INTERNATIONAL JOURNAL OF HUMAN RIGHTS LAW REVIEW
An International Open Access Double Blind Peer Reviewed, Referred Journal

Volume 5 | Issue 4 | 2026 Art. 33

Manufactured Consent in the Age of Artificial Intelligence: Rethinking Criminal Responsibility for Psychological Manipulation

Dr. Surbhi Wadhwa
Assistant Professor,
Vivekananda Institute of Professional Studies, Delhi

Saket Gogia
Advocate

Recommended Citation
Dr. Surbhi Wadhwa and Saket Gogia, *Manufactured Consent in the Age of Artificial Intelligence: Rethinking Criminal Responsibility for Psychological Manipulation*, 5 IJHRLR 475-486 (2026).
Available at www.ijhrlr.in/current-issues/.

Manufactured Consent in the Age of Artificial Intelligence: Rethinking Criminal Responsibility for Psychological Manipulation

ABSTRACT

The proliferation of artificial intelligence technologies has engendered an unprecedented capacity for large-scale psychological manipulation, fundamentally challenging the adequacy of existing criminal law frameworks. Contemporary AI systems, encompassing algorithmic micro-targeting, deepfake generation, emotionally adaptive chatbots, and behavioural prediction engines possess the technical capability to engineer what Edward Herman and Noam Chomsky identified as the 'manufacturing of consent', now at granular, automated, and personalised scale. Where traditional criminal law predicates liability on identifiable human volition, AI-driven manipulation operates through diffuse, probabilistic architectures that obscure the causal nexus between technological design and individual psychological harm. This paper examines the doctrinal and conceptual tensions that arise when artificial intelligence is instrumentalised as a vehicle for psychological coercion, false belief induction, and consent vitiation in contexts ranging from electoral disinformation and financial fraud to coercive control and hate speech amplification. The paper critically interrogates three foundational assumptions of criminal responsibility, the actus reus of manipulation, the mens rea of AI principals and intermediaries, and the causal standard of psychological harm in light of the emergent technical capacities of generative and adaptive AI. It argues that the dominant frameworks of intermediary liability, including Section 79 of the Information Technology Act, 2000, and analogous safe harbour doctrines in comparative jurisdictions, are constitutively inadequate to accommodate algorithmic manipulation as a discrete category of criminally cognisable conduct. Drawing upon the doctrine of innocent agency, the concept of wilful blindness, and comparative analysis of the European Union's Artificial Intelligence Act, 2024, the paper proposes a tripartite liability model predicated on design culpability, deployment intent, and systemic risk threshold. The paper concludes that the criminal law must reconceptualise consent, causation, and agency in a manner cognisant of the non-linear, emergent, and probabilistic nature of AI-mediated harm, and that the absence of a bespoke legislative framework for AI-

driven psychological manipulation constitutes a critical lacuna in India's digital justice architecture.

KEYWORDS

Artificial Intelligence, Criminal Responsibility, Psychological Manipulation, Algorithmic Governance, Consent Vitiation

INTRODUCTION

In 1988, Edward S. Herman and Noam Chomsky wrote of the mass media's capacity to manufacture consent, to shape public opinion not through coercion, but through the quiet engineering of perception.[1] What they described as a structural property of industrial-era media has, in the age of artificial intelligence, become a personalised, automated, and near-instantaneous capability. Today, algorithms do not merely reflect public sentiment, they construct it, one targeted nudge at a time. The emergence of generative AI, emotionally adaptive chatbots, deepfake technologies, and behavioural prediction engines has created an infrastructure for psychological manipulation that operates at a scale and granularity that no prior legal framework anticipated.[2] A deepfake video of a political leader can be generated, distributed, and believed within hours. A financial scam chatbot can exploit a grieving widow's loneliness for weeks before she realises she has been deceived. An algorithmic content feed can radicalise a vulnerable teenager not through a single act of incitement, but through a months-long cascade of emotionally calibrated suggestions. Each of these is a form of psychological manipulation, yet Indian criminal law, structured around identifiable human actors, ascertainable intent, and proximate causation, struggles to engage with any of them.

