Journal of International Commercial Law and Technology
Print ISSN: 1901-8401

Website: https://www.jiclt.com/

Research Article

Legal Incentives and Automated Moderation A Role–Error Model for Synthetic-Content Intermediary Liability

Article History

Name of Authors:
Ayan Roy Chowdhury[1*]

Affiliation:
[1]Ph.D. Candidate, NMIMS, Mumbai, Maharashtra, India

Corresponding Author
Ayan Roy Chowdhury

ayanroychowdhury017@gmail.com

How to cite this article:
Chowdhury, Ayan Roy. Legal Incentives and Automated Moderation A Role–Error Model for Synthetic-Content Intermediary Liability. J. Int. Commer Law Technol., 2026;7(1): 1596–1606,

Received: 22-06-2026
Revised: 19-07-2026
Accepted: 17-08-2026
Published: 01-09-2026

Doi: 10.61336/Jiclt/20267106

Abstract: Legal rules governing synthetic media increasingly depend on automated classification, yet the law rarely specifies how liability weights alter the operating threshold of those systems. This article develops a conceptual role–error model for intermediary liability and compares hosting, curation or amplification and generation enablement across India, the European Union, the United States, Germany and the United Kingdom. Methodologically, it combines doctrinal analysis, functional comparative law and institutional design. The model treats moderation as a threshold choice between the expected costs of false negatives and false positives, showing how compressed deadlines and one-sided penalties create a rational incentive to remove lawful satire, criticism, journalism and other context-dependent speech. The comparative findings indicate that the classical safe harbour remains defensible for genuinely neutral hosts but becomes underinclusive when a service control reach or supplies the means of generation. The article proposes a functional three-tier framework: specific-knowledge duties for neutral hosts; provenance preservation, reason-giving, appeal and proportionate demotion duties for curators; and source marking, consent verification and generation-log duties for synthetic-content enablers. Its principal contribution is to connect legal allocation of responsibility with classifier operating thresholds and the constitutional cost of moderation error. The model is conceptual and should be tested empirically using transparency reports, provenance-persistence studies and platform appeal data.

Keywords: Intermediary Liability, Synthetic Media, Automated Moderation, Safe Harbour, Platform Governance, Provenance

INTRODUCTION

For two decades the audiovisual internet rested upon a bargain. Platforms would not be treated as the authors of what their users posted, provided that they removed unlawful material once they knew of it. That bargain, expressed in the conditional safe harbour, is now being renegotiated in most major jurisdictions simultaneously and the pressure comes from a single source. Generative artificial intelligence has made synthetic audiovisual content cheap, convincing and ubiquitous. When a platform can no longer assume that what it hosts corresponds to anything that occurred in the world, a model of liability triggered by knowledge of specific items begins to strain [1-4].

India has responded more decisively than most. The exemption conferred by section 79 of the Information Technology Act conditions immunity upon due diligence and compliance with lawful orders, a structure long read in parallel with the notice-and-takedown regime of the United States. The Supreme Court fixed the trigger for the duty to act at actual knowledge received through a court order or a government notification, a standard reinforced for copyright by the requirement of knowledge of specific infringing material and by the holding that an intermediary bears no obligation to screen all content before publication. The 2026 amendments to the intermediary rules unsettle that settlement by moving platforms from reaction toward anticipation [5,6].

This paper compares three ways of recalibrating the safe harbour for synthetic content and asks which best reconciles three goods that stand in tension. The first is effective redress for those harmed by synthetic audiovisual media, whose injuries range from commercial misappropriation to fabricated intimate imagery. The second is the operational viability of hosting user content at scale, which no jurisdiction wishes to destroy. The third is the interest in satire, parody, criticism and news reporting, which imperfect filters predictably suppress. The paper argues that the Indian architecture, while responsive, is miscalibrated and that a tiered and role-based model is the sounder design [7,8].

It is worth asking why the intermediary has become the principal object of regulation in this field, since the answer illuminates both the appeal and the limits of the approach. The originator of harmful synthetic content is frequently anonymous, is often outside the jurisdiction and possesses no assets against which a judgment could be enforced. The platform, by contrast, is identifiable, is subject to regulatory pressure and is capable of acting at scale. Regulation therefore gravitates toward the party that can act rather than toward the party that is culpable and the structural consequence is that the intermediary bears duties in respect of conduct it neither performed nor authorised. The scholarship has described the resulting configuration as a triangular relationship in which the state regulates speech indirectly by regulating the infrastructure through which speech travels and that framing explains why the protection of the speaker requires attention that a bilateral state-and-platform model does not naturally supply.

Part 2 sets out the classical safe harbour and the assumption upon which it rested. Part 3 examines the European layered regime. Part 4 considers the United States and the limits of the reactive model. Part 5 analyses the Indian rules of 2026. Part 6 identifies the doctrinal collision with the actual-knowledge standard and the structural incentive toward over-removal. Part 7 proposes a tiered model. Part 8 concludes [9,10].

MATERIALS AND METHODS

The paper addresses three questions. First, when synthetic media makes unlawfulness probabilistic and context-dependent, which features of the classical safe harbour cease to function as intended? Second, how do the European Union, the United States and India distribute the risks of non-removal and wrongful removal among platforms, complainants, speakers and persons depicted? Third, can intermediary duties be redesigned so that legal responsibility tracks a service's actual role in generating, curating or merely hosting the impugned content [11,12]?

Methodologically, the study combines doctrinal analysis with functional comparative law and institutional design. It examines statutory texts, leading judicial decisions and regulatory instruments in India, the European Union, the United States, Germany and the United Kingdom. The comparison is functional rather than taxonomic, each regime is assessed by the legal trigger for intervention, the actor required to decide, the time available for contextual assessment, the procedural protection afforded to the uploader and the allocation of error costs. The proposed model is then tested against constitutional proportionality, operational feasibility and the technical limits of synthetic-content detection [13].

