CLRS International Conference on Law in the Digital Decade

Law in the Digital Decade

Rights, Regulation and Accountability

Edited by

Gyan Prakash Kesharwani · Ritu Verma

VidhiPragaz® Inking Your Brain

Centre For Legal Research And Studies

Law in the Digital Decade

Rights, Regulation and Accountability

Law in the Digital Decade

Rights, Regulation and Accountability

EDITED BY

Gyan Prakash Kesharwani

Chairperson and Publishing Editor at VidhiAagaz, India

Ritu Verma

B.A. LL.B., LL.M., Editorial Executive at VidhiAagaz, India

IN ASSOCIATION WITH

Centre For Legal Research And Studies

Centre for Legal Research and Studies

VidhiAagaz® Inking Your Brain

MAIHAR · INDIA

VIDHIAAGAZ
Maihar, Madhya Pradesh 485771, India
www.vidhiaagaz.com

Published in association with the Centre for Legal Research and Studies. This book collects papers from the CLRS International Conference on Law in the Digital Decade, held online on 13 September 2026.

© VidhiAagaz 2026
Copyright in each chapter is held as set out in the agreement executed for it.
First published 2026

This book is published open access under the Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC 4.0), https://creativecommons.org/licenses/by-nc/4.0/. It may be copied, shared and adapted for non-commercial purposes, provided the authors and the source are credited. Commercial reproduction requires the prior written permission of the publisher.

The views expressed in each chapter are those of its authors. They do not necessarily represent the views of the editors, the Centre for Legal Research and Studies, or the publisher, none of whom accepts responsibility for them.

ISBN 978-81-947620-6-5
DOI https://doi.org/10.63108/VAB.LDD.1

Typeset in Garamond by VidhiAagaz, Maihar
Printed and bound in India

Contents

Preface vii

Notes on Contributors ix

PART I THE AUTOMATED STATE AND THE CONSTITUTION

1 Accountability in AI and Automated Decision-Making: From Explainability to Reviewable Administrative Governance 3
Tzu-Yu Chiou

2 Governing by Algorithm: Constitutional Limits on Automated State Decisions in India 21
Dr. Mini S and Dr. Joby Bhasker

3 Constitutional Limits on Automated Governance: An Article 21 Perspective on Algorithmic Accountability in India 31
Yashwant SC

4 Judicial Disengagement and the Automated Welfare State 39
Nouria Rafi

5 The Right to Digital Access as an Emerging Facet of Article 21 49
Bharatha Vasan G and Hemasuruthi A.S.

PART II REGULATING ARTIFICIAL INTELLIGENCE: FAIRNESS, ACCOUNTABILITY AND LIABILITY

6 Governance by Design, Accountability by Default? India's Techno-Legal Turn in AI Regulation and the Problem of Contestability 61
Gautami Seth and Surabhi Kumari

7 Algorithmic Bias in Automated Decision-Making: An Empirical Fairness Audit and Comparative Legal Analysis Across Healthcare, Labour, Criminal Justice, and Finance 73
Asmita Srivastava and Asmita Singh

8 AI and Legal Accountability in India: Rethinking Responsibility for Automated Decisions 89
Nony Nema

9 Artificial Intelligence and Accountability of Automated Decisions 99
Gaurangi Mehrotra

10 Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems 111
Lakshmi Niveditha S and Anna Mariam Iype

11 Autonomous Weapon Systems in International Humanitarian Law: Command Responsibility and Accountability 125
Madhiroja P S and Vignesh Arun Prakash S R

PART III ARTIFICIAL INTELLIGENCE, COURTS AND THE DELIVERY OF JUSTICE

12 Artificial Intelligence in Judicial and Administrative Decision-Making: Reassessing Rights, Regulation, and Justice 137
Saudamini Gupta and Priyanshu

13 Constitutional Dimensions of AI-Driven Judicial Decision-Making 151
Sarika Baloda and Kirti Goyal

14 Two Black Boxes: Reassessing Judicial Deference in India's Emerging AI-Driven Security Apparatus 173
Mahek Motwani

15 Online Dispute Resolution and Technology in the Delivery of Justice 181
Dr. Karan Singh Gaur and Dr. Garima Yadav

16 Artificial Intelligence, Online Dispute Resolution and the Future of Justice Delivery: Opportunities, Risks and Regulatory Challenges 189
Dr. Yashasvi Sharma and Dr. Tejasvi Sharma

PART IV PRIVACY, DATA PROTECTION AND SPEECH ONLINE

17 The Data Protection Regime in India: Constitutional Privacy to Statutory Compliances 203
Enakshi Guha

18 From Privacy to Practice: Evaluating the Implementation of India's Digital Personal Data Protection Act 217
Dr. Maliya and Santosh Kumar