This paper argues that the existing criminal law framework in India is constitutively inadequate to address AI-driven psychological manipulation as a discrete category of cognizable harm. It examines the doctrinal fault lines in three core areas, the *actus reus* of algorithmic manipulation, the attribution of *mens rea* across distributed AI systems, and the causal standard applicable to diffuse psychological harm, before proposing a tripartite liability model grounded in design culpability, deployment intent, and systemic risk threshold. The paper begins by situating AI-driven psychological manipulation within existing criminal

[1] Edward S. Herman & Noam Chomsky, Manufacturing Consent: The Political Economy of the Mass Media 1-35 (Pantheon Books 1988).

[2] Generative AI refers to AI systems capable of producing novel content; emotionally adaptive chatbots are systems designed to detect and respond to users' emotional states; deepfake technologies involve synthetic media generated using machine learning; behavioural prediction engines analyse user data to forecast and influence decision-making.

law frameworks and identifying the structural gaps that emerge. It then examines the doctrinal inadequacy of intermediary liability as presently constituted under Indian and comparative law. Building on this analysis, the paper develops a proposed tripartite liability model and concludes with reflections on the need for legislative reform.

ARTIFICIAL INTELLIGENCE, PSYCHOLOGICAL MANIPULATION, AND THE CRIMINAL LAW: MAPPING THE GAP

Psychological manipulation, as a legal concept, has historically inhabited the margins of criminal law. Indian courts have recognised it within the context of cheating under Section 318 of the Bharatiya Nyaya Sanhita, 2023 (formerly Section 420 of the Indian Penal Code, 1860), undue influence in civil contracts under Section 16 of the Indian Contract Act, 1872, and the aggravated forms of coercive control now partially addressed under the Protection of Women from Domestic Violence Act, 2005.[3] What connects these provisions is a shared assumption that manipulation is an act performed by an identifiable human agent upon an identifiable human victim, with a discernible intent to deceive or coerce.

AI-driven manipulation fractures each of these assumptions. Consider the following typology of AI-mediated psychological harm. First, *computational propaganda*: the use of algorithms to amplify politically partisan, divisive, or emotionally inflammatory content disproportionately to its organic traction, as documented in the Cambridge Analytica affair and, more recently, in the use of Meta's recommendation algorithms during the 2021 Ethiopian civil conflict.[4] Second, *AI-enabled financial fraud*: the deployment of conversational AI agents to build sustained pseudo-intimate relationships with victims, extract financial and personal information, and manipulate investment decisions, a category that the Reserve Bank of India has flagged as a rapidly growing threat to financial inclusion.[5] Third, *deepfake-facilitated*

[3] Bharatiya Nyaya Sanhita, 2023, No. 45, Acts of Parliament, 2023 (India) S. 318; Indian Contract Act, 1872, No. 9, Acts of Parliament, 1872 (India) S. 16; Protection of Women from Domestic Violence Act, 2005, No. 43, Acts of Parliament, 2005 (India).

[4] Samantha Bradshaw & Philip N. Howard, The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation (Oxford Internet Institute 2019); see also Frances Haugen, Testimony Before the United States Senate Commerce Subcommittee on Consumer Protection, October 5, 2021 (describing Facebook's algorithmic amplification of divisive content for engagement maximisation).

[5] Reserve Bank of India, Annual Report on Payment and Settlement Systems 2022-23, at 47-52 (2023) (noting a 300% increase in AI-facilitated financial fraud complaints between 2020 and 2023); see also Europol, ChatGPT: The impact of Large Language Models on Law Enforcement 14-18 (Europol Innovation Lab 2023).

defamation and non-consensual intimacy: the generation of photorealistic audio-visual content depicting real individuals in fabricated scenarios, deployed to coerce, blackmail, or reputationally destroy.[6] Fourth, *algorithmic radicalisation*: the progressive exposure of vulnerable individuals to extremist content through recommendation systems optimised for engagement rather than truth.[7]

