Page 1237 - 1258

INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

[ISSN 2581-5369]

Volume 9 | Issue 3

2026

Platform Accountability, Encryption, and AI: Tackling Online Child Sexual Abuse in the Digital Age

DR. WASIM AHMAD* AND MANSI PATHAK**

ABSTRACT

Three legal problems have converged, each difficult on its own. Digital platforms hosting child sexual abuse material resist accountability behind safe harbour law. End-to-end encryption creates a detection void that regulators are filling with mandates that are technically impossible to fulfil. And generative AI has broken the one assumption holding the entire global detection infrastructure together: that illegal content, once identified, can be fingerprinted and found again. This paper examines all three through Indian law, with international frameworks used not as models to copy but as pressure tests. The central argument is that India's intermediary liability regime, as operationalised by the IT Rules, 2021, contains a structural impossibility. Its proactive detection obligations apply to end-to-end encrypted platforms without specifying how detection is supposed to happen on a medium that is, by mathematical design, unreadable by the platform itself. Fulfilling those obligations requires breaking encryption, and breaking encryption would violate the constitutional right to privacy that the Supreme Court recognised in K.S. Puttaswamy. The paper analyses Just Rights for Children Alliance v. S. Harish (2024 INSC 716) with doctrinal precision, examines the EU Chat Control debate and the May 2026 AI Act amendment as comparative evidence, identifies the legal vacuum around AI-generated CSAM in Indian law, and proposes a three-part reformulation grounded in what the Constitution permits, what technology can actually do, and what enforcement requires. Child protection does not require giving up on encryption. It requires being honest about what each technical tool can and cannot accomplish.

Keywords: *Child Sexual Abuse Material (CSAM); Platform Accountability; End-to-End Encryption; AI-Generated CSAM; Intermediary Liability; POCSO Act 2012; IT Rules 2021; Client-Side Scanning.*

* Author is an Assistant Professor at Faculty of Law, Integral University, Lucknow, U.P., India.
** Author is a Research Scholar at Department of Law, T.S. Mishra University, Lucknow, U.P., India.

 [ISSN 2581-5369]

I. INTRODUCTION

On 23 September 2024, a two-judge bench of the Supreme Court of India did something that should not have taken twelve years. In *Just Rights for Children Alliance* (2024 INSC 716),[1] Chief Justice D.Y. Chandrachud and Justice J.B. Pardiwala overturned a Madras High Court ruling that had held, in January of the same year, that merely downloading or possessing child sexual abuse material without proof of intent to share it was not a criminal offence. The High Court had quashed a charge sheet. The Supreme Court called that decision an "egregious error," reinstated the prosecution, and clarified that possession under Section 15(1) of the Protection of Children from Sexual Offences Act, 2012 (POCSO) is an inchoate preparatory offence, not a lesser included form of distribution. The Court also directed Indian institutions to stop using the phrase "child pornography" and to adopt instead the term Child Sexual Exploitative and Abuse Material, or CSEAM. The word change matters. It insists that this material is a record of abuse, not a genre of content.

That ruling was right. It was also, in a structural sense, not nearly enough. The insufficiency is not a criticism of the Court. The bench resolved the question put before it with admirable clarity. But there are questions that judgments cannot answer. Who is actually obligated to detect CSAM before it spreads? By what technical means? Under what constitutional constraints? Through what international framework? Those questions remain, in Indian law, almost entirely open. And in the two years since *Just Rights for Children Alliance* was decided, the problem they address has grown considerably worse and considerably more complicated.

Three failures of regulatory architecture are the subject of this paper. They are not separate topics. They are one failure, expressed in three places. The first failure is in platform accountability. India's safe harbour framework, built on Section 79 of the Information Technology Act, 2000, conditions immunity on due diligence. The IT Rules, 2021 operationalise that due diligence through obligations that are procedural rather than substantive, self-certified rather than audited, and worded in language ("endeavour") that no court has ever been asked to define and no regulator has ever been required to enforce against a non-compliant platform.

The second failure is in the treatment of encryption. Rule 4(4) of the IT Rules, 2021 requires significant social media intermediaries to proactively identify CSAM. The rule does not exempt end-to-end encrypted messaging services. WhatsApp and Signal are significant social media

[1]Just Rights for Children Alliance v. S. Harish, 2024 INSC 716 (India) (Chandrachud, C.J., & Pardiwala, J., decided Sept. 23, 2024).

 [ISSN 2581-5369]

intermediaries. End-to-end encryption means the platform cannot read what users send. The rule demands detection on a medium the platform cannot read. This is not a gap that better enforcement can fix. It is a legal obligation built on a technical impossibility, and every year it remains on the books is a year of performed compliance that protects no child.

The third failure is in AI governance. Generative AI systems can now produce photorealistic CSAM without involving any real child, without leaving any trace in existing hash databases, and without requiring the perpetrator to have any technical skill beyond knowing which models to run. India's IT Act, POCSO, and the Digital Personal Data Protection Act, 2023 say nothing about this. The legislature has not acted. The question of whether synthetic CSAM even falls within existing criminal definitions is unresolved. The EU amended its AI Act as recently as May 2026 to explicitly prohibit AI systems that generate CSAM. India has not.

Part II establishes the scale and qualitative transformation of the problem. Part III provides a doctrinal analysis of India's framework and its internal contradictions. Part IV examines the encryption debate through comparative evidence from the EU, the UK, and the United States. Part V analyses AI-generated CSAM as a distinct legal and technical problem. Part VI applies the *Puttaswamy* proportionality test rigorously to detection obligations. Part VII proposes a reformulated framework. Part VIII concludes.

II. THE TRANSFORMATION OF THE PROBLEM: SCALE, STRUCTURE, AND SYNTHETIC CONTENT

A. Global Dimensions

Numbers matter here, though not for the purpose of emotional effect. They establish the mismatch between the scale of the problem and the legal architecture that is supposed to address it. In 2025, the NCMEC CyberTipline received reports containing over 61.8 million images and videos related to child sexual exploitation.[2] That figure represents what platforms detected and chose to report. It says nothing about what went undetected.