19 "You Can Reset a Password, But Not Your Biometrics": The Problem of Irreversible Biometric Data 227
Aakriti Jain and Shreya Agarwala

20 Right to be Forgotten vs. Right to Information: Finding the Digital Balance 239
Dr. Gargi Bhatt

21 Samvaad: Digital Constitutionalism, Article 19 and the Future of Online Dissent 253
Anoushka Girish and Apoorva T

22 Deepfakes, Defamation, and the Limits of Section 79: A Post-Visakha Analysis in Light of the IT Amendment Rules, 2026 267
Raiha Owais and Anha Tahir

Preface

The decade now under way has asked more of the law than most. Automated systems help decide who receives a ration, which post stays online, how a dispute is heard and what an institution learns about a person from data. Each of these decisions was once made, or at least signed, by someone who could be asked for reasons. Much of the work of law in this decade is to make sure that someone still can be.

This book is one of two volumes drawn from the CLRS International Conference on Law in the Digital Decade, which the Centre for Legal Research and Studies at VidhiAagaz held online on 13 September 2026. Its twenty-two chapters deal with rights, regulation and accountability; the companion volume, Law in the Digital Decade: Evidence, Intellectual Property and Markets, takes up evidence, intellectual property and the digital economy.

Part I asks how the Constitution reaches the automated state: decisions made by algorithms for the State tested against Articles 14 and 21, what reviewable administrative governance would require, the courts and the automated welfare state, and whether digital access is part of the right to life.

Part II turns to the regulation of artificial intelligence: India's techno-legal approach and the problem of contestability, an empirical audit of algorithmic bias across four sectors, the allocation of responsibility and liability for automated decisions, and command responsibility for autonomous weapons in international humanitarian law.

Part III considers courts and the delivery of justice: artificial intelligence in judicial and administrative decision-making, the constitutional limits on its use in adjudication, the deference courts give to AI-driven security measures, and online dispute resolution.

Part IV deals with privacy, data protection and speech online: the Digital Personal Data Protection Act, 2023, compromised biometric data, the right to be forgotten against the right to information, online dissent under Article 19, and deepfakes under section 79 of the Information Technology Act.

Every chapter has been read line by line. Where an error of fact, citation or expression could be checked against its source, we have corrected it, and we have tried to leave every argument as its author made it. The views expressed are the authors' own.

We thank the authors, the chairs and co-chairs who led the conference sessions, and the editorial team at VidhiAagaz who prepared the manuscripts for the press. We hope the book will be useful to students, researchers, practitioners and policymakers.

Gyan Prakash Kesharwani
Chairperson and Publishing Editor at VidhiAagaz, India

Ritu Verma
B.A. LL.B., LL.M., Editorial Executive at VidhiAagaz, India

EDITORS

Notes on Contributors

1. TZU-YU CHIOU, Assistant Professor at Department of Law, Hsuan Chuang University, Hsinchu, Taiwan.

2. DR. MINI S, Associate Professor at Government Law College, Ernakulam, Kerala, India.
   DR. JOBY BHASKER, Assistant Professor at Government Law College, Ernakulam, Kerala, India.

3. YASHWANT SC, Student at SASTRA Deemed University, Thanjavur, Tamil Nadu, India.

4. NOURIA RAFI, Independent Researcher at Aligarh Muslim University, Aligarh, Uttar Pradesh, India.

5. BHARATHA VASAN G, Student at The Tamil Nadu Dr. Ambedkar Law University, School of Excellence in Law, Chennai, Tamil Nadu, India.
   HEMASURUTHI A.S., Student at The Tamil Nadu Dr. Ambedkar Law University, School of Excellence in Law, Chennai, Tamil Nadu, India.

6. GAUTAMI SETH, Manager, New Initiatives & Research at India Legal Research Foundation, Noida, Uttar Pradesh, India.
   SURABHI KUMARI, Legal Research and Content Associate at India Legal Research Foundation, Noida, Uttar Pradesh, India.

7. ASMITA SRIVASTAVA, Student at Bennett University, Greater Noida, Uttar Pradesh, India.
   ASMITA SINGH, Student at Bennett University, Greater Noida, Uttar Pradesh, India.

8. NONY NEMA, PhD Scholar at National Law Institute University, Bhopal, Madhya Pradesh, India.

9. GAURANGI MEHROTRA, Advocate.

10. LAKSHMI NIVEDITHA S, Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.
    ANNA MARIAM IYPE, Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.

