
THE ARCHITECTURE OF HATRED: IDENTITY THREAT, PSYCHOLOGICAL RADICALISATION, AND THE CONSTITUTIONAL LIMITS OF HATE-CRIME LAW IN INDIA

Animesh Singh, Assistant Professor, Law, Babu Banarasi Das University

ABSTRACT

Hate crimes in India have intensified alongside communal polarisation, digitally accelerated misinformation, and the normalisation of identity-based rhetoric in public life. Existing scholarship examines this phenomenon predominantly through constitutional, criminological, or sociological lenses, leaving the psychological mechanisms that convert prejudice into criminal violence comparatively unexamined within Indian legal literature. This paper asks why only a small subset of prejudiced individuals become violent offenders, what psychological processes drive that transformation, whether Indian criminal law recognises these dynamics, and how criminal psychology might inform prevention without compromising constitutional guarantees. Adopting a doctrinal, comparative, and interdisciplinary methodology, the paper traces a continuum from identity threat and cognitive bias through moral disengagement, dehumanisation, and group polarisation to organised or spontaneous violence, drawing on social identity theory and Moghaddam's staircase model of radicalisation. It situates this psychological account within India's constitutional framework (Articles 14, 19, 21, and 25), the Bharatiya Nyaya Sanhita's hate-speech and mob-lynching provisions, and Supreme Court jurisprudence in *Tehseen Poonawalla*, *Pravasi Bhalai Sangathan*, and *Amish Devgan*. Comparative analysis of the United Kingdom, the United States, Germany, and Canada reveals a growing, if uneven, integration of behavioural science into hate-crime prevention. The paper concludes that Indian criminal law remains reactive and behaviourally uninformed, and proposes a constitutionally compatible, psychology-informed framework for investigation, sentencing, and community-based deradicalisation that treats criminal psychology as a preventive and evidentiary tool rather than an independent ground for restricting liberty.

Keywords: Hate Crimes; Criminal Psychology; Radicalisation; Social Identity; Dehumanisation; Constitutional Criminal Law; Digital Extremism.

I. Introduction

Polarised democracies across the world are witnessing a resurgence of identity-based violence, in which victims are selected not for what they have done but for who they are. India exemplifies this pattern with particular intensity. Independent monitoring organisations recorded a marked escalation in communal violence through 2023 and 2024: one civil-society tracking initiative documented fifty-nine communal riots in 2024, compared with thirty-two in 2023, alongside at least seventeen suspected religiously motivated killings, eleven of which occurred within a month of the 2024 general election (South Asia Justice Campaign, 2025). Parallel monitoring by the Center for the Study of Organized Hate recorded 1,165 in-person hate-speech events targeting religious minorities in 2024, a 74.4 per cent increase over 2023, with national political figures playing an increasingly prominent instigating role (Center for the Study of Organized Hate, 2025). Officially, this escalation is difficult to verify: the National Crime Records Bureau discontinued discrete reporting of "communal," "lynching," and "hate crime" categories after 2017, folding such incidents into generic classifications such as "riots" and "public disorder" (Citizens for Justice and Peace, 2025a, 2025b). The gap between lived experience and official statistics is itself a governance problem, but it also underscores why an interdisciplinary account of why individuals cross from prejudice into violence is overdue.

Three developments make this inquiry urgent. India has witnessed recurrent episodes of mob violence linked to cow-vigilantism, interfaith relationships, and rumour-driven suspicion, alongside sustained ethnic conflict such as the Manipur violence of 2023. Online platforms have become a primary vector for hate propagation, enabling rapid mobilisation and the diffusion of accountability associated with anonymous or pseudonymous participation. And criminal psychology and behavioural science increasingly inform hate-crime prevention in comparative jurisdictions, yet remain marginal to Indian criminal-justice practice, which continues to respond to hate crimes primarily as *ex post facto* law-and-order failures rather than as the endpoint of an identifiable psychological trajectory.

Four bodies of literature bear on this inquiry, and their relative isolation from one another defines the gap this paper addresses. Criminal law scholarship has examined hate crimes as a distinct offence category defined by bias motivation and identity-based victim selection. Constitutional scholarship has addressed the tension between free expression under Article 19(1)(a) and the state's interest in preventing group-targeted hatred, particularly through the

Supreme Court's evolving hate-speech jurisprudence. Criminal psychology has developed a rich account of prejudice, in-group favouritism, and violence rooted in social identity theory (Tajfel & Turner, 1979), moral disengagement (Bandura, 1999), and dehumanisation research. Radicalisation and behavioural science literature, exemplified by Moghaddam's (2005) staircase model, explains how a minority of individuals progress from grievance to organised violence, while a more recent strand addresses digital extremism and the architecture of online hate ecosystems, including algorithmic amplification and echo-chamber effects (Sunstein, 2017). These bodies of scholarship rarely intersect. Indian hate-crime scholarship examines the phenomenon predominantly through constitutional, criminological, or sociological perspectives, with limited interdisciplinary work explaining the psychological progression from prejudice to criminal violence, and comparative behavioural-science literature on radicalisation, developed largely in relation to terrorism and violent extremism in Western contexts, has not been substantially integrated into Indian hate-crime jurisprudence, which continues to treat bias-motivated violence as an evidentiary question of motive rather than as the product of a traceable psychological process.