The ten-year trajectory makes the situation starker. CyberTipline reports increased by a factor of 77 between 2012 and 2022.[3] In 2024 alone, reports involving generative AI content rose by 1,325 percent, from roughly 4,700 in 2023 to approximately 67,000 in 2024.[4] The Internet

[2]Nat'l Ctr. for Missing & Exploited Children, *CyberTipline Data 2025*, https://www.missingkids.org/gethelpnow/cybertipline/cybertiplinedata (last visited May 2026).
[3]*Stop CSAM Act of 2023 One-Pager*, S. Comm. on the Judiciary (Apr. 2023), https://www.durbin.senate.gov (citing NCMEC CyberTipline data).
[4]*Why Hash Matching Won't Stop AI-Generated CSAM*, CASESCAN (Feb. 2026), https://casescan.com/blog/ai-content-moderation-csam-detection-platforms/ (citing NCMEC CyberTipline 2024 data).

 [ISSN 2581-5369]

Watch Foundation found over 3,000 AI-generated CSAM images in a single month of 2024 and documented a 26,362 percent increase in photorealistic AI-generated abuse videos over the preceding year.[5]

What these numbers establish is not simply that the problem is getting worse. The nature of the problem is changing. The global detection infrastructure rests on Microsoft's PhotoDNA and a set of equivalent hash-matching tools. These systems work by comparing uploaded files against databases of known illegal content. An image that has been identified before can be fingerprinted. If it appears again, it can be caught. Generative AI breaks this logic completely. Every image a diffusion model produces is new. It has never been seen before. It matches nothing in any database. The entire infrastructure was built on the assumption that illegal content recirculates. That assumption no longer holds.

B. The Indian Context: Reactive by Design

India's exposure to this problem is enormous and its institutional capacity to address it is thin. With over 900 million internet users, India is among the three largest internet markets in the world. Its enforcement agencies have acknowledged, directly and in policy documents, that they depend primarily on foreign intelligence, specifically NCMEC CyberTipline reports, for leads on CSAM cases.[6] India has no domestic equivalent of that infrastructure. Public reporting of CSAM circulation is extremely low. Some states have built dedicated child protection units, Telangana's being the most often cited, and that unit has produced hundreds of arrests. But it is a local model, not a national system.

India's posture toward the Budapest Convention on Cybercrime compounds this.[7] The Convention provides the main mutual legal assistance framework for cross-border CSAM investigations. India has declined to ratify it, citing concerns about sharing sensitive data with foreign law enforcement. The practical consequence is that Indian investigators pursuing cases with an international dimension do so without the formal cooperation channels that investigators in most comparable jurisdictions treat as routine. They depend on foreign goodwill and informal arrangements. That is not a foundation for systematic enforcement.

---

[5]Internet Watch Found., *Annual Report 2024*; *Policing the Pixels: The Fight Against AI CSAM*, RESOLVER (Sept. 2025), https://www.resolver.com/blog/ai-csam-moderation-online-safety-act/.

[6]Daniel Manoj et al., *Behind the Screens: Understanding the Gaps in India's Fight Against Online Child Sexual Abuse and Exploitation*, 4 CHILD PROT. & PRAC. 100088 (2024); *Reactive Laws, Relentless Abuse*, OXFORD HUM. RTS. HUB (2025), https://ohrh.law.ox.ac.uk.

[7]Manoj et al., *supra* note 6; *see also Regulations for Online Child Sexual Abuse in India*, IASSCORE (Oct. 2022) (noting India's stated sovereignty concerns regarding the Budapest Convention on Cybercrime).

 [ISSN 2581-5369]

In 2023, the National Commission for the Protection of Child Rights met with representatives of Google, YouTube, Meta, Reddit, Snapchat, and Sharechat and reached what was described as a consensus on mandatory CSAM reporting and KYC compliance under Section 9 of the Digital Personal Data Protection Act, 2023 (DPDPA).[8] A consensus at a meeting, however, is not enforcement. The DPDPA itself is a data protection statute. Section 9 governs what platforms may do with children's personal data. It does not require platforms to detect CSAM. It constrains processing; it does not mandate investigation. The gap between what the DPDPA requires and what CSAM accountability needs is not a detail. It is the central architectural problem.

III. THE INDIAN LEGAL FRAMEWORK: STRUCTURE AND CONTRADICTION

A. The Criminal Law After Just Rights for Children Alliance

The criminal law on CSAM in India has two pillars: Section 67B of the IT Act and Section 15 of POCSO. It is worth understanding both with precision, because the gaps in each become clearer when set against what the Supreme Court said in 2024. Section 67B of the IT Act criminalises publication, transmission, creation, collection, seeking, browsing, downloading, and advertising of material depicting children in sexually explicit conduct. The provision is broad in scope. It covers the supply chain of CSAM distribution in digital environments with comprehensiveness that the 2000 Act's original text did not have. The penalty for a first conviction under Section 67B(a) is imprisonment up to five years and a fine. A second conviction carries up to seven years. These are not trivial penalties.

Section 15 of POCSO, as amended in 2019, creates three graduated possessory offences. The first is storage without reporting to a designated authority. The second is storage with intent to share or transmit. The third is storage with intent to use the material in the commission of an offence. The 2019 amendments strengthened the provision significantly. Before the amendments, courts had read Section 15 narrowly, in some cases requiring evidence of actual distribution for liability to attach.

The Madras High Court's January 2024 ruling took that narrow reading further than any court had before. The respondent, S. Harish, had CSAM on his phone. His defence was that the files had been auto-downloaded through WhatsApp group chats without his volition. The High Court accepted this as sufficient to quash the charge sheet under Section 482 CrPC. It held that possession without demonstrated intent to share was not an offence. The Supreme Court's

[8]Nat'l Comm'n for Prot. of Child Rights, *Meeting with Social Media Platforms on Child Safety* (2024), reported in DECCAN HERALD; The Digital Personal Data Protection Act, 2023, § 9, No. 22, Acts of Parliament, 2023 (India).