What is common to each of these categories is not merely technological sophistication, but a fundamental disruption of the *actus reus* and *mens rea* causation that underlies criminal liability. The manipulative act is not a singular event but a probabilistic process distributed across millions of micro-interactions. The intent is not the deliberate malice of a human actor but the emergent output of a system optimised toward a proximate objective, engagement, conversion, or compliance, that its designers may not have explicitly framed as harmful. The causal pathway between the algorithmic process and the psychological harm sustained by the victim is non-linear, contested, and empirically difficult to isolate. Indian criminal law, as currently constituted, has no adequate doctrinal vocabulary for any of these features. The Bharatiya Nyaya Sanhita, 2023 represents a modernisation of the criminal code in several respects, but it does not introduce any provision specifically addressing AI-mediated harm or algorithmic manipulation.[8] The Information Technology Act, 2000, while addressing cybercrimes including identity theft (Section 66C) and cheating by personation (Section 66D), does not engage with the systemic and structural dimensions of AI-driven psychological manipulation.[9] The Digital Personal Data Protection Act, 2023, India's most recent digital governance legislation, is oriented toward data protection rights rather than psychological harm, and does not create any criminal liability for manipulative use of personal data to influence behaviour.[10]

THE INADEQUACY OF INTERMEDIARY LIABILITY

The dominant legal mechanism through which AI-driven harm is addressed in India is the intermediary liability framework under Section 79 of the Information Technology Act, 2000, as amended by the Information Technology (Intermediary Guidelines and Digital Media

[6] Nina Schick, Deep Fakes and the Infocalypse: What You Urgently Need to Know 23-41 (Monoray 2020); see also Ministry of Electronics and Information Technology, Advisory on Deepfakes, November 2023, No. 11(3)/2022-CERT-In.

[7] Max Fisher, The Chaos Machine: The Inside Story of How Social Media Rewired Our Minds and Our World 78-112 (Little, Brown & Company 2022); see also Jihad Watch v. Union of India, Writ Petition (Civil) No. 44 of 2019 (Supreme Court of India) (pending, raising questions of algorithmic amplification of hate speech).

[8] Bharatiya Nyaya Sanhita, 2023, No. 45, Acts of Parliament, 2023 (India).

[9] Information Technology Act, 2000, No. 21, Acts of Parliament, 2000 (India) S. 66C, 66D.

[10] Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).

Ethics Code) Rules, 2021.[11] Section 79 provides a conditional safe harbour to intermediaries, including social media platforms and AI service providers, from liability for third-party content, subject to compliance with due diligence obligations and take-down requirements upon actual knowledge of unlawful content. Three structural limitations render Section 79 constitutively inadequate as a framework for AI-driven psychological manipulation. The first is the *passive conduit assumption*. Section 79's safe harbour was designed for platforms that transmit or host third-party content without editorial involvement. AI recommendation systems, however, are not passive conduits, they actively curate, amplify, and personalise content in a manner that shapes the informational environment of each individual user. When a platform's algorithm amplifies radicalising content to a vulnerable user, the platform is not merely hosting that content; it is, in a meaningful sense, co-authoring the psychological environment in which harm occurs. The safe harbour, premised on passivity, does not map onto this reality.[12] The second limitation is the *actual knowledge standard*. Section 79(3)(b) requires that liability attach only upon the intermediary's actual knowledge of unlawful content. AI-driven manipulation, however, is constitutionally designed to evade actual knowledge, it operates at the level of algorithmic pattern and systemic design, not individual content items. A platform may have no actual knowledge that its recommendation algorithm is radicalising a specific user, even while that radicalisation is the predictable systemic output of design choices it has consciously made.[13] The third limitation is *jurisdictional and institutional fragmentation*. AI platforms operate across jurisdictions simultaneously. Their training data may be processed in one country, their models hosted in another, and their manipulative outputs felt in a third. The Section 79 framework, premised on national jurisdiction and individual-content take-down, is ill-equipped to address this transnational architecture of harm.[14]

Comparative analysis reinforces this diagnosis. The European Union's Digital Services Act (DSA), 2022 moves beyond the passive conduit model by imposing systemic risk assessment obligations on Very Large Online Platforms, requiring them to evaluate and mitigate the harms caused by their recommendation systems, including to civic discourse