Against this background, the paper is organised around four questions. Why do only certain prejudiced individuals become violent offenders? What psychological mechanisms transform prejudice into hate crimes? Does Indian criminal law adequately recognise these psychological dynamics? And how can criminal psychology contribute to hate-crime prevention without undermining constitutional guarantees? Its central argument is that hate crimes are the culmination of identifiable psychological processes shaped by identity threat, group dynamics, and radicalisation, and that while criminal psychology offers valuable insights for prevention, its application within criminal justice must remain consistent with constitutional principles of equality, liberty, and due process.

The paper adopts a doctrinal, comparative, and interdisciplinary methodology. Its primary focus is Indian constitutional and criminal law, with comparative reference to the United Kingdom, the United States, Germany, and Canada. It relies on judicial decisions, statutory provisions under the Bharatiya Nyaya Sanhita 2023, National Crime Records Bureau data and its documented limitations, and secondary psychological literature on prejudice, group behaviour, and radicalisation. The analysis is necessarily conceptual rather than empirical: it synthesises established psychological models with doctrinal legal analysis rather than presenting new primary data on Indian offenders.

II. Hate Crimes as a Distinct Category of Criminal Offending

The category of "hate crime" emerged in criminological and legislative usage principally in the United States during the 1980s, in response to a recognition that certain offences caused harm disproportionate to their immediate physical consequences because they targeted individuals as representatives of a group. Legislative recognition spread unevenly: the United States adopted federal hate-crime statutes culminating in the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act of 2009 (Cornell Law School Legal Information Institute, n.d.), which extended federal jurisdiction to violence motivated by actual or perceived race, colour, religion, national origin, gender, sexual orientation, gender identity, or disability. The United Kingdom developed a parallel framework through aggravated offences and sentencing enhancements under what is now the Sentencing Act 2020 (House of Commons Library, 2026). India has no comparable standalone hate-crime statute; bias-motivated violence is addressed through general provisions on murder, rioting, and promoting enmity, supplemented since 2023 by a specific mob-lynching provision discussed in Part V.

Across jurisdictions, three elements recur in defining hate crimes: bias motivation, meaning that the offender selects the victim, wholly or partly, because of an actual or perceived protected characteristic; identity-based victim selection, distinguishing hate crimes from crimes in which identity is incidental; and reliance on protected characteristics such as race, religion, caste, ethnicity, sex, or sexual orientation. A persistent doctrinal difficulty lies in distinguishing motive, traditionally irrelevant to criminal liability, from the specific intent or animus that hate-crime provisions require courts to establish, often through circumstantial evidence such as slurs, symbols, or the selection pattern of victims.

Hate crimes are best understood not as a discrete phenomenon but as the visible endpoint of a continuum: prejudice, an internal attitude of negative evaluation toward an out-group; stereotyping, the cognitive simplification that sustains prejudice; discrimination, the behavioural enactment of prejudice in denying opportunities or resources; hate speech, the verbal or symbolic expression of hostility; hate crimes, physical or property-directed violence motivated by bias; and, at the extreme end, organised or ideologically sustained violent extremism. Gordon Allport's (1954) classic account of the "ladder of prejudice" anticipated this continuum, and it remains analytically useful because it demonstrates that criminal violence rarely erupts without antecedent, socially tolerated stages of stereotyping and discrimination.

Hate crimes inflict a dual injury. The individual victim suffers the immediate physical or psychological harm common to any offence. But because the offence communicates a message — that members of a particular identity group are unwelcome, unsafe, or unequal — it also inflicts collective victimisation on the wider community sharing that identity, producing fear, withdrawal from public life, and diminished trust in the state's protective capacity. This symbolic and communicative dimension gives hate crimes a distinctively democratic and constitutional harm: they attack the premise of equal citizenship that underlies Article 14 and the equal protection of laws.

Several structural features complicate the legal response. Under-reporting is endemic, both because victims from marginalised communities distrust law enforcement and because investigating officers frequently decline to record the bias element of an offence. Evidentiary difficulties in proving bias motivation are compounded where offences are prosecuted under generic provisions that do not require proof of animus. There is substantial overlap between communal violence, hate speech, and hate crimes that resists tidy legal categorisation. Finally, digitally facilitated hate offending — coordinated harassment campaigns, doxxing, and online incitement preceding offline violence — increasingly blurs the line between speech offences and violent crime. Together, these features explain why hate crimes differ from ordinary criminal offences: because they are directed against identity groups, they produce harms that extend beyond individual victims to communities and constitutional democracy, and any legal response calibrated only to the individual offence will systematically understate the injury involved.

III. Criminal Psychology of Hate Crimes: From Identity Threat to Violent Radicalisation

Henri Tajfel and John Turner's (1979) social identity theory remains the foundational framework for understanding intergroup hostility. The theory holds that individuals derive part of their self-concept from membership in social groups, and that this membership generates a motivational bias toward favouring the in-group and derogating out-groups, even under conditions of minimal or arbitrary group assignment. In-group and out-group formation occurs rapidly and often automatically, structuring perception along categorical lines well before any material conflict of interest arises. Identity threat — a perceived challenge to the status, distinctiveness, or legitimacy of one's group — heightens in-group solidarity and out-group

hostility, particularly where the threat is framed as existential or civilisational rather than merely material, and individuals and groups respond with defensive cognitive and behavioural strategies, ranging from benign in-group affirmation to hostile action against the perceived source of threat, when identity is framed as under siege.