 [ISSN 2581-5369]

reversal rests on three holdings, each of which has doctrinal significance beyond the facts of the case. The first is the inchoate character of Section 15. The Court held that possession is criminalised as a preparatory act toward distribution or use of the material, not as a lesser form of those more serious offences. The statutory presumption in Section 30 of POCSO, which presumes a culpable mental state once the actus reus is established, operates at trial, not at the pre-trial quashing stage. The High Court conflated these stages.[9]

The second holding concerns the auto-download defence. The Court declined to treat the claim that files arrived automatically through WhatsApp group membership as a ground for quashing. Whether files were received knowingly is a question of fact for trial. This matters practically. Automated viral sharing through group messaging is one of the primary circulation mechanisms for CSAM. If auto-download were a complete defence at the charge stage, it would be an easily manufactured one.

The third holding is terminological but not merely symbolic. The Court directed Indian institutions to replace the phrase "child pornography" with CSEAM throughout legal and regulatory discourse.[10] This is more than a style guide. The word "pornography" frames this material as a category of explicit content, carrying connotations of consensuality and commercial production that are wholly inapplicable. CSEAM insists on the nature of the material as a record of abuse. That framing shapes future interpretive questions, including the question this paper returns to in Part V: whether AI-generated imagery of children that involves no real child is nonetheless CSEAM within the meaning of existing law.

The Court also upheld Sections 19 and 20 of POCSO, which impose mandatory reporting obligations on anyone who encounters CSAM, including hotel staff, hospital personnel, media organisations, and studios. The 2025 Madras High Court ruling in *Ramkumar v. Union of India* reinforced judicial expectations of platforms specifically, holding that intermediaries cannot treat passive compliance as a shield while abuse ecosystems operate on their infrastructure.[11]

B. The Intermediary Liability Architecture and Its Central Defect

India's framework for platform accountability is built on Section 79 of the IT Act, which provides safe harbour immunity to intermediaries for third-party content, subject to the

[9]Just Rights for Children Alliance, *supra* note 1; *see also Comprehensive Commentary on Just Rights for Children Alliance v. S. Harish (2024 INSC 716)*.
[10]Just Rights for Children Alliance, *supra* note 1; LIVELAW SC 728 (2024) (analysing the Court's CSEAM terminology direction).
[11]Ramkumar v. Union of India (Madras H.C. 2025) (India), discussed in *Intermediary Liability in the Digital Age: Judicial Enforcement and the Ramkumar Decision*, MONDAQ (July 2025), https://www.mondaq.com/india/broadcasting-film-tv-radio/1656898.

 [ISSN 2581-5369]

condition that they observe due diligence and respond expeditiously upon acquiring actual knowledge of unlawful content. The 2008 amendments extended this immunity across all areas of law, not just the IT Act itself. This made the protection broader and, in some ways, less accountable.[12]

The IT Rules, 2021 operationalise Section 79's due diligence requirements for Significant Social Media Intermediaries, defined as platforms with more than five million registered users in India.[13] The enhanced obligations include the appointment of a Grievance Officer, a Nodal Contact Person, and a Chief Compliance Officer, all resident in India. Platforms must acknowledge user complaints within 24 hours and resolve them within 15 days. They must submit monthly compliance reports. These are real obligations, and some of them produce real accountability.

Rule 4(4) is the provision that matters most for CSAM, and it is where the framework breaks down. The rule requires significant social media intermediaries to "endeavour" to deploy technology-based measures to proactively identify information depicting child sexual abuse, rape, or content previously removed by order of a court or government authority.

The word "endeavour" has no legal content. No statute defines it. No court has been asked to determine what constitutes adequate endeavour for CSAM detection purposes. The provision specifies no technology type, no detection threshold, no reporting frequency, and no audit mechanism. A platform can appoint its three officers, submit its monthly reports, and truthfully assert that it is endeavouring, while deploying no detection technology at all. The compliance regime is, by design, self-certified and subjectively framed. This is not a minor drafting infelicity. It is the reason that despite Rule 4(4) having been on the books since 2021, India's enforcement agencies remain dependent on American intelligence for CSAM leads.[14]

The deeper problem is that Rule 4(4) applies without differentiation to every significant social media intermediary, including those that use end-to-end encryption. WhatsApp has over 500 million Indian users. Signal, Telegram's secret chat function, and several other encrypted messaging applications are also significant social media intermediaries under the Rules. End-

[12]The Information Technology Act, 2000, § 79, No. 21, Acts of Parliament, 2000 (India), as amended by the Information Technology (Amendment) Act, 2008, No. 10, Acts of Parliament, 2009 (India); Avnish Bajaj v. State (NCT of Delhi), (2005) 3 Comp. L.J. 364 (India) (highlighting the pre-2008 liability gap exposed by the baazee.com prosecution).

[13]The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, rr. 3, 4 (India).

[14]*Safe Harbor Under Section 79 IT Act: Why Legal Immunity for Platforms Is Crumbling*, KSANDK (July 2025), https://ksandk.com/corporate/safe-harbor-intermediary-immunity-indian-law/; *Internet Impact Brief: 2021 Indian Intermediary Guidelines*, INTERNET SOC'Y (Nov. 2021), https://www.internetsociety.org/resources/2021/internet-impact-brief-2021-indian-intermediary-guidelines-and-the-internet-experience-in-india/

 [ISSN 2581-5369]

to-end encryption is not a design preference. It is a mathematical property. A message encrypted on the sender's device cannot be read by the platform in transit. The platform's servers handle ciphertext, not plaintext. Deploying automated tools to proactively identify CSAM in content that the platform cannot read is not possible without inserting scanning technology on the user's device, before encryption occurs. That approach, known as client-side scanning, effectively breaks the guarantee that end-to-end encryption provides.

Mozilla's analysis of the IT Rules, 2021, published immediately after notification, concluded that the automated filtering provisions were fundamentally incompatible with end-to-end encryption and would require platforms to embed monitoring infrastructure into their products.[15] The Internet Society reached identical conclusions.[16] The Dialogue's detailed application of the *Puttaswamy* proportionality test to the related traceability requirement in Rule 4(2) found it constitutionally vulnerable on similar grounds.[17] The Ministry of Electronics and Information Technology issued notices to X, YouTube, and Telegram in 2023 warning that non-compliance with CSAM removal obligations risked safe harbour forfeiture.[18] None of these notices, and none of the academic commentary, resolved the prior question: how is an encrypted messaging platform supposed to comply without destroying its core security architecture?

