that production hubs could be identified, output was stylistically recognizable, distribution was in Arabic and English, and hosting could be located. The post-2019 transformation has significantly undermined each of these assumptions, for instance in the TTP case, but can be seen in all three of the cases studied here.

This analysis makes two contributions to theory. It builds on Sageman's (2004, 2008) network theory of leaderless jihad by identifying the mechanism by which generative AI radicalizes the cost structure of decentralized production. The leaderless logic becomes non-linear resilient when the marginal cost of professional quality vernacular content approaches zero, not only because more cells can produce content, but because each cell's content is stylistically dispersed, thus resisting signature-based detection. In the AI-augmented setup, the Sagemanian “bunch of guys” turns into a bunch of guys with the production power of a small studio and the linguistic reach of a multilingual broadcaster – without any additional capital investment.

The argument also builds upon Gillespie's (2018) platform governance framework in a context that Gillespie did not explicitly discuss. The South Asian configuration introduces a particular constraint: the extreme vernacular moderation capacity deficits and the regulatory pressure from governments with structurally interwoven counter-extremism and political suppression goals. This is not a departure from Gillespie's general account, but a sharpening of it: the commercial incentives, the jurisdictional ambiguities, and the technical-architectural constraints that Gillespie identified are all present in heightened form, with the added wrinkle that the state interlocutors with whom platforms must negotiate are not unambiguously oriented toward counter-extremism.

There are a few caveats to this analysis that should be noted. The empirical evidence of AI-generated extremist content in South Asia is patchy, with far more evidence of Pakistan-related TTP content than of its Bangladeshi or Indian-language counterparts. The cautious inference that the asymmetry is due to documentation gaps as well as empirical differences cannot be tested without dedicated monitoring infrastructure. The Indian case has been addressed here mainly indirectly, by means of the communal-incitement content, and the link between the two needs to be empirically specified. The analysis has been deliberately limited to South Asia, but the structural pattern found could have parallels in other regions which comparative studies would be useful to investigate.

The forward research agenda goes in three directions. The first is systematic documentation of AI-generated extremist content in South Asian vernacular languages, based on the GNET and Tech Against Terrorism infrastructure, but expanded to include Bengali, Sindhi, Punjabi, Balochi, Brahui and Tamil. The second is a comparative analysis of definitional frameworks for AI-generated content in sovereign regulatory systems, which categorically structures most effectively address the evidentiary and procedural challenges of generative production. The third is empirical assessment of technical (as opposed to political) cooperation mechanisms between platforms, states and civil society organizations in areas of high inter-state political tension.

Finally, the argument put forward here is provisional and not conclusive. What does seem analytically defensible is that the governance architecture has not, on the evidence so far, evolved at the same pace as the threat itself, and that the gap, language-deficit, definitional-absence, and coordination-failure do not self-correct. At this writing, it is an open question whether South Asian counter-extremism governance can evolve in time to meet the second-generation threat. The need for ongoing scholarly interest in the question is, on the evidence gathered here, hard to refute.

References:

Amnesty International. (2024). *Repackaging Repression: The Cyber Security Act and the continuing lawfare against dissent in Bangladesh.* Author.

Banaji, S., & Bhat, R. (2019). *WhatsApp vigilantes: An exploration of citizen reception and circulation of WhatsApp misinformation linked to mob violence in India.* London School of Economics and Political Science.

Bashar, I. (2025, Sep. 20). *Asif Adnan: Urban, young, and digital face of Islamist extremism in Bangladesh.* Militant Leadership Monitor, Jamestown Foundation.

Basit, A. (2025, Sep. 15). *Evolution, expansion and diversification: Tehreek-e-Taliban Pakistan's Umar media*. Global Network on Extremism and Technology. https://gnet-research.org

Berger, J. (2018). *Extremism.* MIT Press.

Berger, J., & Morgan, J. (2015). *The ISIS Twitter census: Defining and describing the population of ISIS supporters on Twitter.* Brookings Institution.

Brandom, R. (2024, May 2). WhatsApp could leave India over encryption battle. *Rest of World*.

Conway, M. (2017). Determining the role of the internet in violent extremism and terrorism: Six suggestions for progressing research. *Studies in Conflict and Terrorism, 40*(1), 77-98.

Criezis, M. (2024, Feb. 5). *AI caliphate: The creation of Pro-Islamic state propaganda using generative AI*. Global Network on Extremism and Technology.

Cronin, A. (2009). *How terrorism ends: Understanding the decline and demise of terrorist campaigns.* Princeton University Press.

Cronin, A. (2020). *Power to the people: How open technological innovation is arming tomorrow's terrorists.* Oxford University Press.

Duffy, C. (2021, Oct. 25). Not just Frances Haugen: Whistleblower army accuses Facebook in SEC complaints. *CNN Business*.

Farwell, J. (2014). The media strategy of ISIS. *Survival, 56*(6), 49-55.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

Hegghammer, T. (2017). *Jihadi culture: The art and social practices of militant Islamists.* Cambridge University Press.

Hoffman, B. (2008). The myth of grass-roots terrorism: Why Osama bin Laden still matters. *Foreign Affairs, 87*(3), 133-38.

International Crisis Group. (2018). *Countering jihadist militancy in Bangladesh.* Author.

Khan, A., & Roul, A. (2023). State fragility and dynamism of Islamist extremism in Bangladesh: Recruitment strategies, women and prison radicalization, and future trajectories. In *Non-Western responses to terrorism.* Springer.

Mir, A. (2024). *Pakistan Taliban's evolving social media propaganda*. Observer Research Foundation.

Pakistani, Afghan forces exchange deadly border fire: What we know so far. (2025, Oct. 12). *Al-Jazeera*.

Sageman, M. (2004). *Understanding terror networks.* University of Pennsylvania Press.

Sageman, M. (2008). *Leaderless jihad: Terror networks in the twenty-first century.* University of Pennsylvania Press.

Sayed, A. (2024). *Afghanistan-Pakistan's radical social media ecosystem: Actors, propaganda comparison and implications*. Counter Terrorist Trends and Analyses, RSIS.

Shaikh Ahmad Hassan School of Law (SAHSOL), Lahore University of Management Sciences. (2017). *The Prevention of Electronic Crimes Act 2016: An analysis*.

Siegel, D. (2024, Feb. 19). *AI Jihad: Deciphering Hamas, al-Qaeda and Islamic State's generative AI digital arsenal*. Global Network on Extremism and Technology.

TCAP expansion profile: Tehreek-e-Taliban Pakistan (TTP). (2024). *Terrorist Content Analytics Platform*. https://terroristcontent.org

Tensions between Afghanistan and Pakistan destabilize South Asia. (2025, Nov. 21). Soufan Center: https://thesoufancenter.org

Thapliyal, N. (2021, May 26). *Traceability rule will break end-to-end encryption; Can put privacy of journalists, activists, politicians at risk: WhatsApp tells Delhi high court*. *LiveLaw*.

Traceability in end-to-end encrypted environments: Technical and socio-economic impacts of implementing traceability in India. (2024, Aug. 12). *Internet Society.*

Transparency International Bangladesh. (2023). *Digital Security Act 2018 and the Draft Cyber Security Act 2023: A comparative analysis.* Author.

Weimann, G. (2006). *Terror on the internet: The new arena, the new challenges.* United States Institute of Peace Press.

Winter, C. (2015). *The virtual 'Caliphate': Understanding Islamic State's propaganda strategy.* Quilliam Foundation.

Winter, C. (2015, Oct. 15). *More than just beheadings: How the Islamic State sells itself*. War on the Rocks.

Date of Publication	July 10, 2026