A further theoretical resource, complementary to social identity theory, is realistic conflict theory, which locates intergroup hostility in genuine or perceived competition over scarce material resources such as land, employment, or political representation. In the Indian context, communal violence frequently combines identity-based and resource-based grievance, with land disputes, reservation politics, or competition over municipal resources providing the material substrate onto which identity-threat narratives are subsequently mapped. Terror management theory offers a related insight: reminders of mortality, whether from communal conflict, pandemic disruption, or economic precarity, have been shown experimentally to intensify defensive affirmation of one's own cultural worldview and corresponding hostility toward those perceived to challenge it, a mechanism plausibly implicated in the correlation between periods of acute social anxiety and spikes in identity-based violence.

Several well-documented cognitive biases sustain and amplify intergroup hostility once identity threat has been triggered. The out-group homogeneity effect leads individuals to perceive out-group members as more similar to one another, and more predictably threatening, than in-group members, facilitating collective attribution of blame. Confirmation bias causes individuals to selectively notice and recall information consistent with pre-existing stereotypes while discounting disconfirming evidence. Attribution bias, particularly the fundamental attribution error, leads observers to attribute out-group members' negative conduct to disposition or character while attributing similar in-group conduct to circumstance. Scapegoating channels diffuse economic, social, or political grievance onto an identifiable and relatively powerless out-group, providing a psychologically satisfying, if factually unfounded, explanation for collective hardship.

Group processes intensify individual-level bias. Group polarisation describes the well-established tendency for group discussion to shift individual attitudes toward more extreme positions than those held privately by members before deliberation, particularly where the group is ideologically homogeneous. Mob psychology, associated with classical crowd theory and its modern refinements, describes the loss of individuated moral restraint that can occur in

large, anonymous gatherings, facilitating collective participation in violence that few members would undertake individually. Echo chambers, both offline and online, restrict exposure to counter-attitudinal information, entrenching polarised beliefs, while collective decision-making in radicalised groups often follows a spiral of escalating commitment, in which each member's prior statements and actions create social pressure toward consistency, making de-escalation increasingly costly.

Albert Bandura's (1999) theory of moral disengagement identifies the cognitive mechanisms by which individuals who otherwise hold conventional moral standards come to perform or condone violent acts without experiencing the self-sanctions that would normally inhibit such conduct. Moral justification reframes violence as serving a higher social or religious purpose. Euphemistic labelling substitutes sanitised language — "cleansing," "correction," "self-defence" — for the violent reality of the act. Diffusion of responsibility, particularly acute in mob settings, allows each participant to regard their individual contribution as causally negligible, and victim blaming, the attribution of provocation or culpability to the victim, completes the disengagement by inverting the moral relationship between offender and victim.

Dehumanisation — the psychological process by which out-group members are perceived and represented as lacking full human status, often through animalistic or inhuman metaphor — removes the ordinary moral inhibitions against violence. Historical manifestations of this mechanism are well documented in genocide studies and in the rhetoric preceding mass communal violence, where target communities are habitually described using vermin, disease, or animal imagery. Digital dehumanisation reproduces and accelerates this pattern: hate-speech monitoring in India has recorded increasingly explicit dehumanising and eliminationist rhetoric, including calls for economic boycott and property destruction directed at religious minorities, disseminated through both traditional and social media (Center for the Study of Organized Hate, 2025).

Fathali Moghaddam's (2005) staircase model provides the most widely cited framework for understanding how a small minority of individuals, out of a much larger population holding grievance-based attitudes, progress toward violence. The model conceives radicalisation as a narrowing staircase: the ground floor, comprising a large population, experiences perceived injustice and relative deprivation but takes no violent action; successive floors involve the search for legitimate avenues of redress, the displacement of aggression onto a perceived

enemy, moral engagement with an extremist organisation, categorical "us versus them" thinking reinforced by isolation from moderate influences, and, at the narrowest point, the sidestepping of inhibitory mechanisms that permits the final violent act (Moghaddam, 2005). As individuals ascend, they perceive progressively fewer alternatives, and the psychological and social cost of disengagement rises correspondingly. Subsequent empirical review has qualified the model, finding that while its constituent psychological processes are individually well supported, the strictly linear, sequential character of the staircase receives weaker empirical confirmation, suggesting that radicalisation pathways are better understood as combinatorial rather than uniformly stepwise (Lygre et al., 2011). Collective narcissism — an inflated, ideologically invested view of one's group's importance, combined with a sense that this importance is insufficiently recognised — the need for belonging that radical or extremist collectives readily supply, and uncertainty reduction, whereby rigid ideological frameworks resolve the psychological discomfort of ambiguity, are recurrent facilitating conditions across the staircase.

It is equally important, for a criminal-psychological account intended to inform legal reform, to specify why the overwhelming majority of prejudiced individuals never commit hate crimes. Three protective and inhibitory factors merit note. Cross-group contact under conditions of equal status, common goals, and institutional support — the classical contact hypothesis — measurably reduces prejudice and is inversely correlated with participation in identity-based violence, suggesting that residential and institutional segregation is itself a risk factor for radicalisation rather than a neutral backdrop to it. Individual differences in moral identity centrality — the degree to which an individual's self-concept is organised around moral commitments — moderate susceptibility to the moral-disengagement mechanisms described above; individuals with high moral identity centrality are demonstrably more resistant to euphemistic labelling and diffusion-of-responsibility framing. And the presence of cross-cutting social ties, meaning relationships and group memberships that intersect rather than reinforce group boundaries, disrupts the categorical "us versus them" thinking that the staircase model identifies as characteristic of its higher floors. These protective factors matter for the framework proposed in Part VIII, since preventive intervention is most plausibly effective where it strengthens exactly these three conditions rather than relying solely on deterrence through criminal sanction. Taken together, this literature shows that hate crimes emerge through identifiable psychological mechanisms that progressively transform prejudice into organised or individual acts of violence, and that the same mechanisms which explain

radicalisation also identify the protective conditions capable of interrupting it.