This is not an implementation problem. Implementation problems can be fixed by better enforcement, additional resources, or clearer guidance. This is a design problem. The obligation and the constraint are mutually exclusive. Every year this contradiction remains unaddressed is a year of legal performance that leaves children unprotected and billions of users exposed to the possibility that the government will eventually resolve the tension by mandating the surveillance rather than fixing the law.

C. The DPDPA: Protective in Orientation, Silent on Detection

The Digital Personal Data Protection Act, 2023 and the DPDP Rules, 2025 represent a genuine advance in children's data rights. Section 9 requires verifiable parental consent before processing any personal data of a person under 18. It prohibits behavioural monitoring and

[15] *India's New Intermediary Liability and Digital Media Regulations Will Harm the Open Internet*, MOZILLA (Mar. 2021), https://blog.mozilla.org/netpolicy/2021/03/02/.
[16] Internet Soc'y, *supra* note 14.
[17] *Does the Traceability Requirement Meet the Puttaswamy Test?*, THE DIALOGUE (Sept. 2023), https://thedialogue.co/does-the-traceability-requirement-meet-the-puttaswamy-test/ (applying the four-pronged Puttaswamy test to Rule 4(2) of the IT Rules, 2021).
[18] Oxford Hum. Rts. Hub, *supra* note 6 (noting 2023 MeitY notices to X, YouTube, and Telegram requiring CSAM removal).

 [ISSN 2581-5369]

targeted advertising directed at children. Penalties for violations reach up to 200 crore rupees.[19] For a country with one of the world's largest populations of minor internet users, this matters.

But the DPDPA operates on a different axis than CSAM accountability. Section 9 governs what platforms may do with children's data. It does not ask what platforms must look for on their infrastructure. A platform can be fully compliant with every provision of the DPDPA while hosting CSAM it has made no attempt to detect. The two frameworks occupy different regulatory registers, and their interaction has not been considered in either statute. The gap between data protection and content accountability is where platforms find it easiest to avoid responsibility for the harms that occur on their services.

IV. THE ENCRYPTION DEBATE: LESSONS FROM INTERNATIONAL EXPERIENCE

A. The EU Chat Control: A Cautionary Example

The European Union's proposed Regulation on Child Sexual Abuse, known internationally as Chat Control, offers the most detailed public record available of what happens when a legislature tries to mandate CSAM detection in encrypted communications without first establishing whether it is technically possible. The legislative history is worth examining not to adopt any of it but to understand the failure mode.

The European Commission's 2022 proposal required messaging platforms to scan communications for CSAM. For unencrypted platforms, server-side scanning is feasible and largely uncontroversial among technical experts. For end-to-end encrypted services, it is not. The Commission's answer was client-side scanning: software deployed on users' devices that reads content before it is encrypted. This preserves encryption in a technical sense while eliminating its practical benefit, since the scan occurs before the encryption engages.

Expert opposition was immediate and categorical. The Global Encryption Coalition, comprising the Center for Democracy and Technology, Internet Society, Mozilla, and a coalition of cryptographers and civil society organisations, stated that client-side scanning is fundamentally inconsistent with the promise of end-to-end encryption, regardless of the technical framing.[20] When the Belgian Presidency reformulated the proposal in May 2024 as "upload moderation," scanning content before it is uploaded rather than during transmission, the Coalition's response

[19]The Digital Personal Data Protection Act, 2023, § 9, No. 22, Acts of Parliament, 2023 (India); Digital Personal Data Protection Rules, 2025 (India); *Children's Data Under the DPDP Act, 2023 and DPDP Rules, 2025*, KSANDK (Nov. 2025), https://ksandk.com/data-protection-and-data-privacy/childrens-data-protection-under-indias-dpdp-rules/.

[20]Glob. Encryption Coal. Steering Comm., *Statement on the Belgian Presidency's Compromise Proposal on EU CSAM* (Apr. 2024), https://www.globalencryption.org/2024/04/.

 [ISSN 2581-5369]

was that the relabelling changed nothing.[21] Scanning content on the user's device before encryption engages is client-side scanning. Calling it something else does not alter what it does.

The Max Planck Institute's assessment captured the operational reality in eight words: "more monitoring, but not more protection."[22] The reasoning is straightforward and has not been credibly contested. CSAM offenders are not confined to regulated platforms. They use purpose-built criminal networks, dark web infrastructure, and non-compliant applications. Implementing client-side scanning on WhatsApp does not catch CSAM shared via Signal, via Telegram's secret chats, or via any application operating outside the regulation's reach. The surveillance burden falls on billions of innocent users of regulated services while offenders route around them. The ratio of privacy cost to child protection benefit is deeply unfavourable, and the benefit itself rests on the assumption that offenders will remain on regulated platforms after scanning is introduced, which is not a safe assumption.

As of May 2026, Chat Control is in trilogue negotiations between the Commission, the Parliament, and the Council. The Parliament's position excludes from the regulation's scope any data protected by end-to-end encryption. The regulation's current trajectory is toward voluntary scanning frameworks, with mandatory measures deferred to a future in which the technical objections are somehow resolved. Cryptographers do not believe that future exists. The encryption and detection problems are not amenable to a technological fix that the Parliament can simply order into existence by setting a deadline.

India should treat Chat Control as a cautionary exhibit. The political pressure to mandate CSAM detection in encrypted environments is real and will be experienced by India's parliament as it is experienced by every legislature that takes the problem seriously. The EU's experience demonstrates what that pressure produces if legislators do not first insist on technical honesty from their advisors. It produces mandates that perform child protection without achieving it.

B. The May 2026 EU AI Act Amendment: A More Coherent Model

Separate from Chat Control, the EU produced in May 2026 a more coherent and more immediately relevant development. On 7 May 2026, the co-legislators of the EU AI Act agreed

[21] Glob. Encryption Coal., *Joint Statement on the Dangers of the May 2024 Council of the EU Compromise Proposal on EU CSAM* (May 2024), https://www.globalencryption.org/2024/05/joint-statement-on-the-dangers-of-the-may-2024-council-of-the-eu-compromise-proposal-on-eu-csam/

[22] Max Planck Inst., *More Monitoring, But Not More Protection* (2024), https://www.mpg.de/25788438/chat-control-eu-client-side-scanning.