[11] Information Technology Act, 2000, No. 21, Acts of Parliament, 2000 (India) S. 79; Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, G.S.R. 139(E), February 25, 2021.
[12] *Shreya Singhal v. Union of India, (2015) 5 SCC 1 (Supreme Court of India).*
[13] Information Technology Act, 2000, No. 21, Acts of Parliament, 2000 (India) S. 79(3)(b).
[14] Jack Goldsmith & Tim Wu, Who controls the Internet? Illusions of a Borderless 67-89 (Oxford University Press, 2006).

and psychological well-being.[15] The EU's Artificial Intelligence Act (AIA), 2024 goes further, classifying AI systems that deploy subliminal techniques to influence human behaviour beyond conscious awareness as categorically prohibited.[16] Indian law has no equivalent of either provision. The absence is not incidental, it reflects a legislative choice to treat AI governance as a domain of platform responsibility rather than criminal accountability.

THE COGNITIVE ARCHITECTURE OF HARM: AUTONOMY, VULNERABILITY, AND THE LIMITS OF CONSENT IN AI-MEDIATED ENVIRONMENTS

Before a tripartite liability model can be constructed, it is necessary to examine with greater precision the nature of the harm that AI-driven psychological manipulation inflicts. Legal frameworks for manipulation have traditionally been calibrated to the rational actor model, premised on the assumption that a person of ordinary prudence can, with sufficient information, identify and resist deceptive or coercive influence. This assumption, already contested within behavioural economics, is rendered untenable by the architecture of modern AI systems, which are designed not to deceive the rational agent but to circumvent it entirely.

Cognitive science distinguishes between two systems of human decision-making: System 1, which is fast, automatic, and emotionally driven, and System 2, which is slow, deliberate, and reflective.[17] Traditional forms of manipulation, such as advertising or political persuasion, typically operate at the level of System 1, exploiting established cognitive shortcuts. AI-driven manipulation, however, does not merely exploit existing heuristics; it *engineers* new ones. Recommendation algorithms, trained on granular behavioural data, can identify with statistical precision which emotional triggers are operative in a specific user at a specific moment, and deliver content calibrated to those triggers before the user's System 2 processes can engage. The result is not persuasion in any legally cognisable sense but cognitive bypassing, the redirection of belief formation beneath the threshold of conscious deliberation. This cognitive architecture of harm has direct implications for the doctrinal concept of consent. Indian contract law and criminal law alike treat

[15] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act) 2022.
[16] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down Harmonised Ruled on Artificial Intelligence (Artificial Intelligence Act) 2024 OJ L 2024/1689.
[17] Daniel Kahneman, Thinking, Fast and Slow 20-30 (Farrar, Straus and Giroux 2011). Kahneman's dual-process theory has been applied to AI-mediated behaviour by Cass R. Sunstein & Lucia A. Reisch, Automatically Green: Behavioral Economics and Environmental Protection, 38 Harv. Envt'l. L. Rev. 127, 132 (2014).

consent as valid where it is free, informed, and uncoerced.[18] AI-driven manipulation vitiates each of these conditions without triggering any of the established legal tests for their vitiation. A user who clicks agree on a platform's terms of service has formally consented to the use of their data for personalisation; they have not meaningfully consented to having their emotional vulnerabilities mapped and systematically exploited. A victim of an AI financial fraud chatbot who voluntarily transfers funds has formally acted without coercion; they have been maneuvered into that act through a months-long process of manufactured trust that the law currently has no category to recognise.

The problem is compounded by the differential vulnerability of individuals and communities to AI manipulation. Research in digital epidemiology has established that individuals experiencing grief, social isolation, economic precarity, or mental health challenges exhibit significantly elevated susceptibility to AI-driven emotional manipulation.[19] Algorithmic systems, optimised for engagement, are not indifferent to vulnerability; they are attracted to it, because emotional dysregulation correlates with higher engagement metrics. The consequence is a systematic predation on the most socially marginalised, a dynamic that maps directly onto India's existing digital divide, where first-generation internet users, rural communities, and economically precarious populations disproportionately encounter AI systems without the digital literacy to recognise or resist their manipulative design.[20]

The legal significance of this analysis is threefold. First, it establishes that AI-driven manipulation constitutes a legally cognisable harm to autonomy that is qualitatively distinct from, and more severe than, conventional deception or persuasion, it operates at the level of cognitive architecture, not merely belief content. Second, it identifies vulnerability

[18] Indian Contract Act, 1872, No. 9, Acts of Parliament, 1872 (India) S. 14 (defining free consent as consent uncaused by coercion, undue influence, fraud, misrepresentation or mistake); Bharatiya Nyaya Sanhita, 2023, No. 45, Acts of Parliament, 2023 (India) S. 28 (consent vitiated by fear, misconception, or unsoundness of mind).