IV. Digital Radicalisation and the Psychology of Online Hate

Contemporary hate-crime psychology cannot be separated from the architecture of digital platforms. Recommendation systems are optimised for engagement rather than accuracy or civility, and emotionally activating content — including outrage, fear, and moral condemnation of out-groups — reliably generates higher engagement than neutral content, so that engagement-driven content amplification systematically privileges polarising material, independent of any deliberate ideological design by the platform.

Algorithmic personalisation compounds the pre-existing psychological tendency toward confirmation bias, producing ideological reinforcement in which users are increasingly exposed to attitude-consistent content. Cass Sunstein's (2017) work on group polarisation online demonstrates that homogeneous digital information silos accelerate the same polarisation dynamics observed in offline group deliberation, but at greater speed and scale, since online groups can form and radicalise across geographic distance in ways offline communities cannot. Anonymity and pseudonymity further reduce the accountability that ordinarily constrains public expression of hostility, facilitating the euphemistic labelling and diffusion of responsibility identified above as core moral-disengagement mechanisms (Bandura, 1999); coordinated hate campaigns exploit this diffusion, since no single participant in a pile-on or a viral hashtag campaign experiences themselves as solely responsible for its cumulative harm.

Fear contagion — the rapid, affect-driven spread of threat-related misinformation — and rumour propagation have repeatedly preceded documented incidents of mob violence in India, where unverified claims of child abduction, cattle smuggling, or religious desecration circulated on messaging platforms have triggered lethal attacks before law enforcement could intervene. Digital mobilisation converts online outrage into offline assembly with a speed that outpaces traditional preventive policing, which was designed around slower-moving forms of rumour transmission.

Indian conditions give this general dynamic a particular local inflection. The scale of encrypted messaging use — India remains the largest market for platforms such as WhatsApp — means that much identity-hostile content circulates outside the reach of the public-content moderation

tools on which platform-accountability debates typically focus, since encrypted peer-to-peer forwarding is invisible to both researchers and, absent lawful interception, regulators. This creates a structural mismatch between the regulatory architecture built around public social-media content and the actual vector through which much rumour-driven mob violence has historically been triggered in India. Forwarded-message virality functions psychologically as a trust heuristic: messages arriving through known contacts or community groups inherit the perceived credibility of the sender, short-circuiting the scepticism that identical content might receive from an unknown public account, and thereby accelerating the fear-contagion dynamic described above.

The interaction between human psychological vulnerability and platform architecture is best understood as multiplicative rather than additive: digital platforms do not introduce novel psychological mechanisms so much as they intensify and accelerate pre-existing ones — identity threat, confirmation bias, moral disengagement, and dehumanisation — by removing the natural friction that offline social life imposes on the spread of hostile material. Digital platforms therefore function as accelerants rather than independent causes of radicalisation, a distinction with direct regulatory implications: platform-accountability measures address a contributing structural condition, not the psychological root of hate offending, and the specific architecture of encrypted, high-penetration messaging in India further complicates any conventional platform-accountability response.

V. Indian Criminal Law and the Psychological Blind Spot

India's constitutional architecture provides both the normative foundation for combating hate crimes and the outer limits on how that combat may be pursued. Article 14 guarantees equality before the law and equal protection of the laws, providing the constitutional predicate for treating bias-motivated offences as a distinct category of harm to constitutional equality. Article 19 protects freedom of speech and expression, subject to reasonable restrictions including those relating to public order and incitement, creating the central doctrinal tension in hate-speech regulation. Article 21 guarantees the right to life and personal liberty, encompassing dignity, and has been the textual anchor for judicial intervention against mob violence. Article 25 guarantees freedom of conscience and religion, implicated directly where hate crimes target religious identity or practice.

The Bharatiya Nyaya Sanhita 2023 (BNS), which replaced the Indian Penal Code with effect

from 1 July 2024, restructured but did not fundamentally reconceive India's approach to bias-motivated offending. Section 196 substantially reproduces former Section 153A of the Indian Penal Code, criminalising the promotion of enmity between groups on grounds of religion, race, place of birth, residence, or language, including through electronic communication, with a maximum sentence of three years (Granthaalayah Publication, 2024; Devgan.in, n.d.). Section 197 reproduces former Section 153B, addressing imputations and assertions prejudicial to national integration. Critically, Section 103(2) introduces, for the first time in Indian criminal law, an explicit mob-lynching provision: where five or more persons acting in concert commit murder on grounds of race, caste, community, sex, place of birth, language, personal belief, or any similar ground, each participant is liable to death or life imprisonment (Legallyin, 2024). This provision directly operationalises the Supreme Court's earlier call, in *Tehseen Poonawalla*, for a distinct statutory offence addressing collective, identity-motivated killing. Notably, however, neither Section 196 nor Section 197 defines "hate speech," and commentators have observed that the BNS reproduces the interpretive uncertainty of its colonial-era predecessor provisions rather than resolving it (Citizens for Justice and Peace, 2024).