 [ISSN 2581-5369]

to a new prohibition, taking effect on 2 December 2026, on AI systems that generate or manipulate child sexual abuse material within the meaning of Directive 2011/93/EU.[23]

The structure of the prohibition is worth examining in detail because it resolves a legal problem that India has not yet engaged with. Under the amended Article 5 of the AI Act, the prohibition applies differently to providers and deployers of AI systems. A provider is an entity that develops, trains, or places an AI system on the market. A deployer is an entity that uses a system built by a provider. For providers, the prohibition applies where CSAM generation is the system's intended purpose, or where CSAM output is a "reasonably foreseeable and reproducible" outcome without significant technical modification. For deployers, the prohibition is narrower: it applies only where the deployer intentionally uses the system to generate prohibited material, including by circumventing the provider's safety measures. Accidental generation is expressly excluded.[24]

The provider-deployer distinction does significant legal work. It holds developers accountable for foreseeable misuse of their systems without requiring proof that they intended or approved the misuse. The "reasonably foreseeable and reproducible" standard is demanding: a developer cannot escape liability simply by disclaiming CSAM generation as a use case if the system reliably produces such content with minor modification. At the same time, it does not impose strict liability on developers for genuinely unforeseeable downstream uses. This is a workable and constitutionally defensible calibration of liability along the AI supply chain.

India has nothing equivalent. The IT Act, POCSO, and the DPDPA are all silent on AI model liability for CSAM generation. India's AI governance landscape consists of advisory frameworks without statutory force. The EU's May 2026 amendment provides a template that is technically grounded, legally precise, and free of the encryption complications that made Chat Control so controversial. There is no technical objection to prohibiting AI systems that generate CSAM. The only question is whether India's legislature will act before the next judicial petition forces the issue.

C. The UK and US Frameworks: Different Approaches, Different Gaps

[23]Press Release, Council of the Eur. Union, No. 299/26 (May 7, 2026), https://www.consilium.europa.eu/en/press/press-releases/2026/05/07/; *EU AI Act Update: EU Resolves to Change Rules and Extend Deadlines*, LATHAM & WATKINS (May 2026), https://www.lw.com/en/insights/ai-act-update-eu-resolves-to-change-rules-and-extend-deadlines.

[24]*EU AI Act Update: Timeline Relief, Targeted Simplification, and New Prohibitions*, INSIDE PRIVACY (May 2026), https://www.insideprivacy.com/artificial-intelligence/eu-ai-act-update-timeline-relief-targeted-simplification-and-new-prohibitions/.

 [ISSN 2581-5369]

The UK's Online Safety Act 2023 takes a more adaptive approach than either Chat Control or the US system. It requires platforms to take proactive measures against CSAM and gives Ofcom enforcement authority to impose fines on non-compliant platforms. For encrypted services, it defers detection obligations to the point at which it becomes technically feasible to detect CSAM without compromising end-to-end encryption.[25] Ofcom has already levied fines for CSAM failures. But the encrypted services question remains formally unresolved: the Act defers it to a technology that no cryptographer believes is coming.

The United States framework rests on mandatory reporting rather than mandatory detection. Under 18 USC Section 2258A, platforms that discover CSAM must report it to NCMEC. There is no federal requirement that platforms implement any specific technology to do the discovering. The EARN IT Act, which would have conditioned Section 230 immunity on compliance with child safety standards set by a national commission, has been in legislative limbo since 2020.[26] The current US system works well where major platforms invest voluntarily in hash-matching and classifier tools, because those platforms generate the reporting data on which the global CSAM intelligence infrastructure depends. It fails silently where platforms choose not to invest, and it fails structurally in the face of AI-generated content.

Three frameworks, three distinct failure modes. The EU tries to mandate detection in encrypted environments and produces technically incoherent obligations. The UK defers the hard question. The US creates voluntary excellence and structural gaps simultaneously. India's current framework contains a fourth failure mode: it imposes obligations it cannot enforce because it has never resolved the technical question of what those obligations actually require.

V. AI-GENERATED CSAM: A DISTINCT LEGAL AND TECHNICAL PROBLEM

A. Why Hash-Matching Cannot Address Synthetic Content

Microsoft's PhotoDNA is the industry standard for CSAM detection and has been since roughly 2009. The algorithm converts an image to grayscale, divides it into segments, applies discrete cosine transform processing, and generates a 1,152-bit perceptual hash that is resilient to common manipulations: resizing, compression, colour shifts.[27] The hash is then compared against databases maintained by NCMEC and the Internet Watch Foundation. A match triggers automatic flagging and mandatory reporting. Google's CSAI Match performs equivalent

[25]Online Safety Act 2023, c. 50 (UK); CaseScan, *supra* note 4 (on Ofcom enforcement); Resolver, *supra* note 5 (on the encrypted services provisions).
[26]18 U.S.C. § 2258A; *Stop CSAM Act of 2023* (proposed), S. Comm. on the Judiciary, https://www.durbin.senate.gov.
[27]*Hashing in the Fight Against CSAM: Technology at the Crossroads of Law and Ethics*, 5 J. CYBERSECURITY & PRIV. 92 (2025), https://www.mdpi.com/2624-800X/5/4/92.

 [ISSN 2581-5369]

functions for video. Meta has open-sourced its PDQ and TMK algorithms for adoption by smaller platforms.

This infrastructure has one foundational assumption built into it at every level: the content being detected has been seen before. Known content can be fingerprinted. Re-circulation of known content can be caught. AI-generated CSAM invalidates this assumption completely. A diffusion model produces a new image each time it runs. That image does not match any entry in any database because no entry exists for it. The 1,325 percent increase in AI-CSAM reports at the CyberTipline in 2024 reflects what classifiers and human moderators caught through means other than hash-matching.[28] It does not reflect the volume of AI-generated CSAM that moved through platforms completely undetected. Nobody knows what that figure is.