The paper's original contribution is twofold. It identifies the safe harbour's hidden factual premise that unlawful material is a discrete, observable subset of an otherwise lawful stream—and explains why generative systems destabilise that premise. It then develops a functional three-tier model in which duties attach separately to hosting, curation and generation. This avoids treating a diversified platform as a single legal category and makes capability, rather than corporate identity, the organising principle of liability [14].

Literature and Analytical Framework

The paper sits at the intersection of three bodies of scholarship. The first concerns intermediary liability and the constitutional risks of private speech governance, including the triangular account of platform regulation and the literature on notice-and-takedown over-removal. The second examines differentiated platform responsibility under the Digital Services Act, paper 17 of the Digital Single Market Directive and related jurisprudence. The third addresses provenance, automated moderation and the technical limits of classifiers used to identify manipulated or synthetic media [15].

Existing work commonly asks whether platforms should receive more or less immunity. That binary framing obscures the fact that the same service may perform legally distinct functions, it may store a file, rank it, recommend it, monetise it and supply the generative tool that produced it. The analysis therefore uses role differentiation and error-cost asymmetry as its two organising concepts. Role differentiation asks what the intermediary actually did in relation to the content. Error-cost analysis asks which moderation mistake the law makes expensive and which mistake it leaves effectively costless [16].

This framework also explains why accuracy figures cannot resolve the legal problem by themselves. The incidence of false positives depends on the base rate of unlawful synthetic content, while the operating threshold chosen by a platform responds to the liability structure. Law therefore shapes the technical system's behaviour even when it does not specify the model or classifier.

A Formal Role–Error Model

For analytical purposes, let a platform function f belong to one of three operational classes, H (hosting), C (curation or amplification) and G (generation enablement). Let FN denote the expected cost of retaining unlawful synthetic content and FP the expected cost of removing lawful content. The platform's moderation threshold is not determined by detector accuracy alone; it responds to the ratio of legally imposed FN and FP costs. Where liability attaches almost exclusively to FN, the rational threshold shifts toward removal, increasing FP even when the classifier improves [17].

The model yields a regulatory proposition, obligations should increase with control over creation and reach, while countervailing procedural duties should increase with the probability and social cost of FP. Thus, H attracts specific-knowledge duties, C attracts provenance preservation, reason-giving and proportionate reach-reduction duties and G attracts source marking, consent verification and traceability obligations. This is a conceptual model rather than an empirical validation. Its value lies in making the legal assumptions testable and in identifying variables for future platform-data or simulation-based research [18].

The Classical Safe Harbour and the Assumption Beneath It

The conditional safe harbour was a deliberate allocation of risk. Rather than making platforms strictly liable for the wrongs of their users, which would have rendered large-scale hosting uninsurable, or wholly immune, which would have left victims without a practical remedy, the notice-and-takedown model divided the difference. The platform enjoys immunity so long as it lacks knowledge of specific unlawful material and acts expeditiously once it acquires such knowledge. European law reached a comparable position by conferring the exemption upon providers whose conduct is neutral and merely technical and by prohibiting the imposition of general monitoring obligations.

Indian law embodies the same compromise. Section 79 confers the exemption subject to due diligence and Shreya Singhal supplied its constitutional discipline by requiring that the knowledge triggering the duty to act arrive through a court order or a government notification rather than through the assertion of a private complainant. That requirement was not a technicality. It was a safeguard against the delegation of adjudicative power to private parties and it reflected the recognition that a platform faced with an unverified complaint and the prospect of liability will remove first and consider afterwards [19].

The logic of the settlement depends upon an assumption that is now failing, namely that unlawful content constitutes a discrete and identifiable subset of an otherwise lawful stream, capable of being flagged case by case. Synthetic media defeats the assumption in two ways. It is generated at a volume that outpaces case-by-case notification and it is designed to be indistinguishable from authentic material, so that identifying it requires probabilistic detection rather than observation. A regime built for the knowing removal of specific items is poorly suited to the anticipatory governance of a content type defined by its capacity to deceive [20].

A further strand of the Indian position concerns copyright specifically and it operates alongside the general exemption rather than within it. The Copyright Act contemplates infringement by permitting a place to be used for the communication of infringing works, subject to a defence of absence of knowledge and reasonable belief and it exempts transient or incidental storage arising in the course of electronic transmission or the provision of links. The Delhi High Court reconciled these provisions with the general exemption by holding that liability requires knowledge of specific infringing material rather than general awareness that infringement occurs upon a platform. The convergence of the copyright and information-technology regimes upon a specific-knowledge standard is significant, because it means that the shift toward anticipatory duties disturbs not one settled position but two [21].

It is worth noting how far this settlement was the product of deliberate constitutional design rather than of legislative drafting. The statutory text conditions the exemption upon due diligence and upon expeditious action following receipt of actual knowledge and read literally it might have permitted a private complainant to trigger the duty. The Supreme Court declined that reading precisely because it would have placed the determination of unlawfulness in private hands, with the predictable consequence that intermediaries would remove upon assertion. The point deserves emphasis because the same consequence now arises from a different mechanism. Where detection duties and compressed timelines make assessment impossible, the practical result is removal upon assertion, achieved through operational necessity rather than through the interpretation that the Court rejected [22].

Two further features of synthetic content compound the difficulty. The first is that unlawfulness frequently depends upon context rather than upon the content itself. A fabricated video of a public figure is unlawful if presented as authentic and lawful if presented as satire and the distinguishing feature is often the framing rather than anything within the file. The second is that the harm is frequently done before any notice can operate. A fabricated intimate image or a synthetic statement attributed to a political figure achieves its effect within hours, so that a remedy operating upon notice arrives after the injury is complete. These features explain why legislatures have reached for anticipatory duties and Figure 1 traces that migration [23].