Three decisions frame the Supreme Court's engagement with identity-based violence. In *Tehseen S. Poonawalla v. Union of India* (2018), the Court, per Chief Justice Dipak Misra, held that no individual or mob may take the law into its own hands, and issued preventive, remedial, and punitive directions binding on states under Article 141: the appointment of district nodal officers to prevent lynching and cow vigilantism, automatic registration of first information reports against those inciting violence or spreading provocative material on social media, victim compensation schemes under Section 357A of the Code of Criminal Procedure, and fast-track trials to be completed within six months (Supreme Court Observer, 2023; CaseMine, 2018). The Court explicitly recommended that Parliament create a separate offence of lynching — a recommendation partially realised six years later through BNS Section 103(2). In *Pravasi Bhalai Sangathan v. Union of India* (2014) 11 SCC 477, the Court, while declining to issue fresh guidelines on hate speech, examined the adequacy of existing Indian Penal Code provisions in combating hate-motivated mob violence and urged legislative attention to the gap (CaseMine, 2018). In *Amish Devgan v. Union of India* (2020) SCC OnLine SC 994, the Court undertook its most extensive doctrinal engagement with the concept of hate speech, articulating a content-context-intent-harm framework and holding that "persons of influence," given their reach and impact, owe a heightened duty of responsible speech (Global Freedom of Expression,

2023; SCC Online Blog, 2020). The Court observed that hate speech "repudiates the right to equality" in a polity committed to pluralism, tying the doctrinal analysis of speech restriction directly to Article 14 (LiveLaw, 2020). Critics note, however, that the judgment's reliance on a "reasonable, strong-minded, firm and courageous" person standard leaves the boundary between protected and unprotected speech persistently indeterminate (The Leaflet, 2020).

Beyond judicial directions, institutional responses remain uneven. Preventive policing measures mandated by *Tehseen Poonawalla* — nodal officers, social-media monitoring, and sensitive-district identification — have been only partially and inconsistently implemented across states, as subsequent litigation before the Allahabad High Court and continuing compliance petitions before the Supreme Court illustrate (LiveLaw, 2025). Investigation of bias motivation remains largely untrained and ad hoc; police officers are rarely instructed in recognising or documenting the psychological indicia of hate-motivated offending, such as pre-offence rhetoric, target selection patterns, or symbolic conduct. Existing enforcement challenges are compounded by delayed and, per independent monitors, systematically undercounted official crime statistics, which weaken both accountability and evidence-based policy design (Citizens for Justice and Peace, 2025a).

The doctrinal architecture surveyed above is overwhelmingly reactive. Constitutional adjudication addresses hate crimes after they occur, through remedial compensation and post hoc sentencing; statutory provisions criminalise the completed act or the completed utterance; and judicial guidelines direct institutional response once an offence is imminent or has occurred. What is absent, across constitutional doctrine, statutory drafting, and institutional practice alike, is any explicit engagement with the antecedent psychological process — identity threat, moral disengagement, dehumanisation, and radicalisation — traced in Part III. Indian criminal law neither trains investigators to recognise early behavioural indicators of impending hate-motivated violence, nor incorporates psychological expertise into sentencing determinations regarding the depth and durability of an offender's bias, nor structures preventive policing around the digital vectors of radicalisation examined in Part IV: it primarily responds to the consequences of hate crimes while largely overlooking the behavioural processes that precede them.

VI. Comparative Perspectives on Behaviourally Informed Hate Crime Prevention

The United Kingdom has developed the most institutionally elaborate hate-crime framework

among the jurisdictions surveyed. Its statutory architecture combines aggravated offences and sentencing enhancements — currently under Section 66 of the Sentencing Act 2020, permitting courts to impose more severe penalties within existing maxima where hostility based on race, religion, sexual orientation, disability, or transgender identity is proved (Equality and Human Rights Commission, 2025) — with successive Hate Crime Action Plans since 2016 and a refreshed Hate Crime Plan for 2024–2028 emphasising education, community engagement, and improved victim support (Greater Manchester Combined Authority, 2024). The separate Prevent strategy, targeting radicalisation toward terrorism rather than hate crime as such, illustrates both the potential and the limits of behavioural risk assessment: recent critical scholarship argues that Prevent's vulnerability-based assessment tools generate significant false negatives relative to threat-based models such as the Pathway to Violence Model, a gap starkly illustrated by post-incident review of the 2024 Southport attack (ResearchGate, 2026). The Law Commission's ongoing review of hate-crime legislation, including contested proposals to extend protected characteristics, further illustrates the UK's continuing recalibration of the sentencing-based model (Law Commission, 2025).

The United States federal framework, centred on the Matthew Shepard and James Byrd Jr. Hate Crimes Prevention Act of 2009, expanded federal jurisdiction over bias-motivated violence and removed a prior requirement that the offence interfere with a federally protected activity, extending protection to actual or perceived gender, sexual orientation, gender identity, and disability alongside race, colour, religion, and national origin (Human Rights Campaign, n.d.; Cornell Law School Legal Information Institute, n.d.). The statute embodies a community-impact rationale: legislative history explicitly recognises that bias-motivated violence terrorises entire communities sharing the victim's identity, not merely the immediate victim (Anti-Defamation League, n.d.). Federal grant programmes, including the Bureau of Justice Assistance's hate-crimes programme, fund state and local investigation and prosecution capacity, reflecting a cooperative-federalism model of implementation (Bureau of Justice Assistance, 2024).

Germany's approach centres on tertiary and secondary prevention through counter-extremism and deradicalisation infrastructure rather than solely on enhanced sentencing. The EXIT-Germany programme, operating since 2000, and the Federal Office for the Protection of the Constitution's exit-hotline infrastructure represent one of the world's longest-running state-supported deradicalisation efforts, targeting individuals seeking to leave right-wing extremist

movements (Radicalisation Awareness Network, European Commission, n.d.). German practice distinguishes primary prevention (community resilience and education), secondary prevention (targeted intervention with at-risk individuals before offending), and tertiary prevention (working with convicted or fully radicalised individuals to prevent recidivism), a public-health-derived taxonomy that has influenced comparative programme design internationally for countering violent extremism (Koehler, 2021). Evaluation research, however, cautions against overstating programme efficacy: a systematic review of seventy-four such programme evaluations between 2001 and 2020 found only 32 per cent deemed successful, with 55 per cent showing limited success and some programmes producing negative outcomes such as community stigma (ScienceDirect, 2024).