The technical response is AI-based classification. Classifiers analyse visual content in real time using machine learning models trained on large curated datasets, assessing whether material depicts illegal content without requiring a prior database match. Thorn's classifier technology, deployed by major platforms including Adobe,[29] represents the current state of the art. But classifiers come with their own irreducible complication: training them requires exposure to CSAM, which creates legal and ethical problems even for researchers acting in good faith. The Wilson Center's 2024 report on AI-generated CSAM found that experts across law enforcement, technology, and civil society agreed on the necessity of clean training datasets but disagreed sharply on how effective classifiers for novel synthetic content could be built without creating or maintaining access to illegal material for research purposes.[30]

B. The Open-Source Model Problem

The development of closed-source generative AI models by major companies creates a situation that is difficult but tractable. A closed-source model is accessible to auditors, subject to its developer's safety measures, and can be updated when vulnerabilities are identified. The developer is a known entity with a legal address and significant reputational exposure.

Open-source models are categorically different. Once a generative AI model is released publicly, it can be downloaded, fine-tuned by any technically capable user, and its safety guardrails removed. Forums providing detailed guides for generating CSAM through modified open-source models have been documented by the Internet Watch Foundation and law

[28]CaseScan, *supra* note 4; *What Is CSAM? Definition, Risks and Detection Methods*, DNSFILTER, https://www.dnsfilter.com/glossary/csam.
[29]*Adobe's Commitment to Child Safety*, ADOBE (Nov. 2025), https://www.adobe.com/trust/transparency/child-safety.html.
[30]Wilson Ctr., *Combatting AI-Generated CSAM* (Feb. 2025), https://www.wilsoncenter.org/article/combatting-ai-generated-csam.

enforcement.[31] The NTIA's 2024 risk framework acknowledged that developer liability for downstream fine-tuning of open-source models is among the most complex unresolved questions in AI governance.[32]

India's law cannot answer this question as currently written. The IT Act and POCSO require an identifiable actor who produced, transmitted, or possessed CSAM. A developer who releases an open-source model that is subsequently fine-tuned by a third party to generate CSAM occupies legally uncertain territory: the developer did not generate the material, did not possess it, and may have had no knowledge of the downstream use. The actus reus requirements of existing law, designed for human actors engaged in deliberate conduct, do not map onto the AI supply chain.

C. Does Synthetic CSAM Fall Within Indian Law?

The immediate doctrinal question, which India's courts have not yet been required to answer, is whether AI-generated CSAM is CSAM under existing law at all.

Section 67B of the IT Act uses the phrase "depicting children in obscene or indecent or sexually explicit manner." The verb "depicting" carries an implicit reference to something real: a photograph depicts an event that occurred. A synthetic image may be argued, by someone with a strong interest in the argument, not to depict anything. Section 15 of POCSO frames the offence in terms of "child pornography" (now redirected by the Supreme Court's CSEAM direction), which could be read to require the involvement of an actual child in the material's creation.

The purposive interpretation cuts strongly in the other direction, and this paper argues it should prevail. The Supreme Court's framing in *Just Rights for Children Alliance* treats possession as criminalised because it is preparatory to distribution and use, not because it memorialises a specific past abuse. The harm rationale extends beyond victim-specific injury to include demand suppression, the prevention of normalisation, and the disruption of criminal networks in which synthetic and real material routinely coexist. An interpretation that excludes AI-generated material creates a loophole that offenders can exploit simply by claiming AI origin, a claim that the quality of current diffusion models makes difficult to disprove.

[31]Resolver, *supra* note 5; Internet Watch Found., *Annual Report 2024* (documenting forums providing open-source model misuse guides).
[32]Nat'l Telecomms. & Info. Admin., *Dual-Use Foundation Artificial Intelligence Models with Widely Available Model Weights* (2024), https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report.

 [ISSN 2581-5369]

The UK and EU have resolved this ambiguity through explicit statutory text: both criminalise synthetic CSAM regardless of whether a real child was involved. India's legislature has not acted. The CSEAM terminology direction from the Supreme Court creates an interpretive foundation for including synthetic material, since abuse material can describe the character of the content rather than the method of its production. But a direction is not a statutory amendment, and the gap requires legislative closure.

VI. THE CONSTITUTIONAL FRAMEWORK: PUTTASWAMY APPLIED

Any detection obligation imposed on platforms must pass the four-pronged proportionality test established in *Justice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1.[33] The test, articulated by Chandrachud J. and elaborated by Kaul J. in their respective opinions, requires: (i) the action must be sanctioned by law; (ii) it must serve a legitimate aim in a democratic society; (iii) it must be necessary and proportionate to that aim; and (iv) it must be accompanied by procedural safeguards against abuse.[34] Applying this test to client-side scanning specifically, since that is the only known technical mechanism for CSAM detection in end-to-end encrypted environments, yields results that are not close.

Prong (i): Sanctioned by law. Rule 4(4) of the IT Rules, 2021 provides some basis, but its "endeavour" language does not clearly sanction any specific detection technology. A CSS mandate of sufficient specificity to constitute an actual legal obligation would require either express statutory authority under Section 79 of the IT Act or fresh primary legislation. A vague encouragement to endeavour is not a legal sanction for deploying surveillance technology on the devices of hundreds of millions of users.

Prong (ii): Legitimate aim. Protection of children from sexual exploitation is a legitimate aim of the highest order. This prong is satisfied without difficulty.

Prong (iii): Necessity and proportionality. This is where CSS fails, and fails decisively. Necessity requires the least intrusive means of achieving the legitimate aim. CSS deployed on every user's device without individualised suspicion surveils the communications of hundreds of millions of people who have done nothing wrong in order to detect the content produced by a fraction of that population. The harm to the innocent is certain and immediate. The benefit to child protection is partial and easily circumvented, since offenders move to non-compliant

[33]Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 S.C.C. 1 (India) (nine-judge bench holding privacy a fundamental right under articles 14, 19, and 21 of the Constitution of India).
[34]Puttaswamy, *supra* note 33, ¶¶ 263–265 (Chandrachud, J.; Kaul, J., adding procedural safeguards as the fourth prong); *An Analysis of Puttaswamy*, INDRASTRA GLOB. (Nov. 2017), https://medium.com/indrastra/an-analysis-of-puttaswamy-the-supreme-courts-privacy-verdict-53d97d0b3fc6.