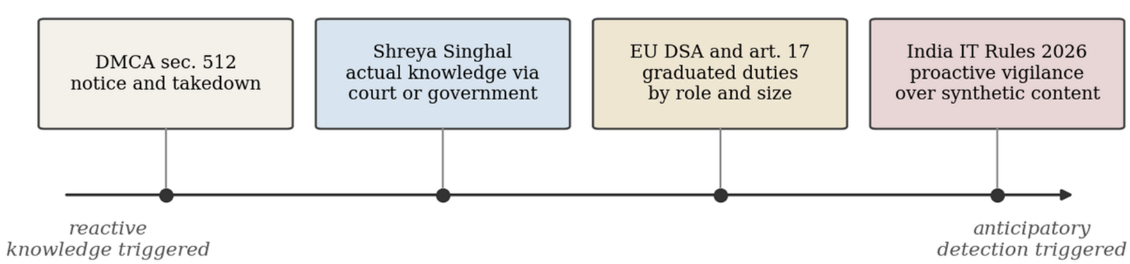

Figure 1: The Migration of the Safe Harbour from Knowledge to Anticipation

The European Layered Regime

The European response is best understood as the differentiation of a single exemption into a graduated structure of duties. The Digital Services Act retains the underlying exemption while imposing obligations that scale with the role and the size of the service, ranging from notice-and-action mechanisms applicable to all hosting providers through to systemic risk assessment and mitigation for very large platforms. It also constitutes a procedural architecture, requiring internal complaint handling and providing for out-of-court dispute settlement, which addresses the position of the user whose content has been removed rather than only that of the complainant [24].

Alongside this general regime, paper 17 of the Digital Single Market Directive imposes a specific liability structure upon platforms hosting user-uploaded protected content, requiring best efforts to obtain authorization and to prevent the availability of notified works. The provision was challenged as incompatible with freedom of expression and the Court of Justice upheld it while making clear that the obligation must be accompanied by safeguards precluding the blocking of lawful content and that filtering systems which cannot distinguish lawful from unlawful uses would not satisfy the requirement. That qualification is the single most important element of European doctrine for present purposes, since it establishes that anticipatory obligations are permissible only where the technology deployed is capable of the distinction the law requires [25].

The Court has taken a comparable approach to injunctions, permitting orders extending to identical and equivalent content while confining the use of automated techniques to cases not requiring an independent assessment of the material. It has also clarified the conditions upon which a platform ceases to be a mere host and performs an act of communication to the public in its own right, which turns upon knowledge and upon the provision of tools specifically facilitating unlawful sharing. The resulting picture is one in which the safe harbour survives but is differentiated by conduct and in which the intensity of the duty tracks the platform's actual involvement [26].

The Artificial Intelligence Act adds a third layer directed specifically at synthetic content, imposing transparency obligations including the disclosure that material has been artificially generated or manipulated. The significance of this layer is that it operates upon a different variable from the others. The Digital Services Act and paper 17 regulate what a platform must do about content, whereas the transparency obligation regulates what the content must disclose about itself. In principle the second reduces the burden of the first, since reliably marked synthetic content is far easier to govern than unmarked content and the provenance standards emerging in parallel supply the technical means [27,28].

Two aspects of the European approach are worth isolating because they are absent from the Indian framework and are load-bearing in the European one. The first is that obligations attach principally to systems rather than to items. A very large platform is required to assess and mitigate systemic risks arising from the design and operation of its service, which is a duty discharged by demonstrating adequate processes rather than by achieving a particular outcome in respect of any individual piece of content. A duty framed in this way does not create the same incentive toward precautionary removal, because compliance is not established by removing more [29].

The second aspect is that the user whose content is removed is a participant in the framework rather than merely its object. The complaint-handling and dispute-settlement provisions give that user a defined procedural position, which serves two functions. It supplies redress in the individual case, and, more importantly for the argument advanced here, it imposes a cost upon wrongful removal that partially offsets the asymmetry examined in Part 6. A platform that must handle appeals, give reasons and submit to external dispute resolution faces a price for over-removal and it is that price rather than any exhortation to proportionality which disciplines the incentive [30].

The United States and the Limits of the Reactive Model

The American position warrants separate treatment, both because it is the origin of the notice-and-takedown template and because it has proved the least adaptable of the three. Two distinct immunities operate. The copyright safe harbour conditions protection upon expeditious removal following notification and upon the absence of specific knowledge, while a broader immunity insulates providers from liability in respect of information provided by others. The breadth of the second has meant that pressure for reform has been channelled almost entirely through legislative proposals rather than through doctrinal development [31-33].

Recent jurisprudence has nevertheless clarified two points of comparative relevance. The Supreme Court declined to hold platforms liable for algorithmic recommendation upon the theories advanced before it, leaving the question of curation largely to legislative resolution. Separately and more significantly for the model proposed in Part 7, the Court has treated editorial curation by platforms as itself expressive activity attracting constitutional protection. That holding complicates any attempt to impose duties upon curation in the American context and it explains why the tiered approach developed here is more readily implemented in jurisdictions whose constitutional traditions permit proportionate regulation of intermediaries [34,35].

The American experience also supplies the clearest evidence of the over-removal dynamic examined in Part 6. Empirical study of notice-and-takedown practice has repeatedly documented removal of lawful material upon unfounded or abusive notification and the courts have responded by requiring a complainant to consider whether the use is permitted before issuing a notification. That requirement is a modest but instructive corrective, since it places a substantive obligation upon the party initiating removal rather than upon the intermediary receiving the demand and Part 7 adopts it [36,37].

Other jurisdictions supply further variations. Germany pioneered the imposition of short removal periods backed by administrative penalty, an approach criticised for the incentive it created toward precautionary removal and instructive precisely for that reason. The United Kingdom has adopted a duty-of-care model resting upon systems and processes rather than upon individual items and European audiovisual regulation had already imposed obligations upon video-sharing platforms before the general instruments arrived. The comparative record therefore contains a reasonably clear lesson, namely that regimes which impose short deadlines without procedural safeguards produce over-removal, while those which regulate systems rather than items produce less of it [38-40].