Canadian practice similarly emphasises rehabilitation and behavioural intervention alongside enhanced sentencing for hate-motivated offences, integrating community-based case-managed intervention models. Comparative evaluation literature, including the Proactive Integrated Support Model, has sought to formalise outcome measurement for individually tailored disengagement interventions, addressing the broader methodological difficulty — evident across jurisdictions — of demonstrating causal programme impact on recidivism to bias-motivated offending (National Institute of Justice, 2024).

These comparative experiences yield three lessons for Indian reform. Adaptable practices exist at every stage of the psychological continuum traced in Part III: sentencing enhancement addresses the completed offence, community-impact statements address the collective harm, and tertiary deradicalisation addresses recidivism risk among identified offenders. Constitutional compatibility nonetheless requires care, since the UK's Prevent experience demonstrates that behavioural risk-assessment tools carry real error costs, cautioning against uncritical importation into a jurisdiction, like India, where preventive detention and profiling already raise acute constitutional concern under Articles 14 and 21. And institutional feasibility in India is constrained by capacity: unlike Germany's decades-old, well-resourced deradicalisation infrastructure, India currently lacks any dedicated institutional mechanism for pre-offence behavioural intervention, meaning reform must be sequenced rather than transplanted wholesale.

VII. Existing Lacunae in Indian Hate-Crime Jurisprudence

Legislative gaps persist despite BNS Section 103(2): India still has no comprehensive,

freestanding hate-crime statute defining bias motivation as an aggravating factor across the full range of offences, unlike the UK's aggravated-offence model or the US's federal framework. Hate speech remains undefined in the BNS itself, reproducing rather than resolving decades of interpretive uncertainty inherited from Sections 153A and 153B of the former Indian Penal Code (Citizens for Justice and Peace, 2024), and no parallel hate-crime data collection obligation accompanies these substantive provisions, perpetuating the statistical invisibility documented in Part I.

Psychological gaps are pervasive: no statutory or procedural provision requires or even authorises psychological assessment of bias depth or radicalisation risk at any stage of investigation, prosecution, or sentencing. Forensic psychological expertise, routinely used in Indian courts to assess fitness to stand trial or diminished responsibility, has no established role in assessing the ideological dimension of bias-motivated offending. Investigative gaps compound this absence: police training curricula do not incorporate behavioural indicators of impending hate-motivated violence, such as pre-offence rhetoric escalation, symbolic target selection, or online mobilisation patterns, despite the *Tehseen Poonawalla* directions contemplating social-media monitoring, and standard operating procedures issued by the Ministry of Home Affairs address post-incident response but say little about pre-incident behavioural triage.

Digital governance gaps remain acute: platform accountability under India's intermediary-liability framework addresses illegal content removal but does not engage algorithmic amplification design, the structural driver identified in Part IV as the principal digital accelerant of radicalisation, nor does it address the encrypted-messaging vector through which much rumour-driven mob violence has historically propagated. Sentencing gaps persist because courts lack structured guidance for weighing ideological motivation, radicalisation depth, or genuine remorse and disengagement potential in sentencing bias-motivated offenders, unlike the UK's Section 66 framework, which at least formalises the hostility-based sentencing uplift.

Preventive justice gaps are perhaps the most consequential: India has no tertiary or secondary intervention infrastructure comparable to Germany's EXIT programme or Canada's case-managed disengagement models, meaning that even identified at-risk or previously convicted individuals have no structured pathway toward disengagement and reintegration. This absence leaves the psychological continuum traced in Part III entirely unaddressed between the point

of initial radicalisation and the point of criminal conviction, and it is this cumulative absence of an integrated legal and psychological framework, more than any single statutory gap, that continues to limit the effectiveness of India's response to hate crimes.

VIII. Towards a Constitutionally Compatible Psychology-Informed Framework

Reform should begin at the investigative stage, where the psychological continuum identified in Part III is most amenable to early intervention. Behavioural science training for investigating officers should cover the recognition of moral-disengagement rhetoric, dehumanising language, and group-polarisation indicators as evidentiary markers of bias motivation, enabling more accurate charging decisions under BNS Sections 103(2), 196, and 197. Scientific assessment of bias motivation, drawing on forensic psychological methodology already used in comparative sentencing practice, could assist courts in distinguishing spontaneous, opportunistic violence from ideologically sustained offending, without displacing the ordinary burden and standard of criminal proof.

Sentencing frameworks should formally recognise ideological motivation as an aggravating factor, following the UK's Section 66 model, while incorporating trauma-informed approaches to victim participation that recognise the collective, community-level harm identified in Part II. This requires legislative amendment rather than judicial improvisation, given the separation-of-powers concerns implicit in courts fashioning sentencing enhancements absent explicit statutory authorisation.

Drawing cautiously on the German and Canadian models, India should pilot community-engagement programmes addressing at-risk individuals and groups before offending occurs, alongside deradicalisation programmes for convicted offenders willing to engage voluntarily, and counter-radicalisation initiatives targeting the specific rumour-and-misinformation vectors documented in Part IV. Given the mixed evaluation record of comparative programming to counter violent extremism, any Indian pilot should be designed with rigorous, independent outcome evaluation built in from inception rather than added retrospectively.