[19] Niam Yaraghi & Srinivasan Rangan, The Current and Future State of Digital Health in Developing Countries, Brookings India Working Paper No. 08 (2018); see also Tim Kendall, Testimony Before the United States Senate Commerce Committee Subcommittee on Consumer Protection, Product Safety, and Data Security, September 30, 2021 (former Facebook Director of Monetisation describing algorithmic targeting of users exhibiting signs of emotional distress).

[20] Internet and Mobile Association of India (IAMAI), Digital Divide Report 2023, at 18-24 (2023) (documenting differential AI literacy and vulnerability across urban-rural and income lines); see also Usha Ramanathan, The Aadhaar Story and What it Means for Us, 53 Economic and Political Weekly 34 (2018) (on structural digital vulnerability of marginalised communities in India).

as a structural feature of AI harm, not an incidental circumstance, which has direct implications for the aggravated liability provisions that the proposed tripartite model should incorporate. Third, it demonstrates that the harm of AI manipulation is not contingent on the victim's subjective awareness of manipulation; harm occurs precisely because awareness is foreclosed. This last point is particularly significant for criminal liability: it means that the absence of a complainant, or the victim's continued belief in the legitimacy of the manipulating system, cannot serve as a bar to prosecution under any reformed framework.

Comparative law has begun to respond to these realities. The EU's AI Act's prohibition on AI systems that deploy subliminal techniques beyond a person's consciousness is explicitly grounded in the cognitive bypassing model described above.[21] The proposed Online Safety Act framework in the United Kingdom introduces a duty of care standard that is sensitive to the vulnerability of users, particularly children and those experiencing mental health difficulties, rather than treating all users as equivalently situated rational actors. Indian law has no equivalent of either provision. The structural gap is therefore not merely a gap in criminal doctrine, it is a failure to recognise the distinctive cognitive architecture through which AI-driven harm operates. Any reform that does not begin from this recognition will address symptoms while leaving the underlying pathology untouched.

TOWARDS A TRIPARTITE LIABILITY MODEL: DESIGN CULPABILITY, DEPLOYMENT INTENT, AND SYSTEMIC RISK THRESHOLD

The paper proposes a tripartite liability model for AI-driven psychological manipulation, grounded in three analytically distinct but mutually reinforcing bases of criminal accountability. The first limb is *design culpability*. Drawing upon the doctrine of innocent agency under which a person who uses an innocent third party as the instrument of a crime is treated as the principal offender[22], this paper argues that AI developers and deployers who design systems with foreseeable manipulative effects should attract criminal liability as principals, even where the proximate harm is mediated by the autonomous operation of the AI system itself. The analogy is imperfect but instructive, just as a

[21] Regulation (EU) 2024/1689 (Artificial Intelligence Act), supra note 15, art. 5(1)(a) (prohibiting AI systems that deploy subliminal techniques beyond a person's consciousness or purposefully manipulative techniques exploiting persons' vulnerabilities); see also European Parliament, Artificial Intelligence Act: MEPs Adopt Landmark Law, March 13, 2024, PE 760.107.

[22] The doctrine of innocent agency is codified in the Indian context through the general principle that a person who causes an act to be done through an innocent agent is deemed to have done the act himself. See also R v. Michael (1840) 9 C & P 356 (establishing the innocent agency principle in English Criminal law).

person who sets a mantrap incurs liability for its foreseeable victims without requiring a specific act against each, a designer who builds a manipulation architecture incurs liability for its foreseeable psychological victims without requiring a specific act against each.

The second limb is *deployment intent*. The concept of wilful blindness under which a person who deliberately avoids acquiring knowledge of facts that would establish criminal liability is treated as having that knowledge[23] provides a doctrinal bridge between AI operators who claim ignorance of their systems' manipulative outputs and the standard of constructive knowledge. A platform operator who has access to data demonstrating that its recommendation algorithm is causing systematic psychological harm and chooses not to act upon it is, on any reasonable application of the wilful blindness doctrine, in a position of constructive criminal knowledge.