India's 2026 Recalibration

The Indian amendments introduce a defined category of synthetically generated information and attach to it a series of obligations. Content of that character must be labelled, provenance metadata must be embedded and preserved, significant social media intermediaries must deploy verification measures directed at identifying such content and removal upon notification must be effected within periods considerably shorter than those previously applicable. The amendments also clarify that removal or disablement effected in compliance with the rules does not by itself defeat the statutory exemption, which addresses a genuine uncertainty under the earlier framework as to whether proactive moderation compromised an intermediary's neutrality [41,42].

Three features of the recalibration deserve analysis. The first is the shift in the trigger. Under the previous architecture the duty to act arose upon knowledge and the 2021 rules constituted a due-diligence framework operating largely upon that basis. The new obligations attach irrespective of notification, requiring the intermediary to identify a class of content by its technical characteristics. This is not a marginal extension of the duty of care but a change in its nature, from responding to what one has been told to detecting what one has not [43].

The second feature is the compression of time. Very short removal periods are effective against harms that propagate rapidly, which is a genuine and important consideration for fabricated intimate imagery and for synthetic political content in an election period. They also, however, eliminate the interval within which any assessment of lawfulness could occur. A platform required to act within a few hours cannot evaluate whether the impugned material is satire, criticism or news reporting and will therefore not attempt to do so.

The third feature is the conditioning of immunity upon detection. Where the safe harbour depends upon the deployment of measures to identify synthetic content, the platform's legal position becomes a function of the accuracy of its detection tools. The empirical literature upon automated content detection consistently reports substantial error rates, particularly where the assessment requires contextual judgment rather than matching against a known reference and synthetic-content detection is presently among the least reliable applications of the technology. Conditioning a legal exemption upon a capability that does not reliably exist transfers the risk of technical failure onto the intermediary, which will respond in the manner that Part 6 describes [44].

A fourth feature concerns the interaction with the existing blocking machinery. Indian law already provides for the blocking of public access to information through a statutory procedure involving governmental determination and the courts have considered the limits of directions requiring automated blocking in the context of prohibited advertising. The new obligations do not displace that machinery but operate alongside it, with the consequence that synthetic content may be removed through three distinct routes, namely governmental blocking, court-ordered removal and platform detection under the amendments. The multiplication of routes is not objectionable in itself. What is objectionable is that only the first two involve any determination by a public authority, while the third, which will in practice account for the overwhelming majority of removals, involves none [45,46].

A fifth feature concerns the position of the person depicted, which the framework addresses only indirectly. The obligations run between the state and the intermediary and a person whose likeness has been synthetically appropriated is a beneficiary of the framework rather than a participant in it. That person acquires no defined entitlement, no right to require removal upon proof of non-consent and no remedy where removal does not occur. Indian law protects narrow categories of such conduct through criminal provisions addressing voyeurism and the transmission of private images and data-protection legislation supplies a further partial route, yet neither supplies a general civil entitlement in one's own likeness. The regulatory turn therefore governs distribution comprehensively while leaving the underlying right undefined, which is a gap that intermediary regulation cannot itself repair [47].

The Doctrinal Collision and the Structure of Error

The first difficulty is doctrinal. Shreya Singhal read down the statutory obligation precisely in order to prevent private assertion from operating as the trigger for removal, holding that the knowledge required must come through a court order or a government notification. Subsequent authority confirmed that an intermediary is not obliged to screen material before publication. An obligation to deploy measures identifying synthetic content and to remove within periods that preclude assessment, is difficult to reconcile with either holding. The clarification that compliant removal does not defeat the exemption resolves the intermediary's uncertainty about the consequences of removing and does nothing at all to protect the user whose lawful content is removed.

It may be argued that the earlier authorities addressed a different problem, being unlawfulness that is contestable and content-dependent, whereas the synthetic character of a file is a technical fact susceptible of objective determination. The argument has some force in respect of labelling obligations, which require the identification of provenance rather than the assessment of legality. It has none in respect of removal obligations, because whether synthetic content is unlawful is precisely a contestable and context-dependent question. The distinction between marking and removing is therefore constitutionally significant and a framework that imposes both should not assume that the justification for the first extends to the second.

The second difficulty is structural and concerns the asymmetry of error costs. Any moderation system produces two kinds of error, namely the removal of lawful content and the retention of unlawful content. Figure 2 represents the resulting matrix.

The legal architecture attaches consequences to only one of these errors. A platform that retains unlawful content loses its exemption and faces liability, whereas a platform that removes lawful content faces, in most jurisdictions and certainly in India, no comparable sanction. Where the costs of the two errors are asymmetric in this way, the rational platform will minimise exposure by erring toward removal and will do so more strongly as the compliance period shortens and as detection accuracy falls. The over-removal that follows is not a failure of implementation to be corrected by better guidance. It is the predicted response to the incentive structure that the law creates and the empirical literature on notice-and-takedown regimes documents it consistently.

A fourth difficulty is jurisdictional and has attracted less attention than it deserves. Synthetic content is generated, hosted and consumed across borders and Indian courts have responded by granting relief extending beyond the territory where the material was uploaded from within India. The European Court of Justice reached a comparable position in permitting injunctions to extend to identical and equivalent content, while confining automated identification to material that does not require independent assessment. The qualification matters here for the same reason it matters domestically, since equivalence assessed by a classifier is precisely the operation whose error characteristics Part 6 has examined. An obligation to prevent the reappearance of equivalent synthetic content, enforced across jurisdictions and implemented by automated matching, compounds the over-removal problem rather than confining it, because a single erroneous classification propagates to every subsequent upload it resembles.