Digital literacy programmes addressing rumour verification and platform-manipulation awareness should be integrated into the preventive-policing framework contemplated by *Tehseen Poonawalla*. Platform accountability measures should extend beyond illegal-content takedown to require transparency regarding algorithmic amplification of high-engagement,

identity-hostile content, while algorithmic transparency obligations should be calibrated to avoid displacing accountability onto private intermediaries in ways that create incentives for over-removal of legitimate speech.

Any such framework, however, must operate within firm constitutional limits. Article 14 and the guarantee of non-discrimination require that behavioural risk indicators not become a proxy for profiling particular religious or ethnic communities as inherently more prone to radicalisation — a risk starkly illustrated by the UK Prevent experience, and one with particular resonance in the Indian context given the documented pattern, noted in Part I, of disproportionate criminal-justice targeting of minority and marginalised communities. Article 21 and the guarantee of due process require that psychological assessment inform, but never substitute for, proof beyond reasonable doubt of the statutory elements of an offence. The presumption of innocence must remain undiluted: risk indicators may justify enhanced investigative attention, but never a lowered evidentiary threshold for conviction. Judicial oversight of any behavioural-risk-assessment tool, including periodic review of its accuracy and disparate impact, is essential, following the cautionary lesson of Prevent's documented false-negative and, implicitly, false-positive rates (ResearchGate, 2026), and proportionality must govern the relationship between the individualised threat a person poses and the intrusiveness of any preventive measure applied to them.

Two further institutional design principles follow from this constitutional analysis. Any behavioural-training or risk-assessment initiative should be piloted under an independent evaluation mandate from inception, with published, disaggregated outcome data, precisely because the comparative evidence surveyed in Part VI shows how easily such programming can produce unmeasured or counterproductive effects, including community stigma, without rigorous external scrutiny. And the institutional home for any such initiative matters constitutionally: locating behavioural expertise within the ordinary criminal investigation and victim-support apparatus, subject to judicial review at each stage, is markedly more consistent with due-process norms than locating it within a parallel, preventive-security apparatus of the kind that has attracted sustained constitutional criticism in other contexts of Indian counter-terrorism law. A psychology-informed framework that migrates hate-crime prevention from the ordinary criminal process into an executive-controlled preventive-detention regime would replicate, rather than solve, the psychological blind spot identified in Part V, by substituting one form of reactive, poorly targeted intervention for another. Criminal psychology, in short,

should function as a preventive and investigative tool rather than an independent basis for restricting liberty.

IX. Conclusion

This paper has traced a continuum running from identity threat and cognitive bias, through moral disengagement and dehumanisation, to organised or spontaneous radicalisation, and has shown that this continuum is intensified, though not created, by digital platform architecture. It has demonstrated that Indian constitutional and criminal law — from the BNS's new mob-lynching provision to the Supreme Court's jurisprudence in *Tehseen Poonawalla*, *Pravasi Bhalai Sangathan*, and *Amish Devgan* — remains structured around the reactive consequences of hate-motivated violence rather than its antecedent psychological process, leaving a persistent blind spot in investigation, sentencing, and prevention alike.

Hate crimes are the culmination of identifiable psychological processes, and criminal psychology can meaningfully inform their prevention, but only where its application is constrained by the constitutional principles of equality, liberty, and due process that distinguish a rights-respecting criminal justice system from a preventive security state. The theoretical contribution of this analysis lies in mapping social identity theory, moral-disengagement theory, dehumanisation research, and the staircase model of radicalisation onto the specific doctrinal architecture of Indian hate-crime law — an integration largely absent from existing Indian legal scholarship. Its policy implications point toward behaviourally informed investigation, psychology-aware sentencing, cautiously designed community-based deradicalisation pilots, and digital-governance reform targeting algorithmic amplification, each bounded by explicit constitutional safeguards against profiling and evidentiary dilution.

Future research should pursue empirical study of Indian hate-crime offender pathways specifically, rather than relying, as this paper has, on models developed principally in Western contexts; should evaluate any pilot deradicalisation or behavioural-training programme with the rigour that comparative evaluation literature shows has too often been absent; and should examine the disparate-impact risks of behavioural risk-assessment tools within India's specific communal and constitutional context. Hate crimes, in the end, cannot be effectively addressed through punitive criminal law alone. A constitutionally compliant response requires recognising hatred as a dynamic psychological process while ensuring that behavioural science strengthens rather than compromises the principles of equality, liberty, and due process.

References

Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.

Amish Devgan v. Union of India, 2020 SCC OnLine SC 994.

Anti-Defamation League. (n.d.). *Questions and answers: The Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act* [Lesson resource, fifth anniversary edition].

<https://www.adl.org/sites/default/files/documents/assets/pdf/education-outreach/hate-crimes-prevention-act-fifth-anniversary.pdf>

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, 3(3), 193–209.

Bharatiya Nyaya Sanhita, 2023, Act No. 45 of 2023 (India).

Bureau of Justice Assistance. (2024). *FY24 Matthew Shepard and James Byrd, Jr. Hate Crimes Program*. U.S. Department of Justice. <https://bja.ojp.gov/funding/opportunities/o-bja-2024-172097>

CaseMine. (2018). *Supreme Court establishes comprehensive framework to combat mob vigilantism and lynching: Supreme Court of India*.