 [ISSN 2581-5369]

platforms. Chandrachud J.'s Aadhaar dissent, applying the same *Puttaswamy* standard, described treating "every citizen as a potential criminal without even requiring the State to draw a reasonable belief" as disproportionate.[35] That description fits CSS precisely. Targeted alternatives exist: metadata analysis, financial intelligence, risk-based profiling, and undercover operations do not require reading the communications of users who are not under suspicion. CSS is not the least intrusive means. It is among the most intrusive means imaginable for a communication medium.

Prong (iv): Procedural safeguards. No version of any CSS proposal, in India or internationally, has included judicial authorisation for individual scanning operations, meaningful audit mechanisms, or effective redress for users whose communications are falsely flagged. The Belgian Presidency's "consent" requirement was described by critics as structurally mandatory: users who declined would lose access to platform features, making it forced consent inconsistent with the standard of freely given consent under EU law.[36] An Indian CSS mandate would face the same objection.

The *Puttaswamy* analysis does not conclude that child protection cannot justify detection obligations. It concludes that only those obligations which are technically achievable without mass suspicionless surveillance of the general population survive the proportionality test. Hash-matching and classifier deployment on unencrypted platforms are proportionate: they target content rather than communications, they do not read personal messages between identified individuals, and they generate specific signals from specific content rather than monitoring all traffic. CSS on encrypted platforms is not proportionate, and cannot be made so without primary legislation that itself survives proportionality review.

VII. Reformulated framework

What follows is not an incremental improvement to the existing Rules. It is a reconception of the framework's basic logic. Three premises drive it. First, obligations must be calibrated to what each platform type can technically accomplish, not to an undifferentiated standard that is technically impossible for some and meaninglessly vague for others. Second, all obligations must survive the *Puttaswamy* proportionality test on each of its four prongs. Third, obligations must be measurable, auditable, and subject to specified consequences for non-compliance. An obligation that no regulator can verify is not an obligation. It is a gesture.

[35]Justice K.S. Puttaswamy (Retd.) v. Union of India, (2019) 1 S.C.C. 1 (India).

[36]*Joint Statement: EU Compromise CSAM Scanning*, TUTA (June 2024), https://tuta.com/blog/joint-statement-eu-compromise-csam-scanning (noting that the Belgian Presidency's so-called consent requirement was structurally mandatory and therefore inconsistent with the GDPR standard of freely given consent).

A. Proposal 1: A Differentiated Detection Regime

India's "endeavour" standard must be replaced by a tiered framework calibrated to platform architecture.

Tier A: Unencrypted content platforms. Significant social media intermediaries processing unencrypted user-generated content, including social media feeds, video hosting platforms, cloud storage services, and image sharing applications, should face statutory and specific detection obligations. These should include mandatory deployment of hash-matching against NCMEC and IWF databases updated at defined intervals; mandatory deployment of at least one AI classifier for novel content detection; mandatory reporting to a designated Indian authority within 24 hours of confirmed detection; and mandatory account termination for confirmed uploaders. Failure to deploy required technologies should trigger graduated civil liability and administrative penalty structures, not merely the threat of safe harbour forfeiture. The "endeavour" standard should be removed from this tier entirely.

Tier B: End-to-end encrypted messaging services. The current Rule 4(4) obligation as applied to E2EE platforms must be replaced with obligations calibrated to what encryption genuinely permits. These should include mandatory hash-matching of unencrypted elements such as profile images, group names, and publicly visible content; mandatory cooperation with court-ordered disclosure of available metadata; mandatory removal on receipt of a court-verified report identifying a specific CSAM-associated account; and mandatory semi-annual transparency reporting on government data requests received and responded to. Client-side scanning should not be required unless primary legislation, following a public consultation informed by independent technical assessment and formal parliamentary debate, determines that a specific technology has been certified to achieve detection without compromising end-to-end encryption. That standard cannot currently be met.

Tier C: Hosting and file-sharing infrastructure. Large file-sharing services, forums, and hosting infrastructure should face obligations calibrated to whether user content is encrypted at rest. Where the platform can access content, full Tier A obligations apply. Where content is encrypted by the uploader before it reaches the platform, Tier B obligations apply.

B. Proposal 2: AI Model Liability for CSAM Generation

Drawing directly on the EU AI Act's May 2026 amendment, India should enact a statutory provision establishing liability for AI model providers and deployers in relation to CSAM generation. The amendment to the IT Act or POCSO should adopt the provider-deployer distinction with appropriate adaptation.

 [ISSN 2581-5369]

A provider of a generative AI system should be liable for AI-generated CSAM where CSAM generation is the system's intended purpose, or where CSAM output is a reasonably foreseeable and reproducible result of the system without significant technical modification. This standard holds developers accountable for foreseeable misuse without requiring proof of intent. A deployer should be liable where the deployer intentionally uses a system to generate CSAM, including by circumventing the provider's safety measures. Accidental generation should be expressly excluded, for the same reasons the EU amendment excludes it: strict liability for inadvertent outputs would deter beneficial AI development without protecting children.

For open-source models, a risk-based framework is more appropriate than blanket developer liability. Developer liability should attach where the model was specifically optimised for CSAM generation, or where the developer had actual knowledge of widespread fine-tuning for CSAM purposes and failed to take available mitigation steps. Liability should not attach to open-source releases where the CSAM capability arises through fine-tuning that the developer could not reasonably have foreseen.

The definition of CSAM in both the IT Act and POCSO must be amended explicitly to include AI-generated imagery regardless of whether a real child was involved in its creation. Purposive interpretation supports inclusion, but statutory clarity forecloses litigation.

C. Proposal 3: A Domestic Reporting Infrastructure

India's dependence on NCMEC's CyberTipline is a structural vulnerability. Intelligence about CSAM on Indian platforms and Indian accounts flows to a US non-governmental organisation before it reaches Indian law enforcement. This arrangement is acceptable as a global coordination mechanism. It is not acceptable as the primary domestic intelligence system for a country of India's scale.

Three institutional investments follow from this. A domestic CSAM reporting mechanism should be established, either as a division of the National Crime Records Bureau or as a dedicated unit under MeitY, to which platforms report confirmed detections. This unit should maintain a national hash database that contributes to and receives from the NCMEC and IWF global infrastructure. India should engage with the Budapest Convention on Cybercrime through formal accession or through negotiated bilateral mutual legal assistance arrangements with key partner jurisdictions. The sovereignty concerns that have prevented ratification are real, but they are being used to justify complete non-participation in the primary international cooperation framework for cybercrime, and that tradeoff harms Indian investigators far more than it protects Indian sovereignty. Finally, dedicated child protection units in law enforcement,

 [ISSN 2581-5369]

of the kind Telangana has developed, should be systematised nationally through minimum staffing and funding obligations on state governments coordinated through the national reporting mechanism.