The magnitude of the resulting over-removal depends upon two variables that deserve to be stated explicitly, because policy discussion frequently proceeds as though detection accuracy alone were determinative. The first is the base rate. Where unlawful synthetic content forms a small proportion of the total synthetic content in circulation and lawful synthetic satire, advertising, entertainment and educational material forms the remainder, even a highly accurate classifier will generate false positives that substantially outnumber true positives. This is an arithmetic consequence of the base rate rather than a deficiency of the technology and it means that improvements in detection accuracy, while valuable, cannot by themselves resolve the difficulty.

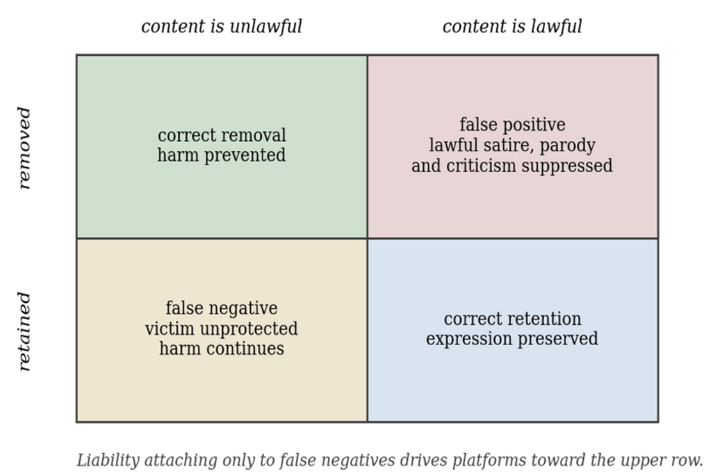

Figure 2: Error Costs in Synthetic-Content Moderation

The second variable is the threshold at which the platform sets its classifier, which is not a technical constant but a business decision responding to the legal incentive. A platform facing liability for retention and none for removal will set the threshold so as to minimize false negatives, accepting a high rate of false positives as the cost of compliance. The legal framework therefore determines the operating point of the technology and a framework that attaches consequences to only one error will drive the threshold toward the suppression of lawful material regardless of how accurate the underlying classifier becomes. It follows that the remedy lies in the structure of liability rather than in the state of the art, which is the central design claim of this paper.

The categories most exposed are identifiable in advance. Satire and parody are synthetic by construction and are distinguished from deception only by framing. Political criticism attracts complaints from those criticized, who under a rapid-removal regime need only assert that the content is synthetic. Investigative journalism concerning synthetic media must frequently reproduce the material it examines. Artistic and educational uses share the same characteristics. In each case the content is technically synthetic, its lawfulness depends upon context and the compliance period allows no time for the contextual assessment upon which lawfulness turns.

Indian constitutional doctrine supplies the standard against which this must be measured. Restrictions upon expression must fall within the permitted grounds and must be proportionate and proportionality has been elaborated as the governing standard for intrusions upon protected interests generally. A framework that predictably suppresses lawful expression as a systemic consequence of its incentive structure and that provides no procedural mechanism by which the affected speaker may be heard, is vulnerable upon that standard. The comparative position reinforces the point, since the European instrument was upheld only upon the express condition that lawful content be safeguarded and the international soft-law consensus has consistently required notice to the affected user, reasons and an avenue of appeal.

A third difficulty concerns institutional competence. The determination whether a particular synthetic depiction is satire or deception is an adjudicative judgment. Under the present architecture it is made by a platform, under time pressure, without hearing the speaker and with a financial incentive pointing in one direction. That is not a criticism of platforms, which have neither the mandate nor the procedural apparatus for adjudication. It is an argument that the architecture allocates the judgment to the wrong actor and that any recalibration should either return the judgment to a public body or supply the procedural safeguards that make private determination tolerable.

A Tiered and Role-Based Model

The alternative proposed here rests upon a single principle, namely that the intensity of an intermediary's duty should track its actual capability to govern the content in question and its actual role in bringing that content into circulation. European law has moved in this direction by differentiating obligations according to role and scale and Indian law has an indigenous foundation for the same approach in the distinction between active and passive intermediaries developed in the electronic-commerce context. Figure 3 sets out the resulting structure.

Tier one comprises the genuinely neutral host, being a service that stores or transmits material without curating, ranking or amplifying it. For such a service the classical settlement should be preserved substantially intact.

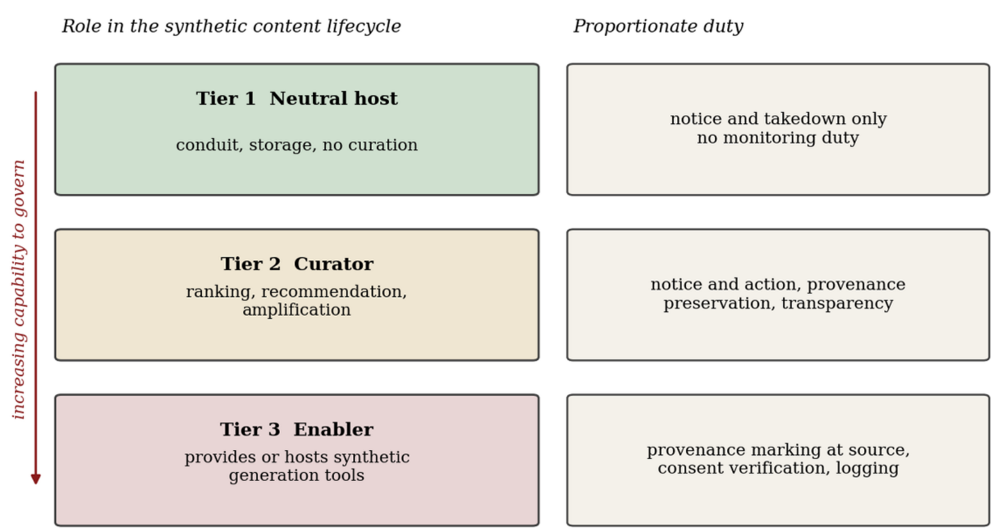

Figure 3: A Tiered, Role-Based Model of Intermediary Duty

The duty should arise upon actual knowledge in the sense established by Shreya Singhal, no general monitoring obligation should be imposed and removal periods should allow time for the assessment of manifestly contestable claims. The justification is straightforward, since a service that does not select content has neither the informational advantage nor the operational capacity that would make anticipatory duties reasonable.