<https://www.casemine.com/commentary/in/supreme-court-establishes-comprehensive-framework-to-combat-mob-vigilantism-and-lynching/view>

Center for the Study of Organized Hate. (2025). *Report 2024: Hate speech events in India*. India Hate Lab. <https://www.csohate.org/2025/02/10/hate-speech-report-india-2024/>

Citizens for Justice and Peace. (2024, July 10). *BNS 2023 does nothing to bring in a nuanced effective understanding of hate speech, making its prosecution even more difficult*. <https://cjp.org.in/>

Citizens for Justice and Peace. (2025a, October 13). *Counting crimes, discounting justice: The NCRB's statistical blind spots*. <https://cjp.org.in/>

Citizens for Justice and Peace. (2025b, October 30). *The myth of neutral data: The disappearance of communal violence in NCRB data*. <https://cjp.org.in/>

Cornell Law School Legal Information Institute. (n.d.). *Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act*. Wex.

https://www.law.cornell.edu/wex/matthew_shepard_and_james_byrd_jr_hate_crimes_prevention_act

Devgan.in. (n.d.). *BNS Section 196 — Promoting enmity between different groups on ground of religion, race, place of birth, residence, language, etc., and doing acts prejudicial to maintenance of harmony*. <https://devgan.in/bns/section/196/>

Equality and Human Rights Commission. (2025). *Hate crime and hate speech: UK government action*. <https://www.equalityhumanrights.com/>

Global Freedom of Expression, Columbia University. (2023). *Devgan v. Union of India*. <https://globalfreedomofexpression.columbia.edu/cases/devgan-v-union-of-india/>

Granthaalayah Publication. (2024). Hate speech as a threat to inclusive societies. *ShodhKosh: Journal of Visual and Performing Arts*. <https://www.granthaalayahpublication.org/Arts-Journal/ShodhKosh/article/download/6199/5642/32197>

Greater Manchester Combined Authority. (2024). *Hate crime plan 2024–2028*. <https://www.greatermanchester-ca.gov.uk/>

House of Commons Library. (2026). *Hate crime* (Briefing Paper CBP-8791). UK Parliament. <https://commonslibrary.parliament.uk/research-briefings/cbp-8791/>

Human Rights Campaign. (n.d.). *Questions and answers: The Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act*. <https://www.hrc.org/>

Koehler, D. (2021). Preventing and countering violent extremism in Germany. *Revista CIDOB d'Afers Internacionals*, 128, 59–78.

Law Commission (England and Wales). (2025). *Hate crime*. <https://lawcom.gov.uk/project/hate-crime/>

Legallyin. (2024, December 1). *New provision of mob lynching in BNS, 2023: Explained*. <https://legallyin.com/>

LiveLaw. (2020, December 7). *Hate speech repudiates right to equality in a polity committed to pluralism: Supreme Court*. <https://www.livelaw.in/top-stories/hate-speech-supreme-court-equality-right-amish-devgan-166922>

LiveLaw. (2025, July 25). *SC's "Tehseen Poonawalla" directives binding on state & centre; can't monitor mob lynching incident in PIL: Allahabad HC*. <https://www.livelaw.in/high-court/allahabad-high-court/allahabad-high-court-mob-lynching-pil-monitor-govt-jamiat-ulema-tehseen-poonawalla-directives-supreme-court-298709>

Lygre, R. B., Eid, J., Larsson, G., & Ranstorp, M. (2011). Terrorism as a process: A critical review of Moghaddam's "Staircase to Terrorism." *Scandinavian Journal of Psychology*.

Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist*, 60(2), 161–169.

National Institute of Justice. (2024). *Assessing the effectiveness of programs to prevent and counter violent extremism*. U.S. Department of Justice.

<https://nij.ojp.gov/topics/articles/assessing-effectiveness-programs-prevent-and-counter-violent-extremism>

Pravasi Bhalai Sangathan v. Union of India, (2014) 11 SCC 477.

Radicalisation Awareness Network, European Commission. (n.d.). *RAN practices: Migration and home affairs*. <https://home-affairs.ec.europa.eu/>

ResearchGate. (2026). *Pathway to Violence Model as an augmentation to the UK Prevent strategy: A scoping review* [Figure repository entry referencing Moghaddam's staircase model and the 2024 Southport attack]. https://www.researchgate.net/figure/Moghaddams-Staircase-to-Terrorism_fig2_241835430

SCC Online Blog. (2020, December 8). *Freedom & rights cannot armour those who promote & incite violence: 15 notable excerpts on "hate speech" from Supreme Court's verdict in Amish Devgan case*. <https://www.scconline.com/blog/post/2020/12/08/freedom-rights-cannot-armour-those-who-promote-incite-violence-15-notable-excerpts-on-hate-speech-from-supreme-courts-verdict-in-amish-devgan-case/>

ScienceDirect. (2024). Evaluations of countering violent extremism programs: Linking success to content, approach, setting, and participants. *Aggression and Violent Behavior*.

South Asia Justice Campaign. (2025). *India persecution tracker 2024: Overview*.
<https://southasiajusticecampaign.org/ipt2024/>

Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.

Supreme Court Observer. (2023, July 27). *Plain English summary of judgment: Tehseen S. Poonawalla v. Union of India*. <https://www.scobserver.in/reports/tehseen-poonawalla-union-of-india-tushar-gandhi-plain-english-summary-of-judgment/>

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole.

Tehseen S. Poonawalla v. Union of India, Writ Petition (Civil) No. 754 of 2016, decided 17 July 2018 (Supreme Court of India).

The Leaflet. (2020, December 23). *Supreme Court's curious judgment in the Amish Devgan case: Numerous questions*. <https://theleaflet.in/>