The third limb is *systemic risk threshold*. Rather than requiring proof of harm to an identifiable individual victim, which the diffuse and probabilistic nature of AI-driven manipulation renders practically impossible, this paper proposes a harm standard based on aggregate systemic risk. An AI system that, by design or in deployment, demonstrably increases the probability of psychological harm across a population above a legislatively defined threshold should attract criminal liability regardless of whether any specific victim can be identified and proved. This approach borrows from the public health model of harm, which recognises that population-level risk can ground regulatory and legal intervention even in the absence of identified individual victims.[24]

Together, these three limbs construct a liability framework that is responsive to the distinctive features of AI-driven harm without abandoning the foundational architecture of criminal law. They preserve the requirement of fault, through design culpability and wilful blindness, while reconceptualising its application to systemic, probabilistic, and distributed harms. They preserve the requirement of harm, through the systemic risk threshold, while acknowledging that individual harm identification may be epistemically unreachable in the context of algorithmic manipulation.

The model has precedent in comparative law. The AIA's prohibition on subliminal AI manipulation operates precisely on a systemic risk logic, it

[23] *Global-Tech Appliances, Inc. v. SEB S.A., 563 U.S.754 (2011).*
[24] Malcol Feeley & Jonathan Simon, The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications, 30 Criminology 449 (1992).

does not require proof of harm to a specific individual, but treats the deployment of the manipulative system as itself constituting the prohibited conduct.[25] India's own experience with the Unlawful Activities Prevention Act's conspiracy provisions, which criminalise the planning of harm irrespective of its completion, offers a domestic analogue for the proposition that criminal liability may attach to the creation of risk rather than the achievement of harm.[26]

CONCLUSION

The manufacturing of consent has always been a form of power. What artificial intelligence has changed is the scale, the precision, and the invisibility with which that power is exercised. A deepfake does not announce itself. An algorithmic nudge does not leave fingerprints. A chatbot designed to exploit loneliness does not carry a weapon. Yet the psychological harm they inflict, the erosion of autonomous belief formation, the vitiation of informed consent, the systematic distortion of individual and collective decision-making, is real, measurable, and in many cases irreversible.

Indian criminal law is not equipped for this challenge. Its categories of act, intent, and causation were constructed for a world in which harm was proximate, agents were human, and victims were identifiable. The architecture of AI-driven psychological manipulation is antithetical to each of these assumptions. The doctrinal gaps identified in this paper, the passive conduit fallacy of Section 79, the actual knowledge threshold's blindness to systemic design, the absence of a harm standard calibrated to population-level risk are not peripheral technical deficiencies. They are structural features of a legal framework that was not designed for, and cannot be retrofit to accommodate, the distinctive harms of the algorithmic age.

The tripartite model proposed in this paper design culpability, deployment intent, systemic risk threshold offers a principled basis for criminal accountability that is both doctrinally coherent and practically workable. Its implementation would require legislative action, an amendment to the Information Technology Act or a dedicated statute addressing AI accountability that moves beyond the intermediary liability paradigm. It would also require institutional investment in the technical expertise necessary for regulators and courts to evaluate the design properties of AI systems a capacity that India's criminal justice infrastructure currently lacks.

[25] Regulation (EU) 2024/1689 (Artificial Intelligence Act), supra note 15, art. 5(1)(a)-(b).
[26] Unlawful Activities (Prevention) Act, 1967, No. 37, Acts of Parliament, 1967 (India) S. 18

These are not small asks. But the alternative allowing the psychological manipulation of millions of people to proceed beneath the threshold of legal accountability, shielded by the complexity of the systems through which it is delivered is not an acceptable position for a constitutional democracy committed to the dignity and autonomy of every person. The law, as Oliver Wendell Holmes observed, must be made for men as they are.[27] In the age of artificial intelligence, it must also be made for the systems that shape them.

[27] Oliver Wendell Holmes, The common Law 1 (Little, Brown & Company 1881).