Tier two comprises the curator, being a service that ranks, recommends or amplifies content and thereby determines what its users see. Such a service exercises editorial functions in substance and comparative jurisprudence has increasingly recognized curation as a distinct activity attracting distinct treatment. Its duties should accordingly be greater, extending to notice-and-action mechanisms with defined timelines, preservation of provenance metadata rather than the routine stripping that presently accompanies transcoding, transparency reporting sufficient to permit external evaluation and a duty not to amplify content that has been flagged as synthetic pending assessment. The last of these deserves emphasis, because it is a proportionate intermediate measure. Demotion is materially less restrictive of expression than removal, while being substantially effective against harms whose gravity depends upon reach.

Tier three comprises the enabler, being a service that provides or hosts the tools by which synthetic content is generated. Such a service stands in a different relation to the harm, since it does not merely host material created elsewhere but supplies the means of its creation and it is uniquely placed to act at the point of generation. Its duties should include provenance marking at source, so that synthetic content enters circulation already identified, verification of consent where the output depicts an identifiable person and retention of generation logs sufficient to permit identification of the originator where unlawful material is subsequently traced. Duties of this kind are not available to downstream hosts at any price and imposing them at the point of generation is both more effective and less restrictive than requiring detection after the fact.

Three cross-cutting elements complete the model. The first is procedural, requiring that removal be accompanied by notice to the uploader, a statement of reasons and an avenue of appeal, following the structure adopted in the European instrument and the international consensus. The second is a good-faith consideration requirement imposed upon complainants, obliging a party issuing a notification to consider whether the material falls within an established exception before doing so, on the model developed in American copyright jurisprudence. The third is a counter-incentive addressing the asymmetry identified in Part 6, whether through a remedy for wrongful removal, a transparency obligation exposing removal rates and reversal rates to public scrutiny, or a safe harbour for good-faith retention pending assessment analogous to the protection afforded elsewhere for good-faith restriction.

Implementation raises three practical questions on which the model should be explicit. The first concerns the treatment of services that occupy more than one tier, which will be the common case, since a large platform typically hosts material, curates a feed and increasingly offers generative tools. The appropriate response is to apply the tiers functionally rather than to classify the entity as a whole, so that the same service bears tier-three duties in respect of the content its own tools generate, tier-two duties in respect of what it amplifies and tier-one duties in respect of what it merely stores. Functional application is more demanding to administer than entity classification and it is the only approach that matches duty to capability, which is the organizing principle of the model.

The second question concerns small services, for which even tier-two duties may be disproportionate. The European instrument addresses this through exemptions calibrated to size and a comparable calibration is appropriate here, with the important qualification that the exemption should attach to the burden of systemic obligations rather than to the procedural safeguards owed to users. A small platform may reasonably be excused from transparency reporting. It should not be excused from giving notice and reasons to a user whose content it removes, since that obligation protects a third party rather than reducing the platform's costs.

The third question concerns transition. Detection capability and provenance infrastructure are both developing and duties defined by reference to the current state of either will be obsolete quickly. The framework should therefore express tier-three obligations in terms of outcomes, requiring that generated content carry provenance information conforming to a prescribed standard, while leaving the standard itself to subordinate instruments capable of revision. A staged commencement is likewise appropriate, with procedural safeguards and provenance marking taking effect first, since neither depends upon detection capability and detection-contingent duties commencing only upon a determination that the necessary capability exists.

The model may finally be tested against the three goods identified at the outset. Effective redress is improved, because tier-three duties operate at the point of generation where prevention is possible rather than at the point of hosting where only removal is available and because the person depicted acquires a determinate route to action rather than depending upon a platform's discretionary assessment. Operational viability is preserved, because duties are matched to capability and no service is required to do what it cannot do. And the expression interest is protected structurally rather than by exhortation, through demotion as an intermediate measure, procedural safeguards for the affected speaker and a counterweight to the asymmetry of error costs. No framework will eliminate the tension among the three, since they genuinely conflict. The claim advanced here is the more modest one that a framework matching duty to role manages the tension better than one imposing uniform detection obligations upon services whose capacities differ profoundly.

Two objections merit response. The first is that tiering invites classification disputes, since services will contend that they belong in a lower tier. The objection is real and is answered by defining the tiers by reference to observable conduct rather than to self-description, in the manner that European and Indian jurisprudence has already developed for the active and passive distinction. The second objection is that a tiered structure is administratively demanding for a regulator with limited capacity. That is true and it is an argument for defining the tiers in the primary rules rather than leaving classification to case-by-case determination and for concentrating supervisory attention upon the third tier, which is both the smallest class and the one where intervention is most effective.

The model also has an international dimension that a purely domestic recalibration lacks. Generation frequently occurs in one jurisdiction, hosting in another and harm in a third and duties imposed upon downstream hosts alone are readily evaded by relocation. Provenance marking at the point of generation, by contrast, travels with the content across jurisdictions, which is why the alignment of national frameworks with emerging technical standards matters more than the stringency of any individual national rule.

Limitations and Future Research

This is a doctrinal and institutional-design study. It does not report platform-level removal data, classifier benchmarks or interviews with moderators, complainants and affected users. Publicly available evidence on synthetic-content detection remains heterogeneous and platform systems are proprietary and change rapidly. The paper therefore does not claim that a single error rate applies across services or content types. Its narrower claim is structural, where the legal consequences of false negatives substantially exceed those of false positives, a rational intermediary will move the operating threshold toward removal. Future research should test the model using transparency-report data, controlled studies of provenance persistence and interviews concerning escalation and appeal practices. A submission to an AI-and-law venue would be materially strengthened by a reproducible simulation, a labelled audiovisual test set, or an empirical study of how legal penalty weights alter classifier thresholds and appeal outcomes (Table 1).