VIII. CONCLUSION: HONEST REGULATION IN PLACE OF PERFORMANCE

The three convergent failures this paper opened with are not primarily failures of enforcement. They are failures of honesty. A regulation that requires CSAM detection from encrypted platforms without specifying how such detection is achievable is not a child protection measure. It is a performance of regulatory attention, designed to satisfy a political demand without engaging with the technical constraint that makes the demand impossible to meet in the way it is framed. A legal framework that has not addressed AI-generated CSAM is not a framework for the digital age. It is a framework for 2012 applied to a world that was transformed in 2022 and has continued transforming since. A criminal law that has not clearly criminalised synthetic imagery offers perpetrators a terminological escape route in the very cases where the harm is most scalable and most difficult to stop.

Just Rights for Children Alliance demonstrates that Indian courts are willing to read child protection legislation expansively and purposively. That willingness is valuable but cannot substitute for legislative action. A court cannot specify detection technology requirements. It cannot define the provider-deployer liability boundary for AI model developers. It cannot establish a domestic CSAM reporting authority. These are functions of legislation and executive action, and they have been left unperformed.

The EU amended its AI Act in May 2026 to explicitly prohibit AI systems that generate CSAM. That development occurred while this paper was being written. It is a reminder that international law is moving, with or without India. The gap between India's framework and what child protection in the AI era actually requires is not a gap that is shrinking on its own.

The framework this paper proposes does not promise the elimination of CSAM from the internet. No legal framework can do that. It promises something more modest and more achievable: regulatory honesty. Obligations that are calibrated to what technology can actually accomplish. A constitutional analysis applied rigorously rather than invoked rhetorically. Institutional investment in enforcement capacity that is proportional to the scale of the problem. And a legislative response to AI-generated CSAM that does not wait for the next Supreme Court petition to force the question into the open.

 [ISSN 2581-5369]

The law's job is not to perform concern. It is to close the gap between what the technology does and what the children it harms require. India's law has not done that job. This paper has tried to describe, as precisely as possible, what doing it would take.

*****

 [ISSN 2581-5369]

IX. BIBLIOGRAPHY

Legislation

- The Protection of Children from Sexual Offences Act, 2012, §§ 14, 15, 19, 20, 30, No. 32, Acts of Parliament, 2012 (India).
- The Information Technology Act, 2000, §§ 67B, 79, No. 21, Acts of Parliament, 2000 (India), as amended by the Information Technology (Amendment) Act, 2008, No. 10, Acts of Parliament, 2009 (India).
- The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, rr. 3, 4(4) (India).
- The Digital Personal Data Protection Act, 2023, § 9, No. 22, Acts of Parliament, 2023 (India); Digital Personal Data Protection Rules, 2025 (India).
- Online Safety Act 2023, c. 50 (UK).
- Communications Decency Act, 47 U.S.C. § 230 (U.S.).
- 18 U.S.C. § 2258A (U.S.).
- Regulation 2024/1689, of the European Parliament and of the Council (Artificial Intelligence Act), 2024 O.J. (L 1689) (as amended by provisional agreement of May 7, 2026 introducing the Article 5 prohibition on CSAM-generating AI systems).
- Proposed Regulation on Child Sexual Abuse (CSA Regulation, Chat Control), COM (2022) 209 final (in trilogue negotiations as of May 2026).
- Directive 2011/93/EU, of the European Parliament and of the Council, on Combating the Sexual Abuse and Sexual Exploitation of Children, 2011 O.J. (L 335) 1.

Cases

- Just Rights for Children Alliance v. S. Harish, 2024 INSC 716 (India).
- Ramkumar v. Union of India (Madras H.C. 2025) (India).
- Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 S.C.C. 1 (India).
- Avnish Bajaj v. State (NCT of Delhi), (2005) 3 Comp. L.J. 364 (India).

 [ISSN 2581-5369]

Reports and Secondary Sources

- Daniel Manoj, Ranjit Immanuel James et al., *Behind the Screens: Understanding the Gaps in India's Fight Against Online Child Sexual Abuse and Exploitation*, 4 CHILD PROT. & PRAC. 100088 (2024).
- Wilson Ctr., *Combatting AI-Generated CSAM* (Feb. 2025).
- Oxford Hum. Rts. Hub, *Reactive Laws, Relentless Abuse: Why Child Sexual Abuse Material Still Circulates Freely in India's Digital Spaces* (2025).
- *Hashing in the Fight Against CSAM: Technology at the Crossroads of Law and Ethics*, 5 J. CYBERSECURITY & PRIV. 92 (2025).
- *How Europe's Chat Control Regulation Could Compromise American Communications*, TECH POL'Y PRESS (Feb. 2026).
- Max Planck Inst., *More Monitoring, But Not More Protection* (2024).
- Internet Soc'y, *Internet Impact Brief: 2021 Indian Intermediary Guidelines* (Nov. 2021).
- Thorn & All Tech Is Human, *Safety by Design: Principles to Combat AI-Generated CSAM* (2024).
- Nat'l Telecomms. & Info. Admin., *Dual-Use Foundation Artificial Intelligence Models with Widely Available Model Weights* (2024).
- Internet Watch Found., *Annual Report 2024*.
- Nat'l Ctr. for Missing & Exploited Children, *CyberTipline Data 2025*.
- Glob. Encryption Coal., *Joint Statement on the EU CSAM Regulation* (May 2024).
- *Safe Harbor Under Section 79 IT Act: Why Legal Immunity for Platforms Is Crumbling*, KSANDK (July 2025).
- *Does the Traceability Requirement Meet the Puttaswamy Test?*, THE DIALOGUE (Sept. 2023).
- *An Analysis of Puttaswamy: The Supreme Court's Privacy Verdict*, INDRASTRA GLOB. (Nov. 2017).

*****

 [ISSN 2581-5369]