The classical safe harbour rested upon an assumption that synthetic media has falsified, namely that unlawful content forms a discrete and identifiable subset of an otherwise lawful stream. Legislatures are right to conclude that a purely reactive model cannot govern content generated at scale and designed to deceive and the Indian recalibration responds to a genuine and urgent problem. The difficulty lies not in the recognition of that problem but in the design of the response.

Conditioning a legal exemption upon the accuracy of detection technology that is not yet accurate, while compressing compliance periods below the interval required for any assessment of lawfulness and attaching consequences to only one of the two errors that moderation can produce, together generate a predictable outcome. Platforms will remove lawful satire, parody, criticism and journalism, not through misunderstanding but through rational response to the incentives the framework creates. The constitutional standard of proportionality, the express condition upon which the comparable European obligation was upheld and the international consensus upon procedural safeguards all indicate that this outcome is not merely unfortunate but legally vulnerable.

A tiered and role-based model offers a better settlement. It preserves the classical exemption for services that genuinely do no more than host, imposes intermediate and proportionate duties upon those that curate and amplify and concentrates the most demanding obligations upon those that supply the means of generation and are uniquely able to mark content at source.

Table 1: Comparative Allocation of Synthetic-Content Governance

Regime	Primary trigger	Decision-maker	Ex ante duty	User safeguards	Dominant risk
EU layered model	Notice; systemic risk; notified works	Platform under statutory procedures and regulator oversight	Role- and scale-sensitive	Reasons, complaint handling, external settlement	Complexity and compliance burden
United States	Specific notice or knowledge	Platform, followed by courts	Limited; mainly reactive	Counter-notice and limited complainant duties	Delay and under-removal
India 2026	SGI identification, notice and lawful directions	Platform under compressed timelines	Verification, labelling and rapid action	Comparatively underdeveloped	Precautionary over-removal
Proposed model	Function performed in relation to content	Platform with review and appeal safeguards	Capability-matched by function	Notice, reasons, appeal and counter-incentives	Classification and administrative cost

Coupled with procedural safeguards for affected speakers, a good-faith requirement for complainants and a counterweight to the asymmetry of error costs, such a model would hold intermediaries accountable for the synthetic content they are actually placed to govern, without conscripting them into the suppression of lawful expression.

A final observation concerns the relationship between the framework proposed here and the copyright dimension of the same problem. Much synthetic audiovisual content is infringing as well as deceptive, since it frequently incorporates protected material or reproduces protected expression and the copyright safe harbour operates alongside the general exemption. The tiered model translates readily into that setting. A neutral host should continue to enjoy protection upon the specific-knowledge standard, a curator that amplifies infringing material should bear notice-and-action duties calibrated to its editorial role and a service supplying generative tools stands in a different relation again, since it is placed to prevent infringement at the point of creation rather than to remove it afterwards. Treating the two regimes as a single architecture is preferable to the present position, in which a platform faces one standard for copyright and another for synthetic content, with the more demanding of the two setting its operating threshold for both.

The wider lesson concerns the sequencing of regulatory attention. Intermediary liability is the instrument that legislatures reach for first, because platforms are visible, are within the jurisdiction and are capable of acting. It is not, however, the instrument best suited to every dimension of the problem. The harms that synthetic media inflicts upon identifiable persons require a defined entitlement in the person, which intermediary rules do not supply. The harms it inflicts upon right holders require licensing and remuneration mechanisms, which intermediary rules likewise do not supply. Recalibrating the safe harbour is necessary and is not sufficient and a framework relying upon it alone will impose substantial compliance burdens while leaving the underlying entitlements as undefined as it found them.

The safe harbour does not need to be abolished. It needs to be differentiated.

Declarations

Use of generative AI. Generative-AI tools were used for language editing, structural comparison and formatting assistance. All legal propositions, sources, citations, interpretations and conclusions remain the author's responsibility.

Preprint notice. The manuscript has not been published and is not under consideration elsewhere. Any preprint deposit will be disclosed in accordance with the journal's policy.

REFERENCES

1. *17 U.S.C. § 512*. 2018. "Limitations on Liability Relating to Material Online."
2. *47 U.S.C. § 230*. 2018.
3. *Arijit Singh v. Codible Ventures LLP*, Interim Application (L) No. 23724 of 2024, Bom. H.C., 26 July 2024; *Anil Kapoor v. Simply Life India*, 2023 SCC OnLine Del 6914.
4. *Bharatiya Nyaya Sanhita*, 2023, § 77, "Voyeurism"; *Information Technology Act*, no. 21 of 2000, § 66E.
5. *Case C-18/18, Glawischnig-Piesczek v. Facebook Ireland Ltd.*, ECLI:EU:C:2019:821, 2019. "Injunction extending to identical and equivalent content, subject to the use of automated techniques not requiring independent assessment."
6. *Case C-236/08, Google France SARL v. Louis Vuitton Malletier SA*, 2010 E.C.R. I-2417; *Case C-324/09, L'Oréal SA v. eBay International AG*, 2011 E.C.R. I-6011. "The exemption is available to a provider whose conduct is neutral and merely technical."
7. *Case C-360/10, Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA v. Netlog NV*, ECLI:EU:C:2012:85, 2012. "General filtering obligations incompatible with the prohibition upon general monitoring."
8. *Case C-401/19, Republic of Poland v. Parliament and Council*, ECLI:EU:C:2022:297, 2022. "Upholding article 17 subject to safeguards precluding the blocking of lawful content."
9. *Christian Louboutin SAS v. Nakul Bajaj*, 2018 SCC OnLine Del 12215. "Distinguishing active from passive intermediaries by reference to the degree of involvement in the transaction."
10. Coalition for Content Provenance and Authenticity. *C2PA Specification*. Version 2.x.
11. *Communications Decency Act*, § 230(c)(2), 2018. "Protection for good-faith restriction of material."
12. *Constitution of India*, art. 19(1)(a), art. 19(2).
13. *Copyright Act*, 1957, § 51(a)(ii), India. "Infringement by permitting a place to be used for communication of infringing works."
14. Keller, Daphne. *Empirical Evidence of Over-Removal by Internet Companies under Intermediary Liability Laws*. Stanford Center for Internet and Society, 2015 and updates. SSRN.
15. *Digital Personal Data Protection Act*, no. 22 of 2023, India.
16. *Directive 2000/31/EC (Electronic Commerce Directive)*, art. 15, 2000, *Official Journal of the European Union*, L 178, pp. 1. "No general obligation to monitor."
17. *Directive 2010/13/EU (Audiovisual Media Services Directive)*, as amended by *Directive 2018/1808*, art. 28b, 2018, *Official Journal of the European Union*, L 303, p. 69. "Obligations of Video-Sharing Platform Services."

18. *Directive 2019/790 on Copyright and Related Rights in the Digital Single Market*, art. 17, 2019, *Official Journal of the European Union*, L 130, pp. 92.
19. Engstrom, Evan and Nick Feamster. *The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools*. Engine, 2017. ResearchGate.
20. Frosio, Giancarlo. "Reforming intermediary liability in the platform economy." *Northwestern University Law Review Online*, vol. 112, 2017, pp. 19. SSRN Working Paper no. 3011276.
21. *Gonzalez v. Google LLC,* 598 U.S. 617 (2023); Twitter, Inc. v. Taamneh, 598 U.S. 471 (2023).
22. *Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, India*. "Due-diligence obligations, grievance redressal and additional duties for significant social media intermediaries."
23. *Information Technology Act, no. 21 of 2000, § 69A, India*. "Blocking of public access to information."
24. *Information Technology Act, no. 21 of 2000, § 79, India.* "Conditional exemption from liability for intermediaries observing due diligence and acting upon lawful orders."
25. Balkin, Jack M. "Free speech is a triangle." *Columbia Law Review*, vol. 118, 2018, pp. 2011*. SSRN Working* Paper no. 3186205.
26. *Joined Cases C-682/18 and C-683/18, YouTube and Cyando, ECLI:EU:C:2021:503, 2021*. "Conditions upon which a platform performs an act of communication to the public."
27. Quintais, João Pedro. "Generative AI, copyright and the AI act." *Computer Law & Security Review*, vol. 56, 2025, p. 106107. *SSRN Working Paper* no. 4912701.
28. K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1. "Proportionality as the governing standard."
29. Kent RO Sys. Ltd. v. *Amit Kotak*, 2017 SCC OnLine Del 7201. "An intermediary is not obliged to screen all content before publication."
30. Lenz v. Universal Music Corp., 815 F.3d 1145, 9th Cir., 2016. "A copyright owner must consider fair use before issuing a takedown notification."
31. *Manila Principles on Intermediary Liability, 2015; Santa Clara Principles on Transparency and Accountability in Content Moderation*, 2nd ed., 2021.
32. *Ministry of Electronics and Information Technology. Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules*, 2026. 2026.
33. Moody v. NetChoice, LLC, 603 U.S. 707, 2024. "Editorial curation by platforms as expressive activity."
34. Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], 1 September 2017, BGBl. I, pp. 3352, Germany.
35. Online Safety Act 2023, c. 50, *United Kingdom*. "Duties of care in respect of illegal and harmful content."
36. *Organisation for Economic Co-operation and Development. AI, Data Governance and Privacy: Synergies and Areas of International Co-operation. OECD Artificial Intelligence Papers*, no. 22, 2024.
37. *Organisation for Economic Co-operation and Development. Transparency Reporting on Terrorist and Violent Extremist Content Online*. *OECD Digital Economy Papers*, no. 313, 2021.
38. *Regulation (EU)* 2022/2065 *(Digital Services Act), 2022, Official Journal of the European Union*, L 277, pp. 1.
39. *Regulation (EU) 2024/1689 (Artificial Intelligence Act), art. 50, 2024, Official Journal of the European Union, L 2024/1689.* "Transparency obligations, including disclosure of artificially generated or manipulated content."
40. *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N*. Doc. A/HRC/38/35, 6 April 2018. "Content regulation in the digital age."
41. S. Rangarajan v. P. Jagjivan Ram, (1989) 2 SCC 574; *Anuradha Bhasin v. Union of India*, (2020) 3 SCC 637 (proportionality in restrictions upon expression and upon access to the internet).
42. Sabu Mathew George v. *Union of India*, (2018) 3 SCC 229 (auto-blocking directions and their limits).
43. Shreya Singhal v. *Union of India*, (2015) 5 SCC 1, paras. 117–119 (reading down § 79(3)(b) so that the obligation to act arises upon actual knowledge received through a court order or a government notification).
44. *Super Cassettes Indus*. Ltd. v. *My Space Inc*., 2016 SCC *OnLine Del* 6382. "Knowledge of specific infringing material required before liability attaches."
45. *Swami Ramdev v. Facebook, Inc*., 2019 SCC *OnLine Del* 10701. "Global disabling of content uploaded from within India."
46. *World Intellectual Property Organization*. "WIPO conversation on intellectual property and frontier technologies: Session papers on generative AI". 2024–2025.
47. *Zeran v. Am. Online, Inc., 129 F.3d 327, 4th Cir., 1997; Viacom Int'l, Inc. v. YouTube, Inc*., 676 F.3d 19, 2nd Cir., 2012. "Specific rather than general knowledge required to defeat the safe harbour."