11. MADHIROJA P S, LL.M. Student at Government Law College, Madurai, Tamil Nadu, India.
    VIGNESH ARUN PRAKASH S R, LL.M. Student at Government Law College, Madurai, Tamil Nadu, India.

12. SAUDAMINI GUPTA, Junior Standing Counsel at Income Tax Department.
    PRIYANSHU, Junior Research Fellow at National Law University Odisha, Cuttack, Odisha, India.

13. SARIKA BALODA, Research Scholar at Symbiosis International (Deemed University), Pune, Maharashtra, India.
    KIRTI GOYAL, Research Scholar at Amity University Rajasthan, Jaipur, Rajasthan, India.

14. MAHEK MOTWANI, Student at SNDT Women's University, Juhu, Mumbai, Maharashtra, India.

15. DR. KARAN SINGH GAUR, Assistant Professor at Anangpuria Group of Institutions, Faridabad, Haryana, India.
    DR. GARIMA YADAV, Assistant Professor at Anangpuria Law School, Faridabad, Haryana, India.

16. DR. YASHASVI SHARMA, Assistant Professor at Rajdhani Law College, Jaipur, Rajasthan, India.
    DR. TEJASVI SHARMA, Assistant Professor at Kanoria School of Law for Women, Jaipur, Rajasthan, India.

17. ENAKSHI GUHA, Practicing Advocate.

18. DR. MALIYA, Assistant Professor at School of Law, NIILM University, Kaithal, Haryana, India.
SANTOSH KUMAR, Research Scholar at School of Law, NIILM University, Kaithal, Haryana, India.

19. AAKRITI JAIN, Student at Unitedworld School of Law, Karnavati University, Gandhinagar, Gujarat, India.
SHREYA AGARWALA, Student at Unitedworld School of Law, Karnavati University, Gandhinagar, Gujarat, India.

20. DR. GARGI BHATT, Assistant Professor & Head, Training & Placement Office at DME Law School, Delhi Metropolitan Education, Noida, Uttar Pradesh, India.

21. ANOUSHKA GIRISH, Student at CMR University, School of Legal Studies, Bengaluru, Karnataka, India.
APOORVA T, Assistant Professor at CMR University, School of Legal Studies, Bengaluru, Karnataka, India.

22. RAIHA OWAIS, Faculty of Law, Aligarh Muslim University, Aligarh, Uttar Pradesh, India.
ANHA TAHIR, Faculty of Law, Aligarh Muslim University, Aligarh, Uttar Pradesh, India.

PART I

The Automated State and the Constitution

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.1

CHAPTER 1

Accountability in AI and Automated Decision-Making: From Explainability to Reviewable Administrative Governance

TZU-YU CHIOU*

This paper argues that the prevailing discourse on algorithmic accountability, which centers on technical explainability, offers an insufficient normative foundation for governing AI-driven automated decision-making (ADM) within public administration. Drawing on Lorenz von Stein's dialectical theory of state and society (Staat-Gesellschaft-Dialektik) and his conception of administration (Verwaltung) as the institutional mediator between state will and social interests, this paper reframes algorithmic governance as a contemporary iteration of Stein's core concern: preventing the capture of state functions by dominant social forces. Where Stein warned that unchecked class interests could colonize administrative power, this paper contends that concentrated ownership of data, models, and computational infrastructure by technology capital poses a structurally analogous risk to public administration today. Explainability-focused accountability regimes, by reducing legitimacy to procedural transparency of decision logic, mirror a narrowly technocratic conception of administration that Stein explicitly rejected. Instead, this paper proposes that administrative legitimacy in the age of AI-ADM requires institutionalized, continuous mechanisms of review—echoing Stein's insistence on the reciprocal feedback between administration and society (Selbstverwaltung)—that allow public institutions to intervene when automated governance drifts from the public interest. Using Taiwan's emerging AI governance framework as a case study, the paper examines how administrative law instruments (judicial review, procedural safeguards, oversight bodies) might be reconstructed to operationalize this Stein-inspired model of reviewable governance. The paper concludes that accountability for automated decision-making should be evaluated not merely by the interpretability of algorithmic outputs, but by the robustness of institutional channels through which state and society can jointly contest and correct administrative power exercised through AI systems.

Keywords: algorithmic accountability; automated decision-making; Lorenz von Stein; administrative law; AI governance; Taiwan

1 INTRODUCTION

The governance of artificial intelligence in public administration has, over the past decade, converged on a single dominant vocabulary: explainability. From the European Union's General Data Protection Regulation and its much-debated "right to explanation," to the OECD's AI Principles, to the technical standards issued by national metrology institutes, the policy consensus holds that an automated decision becomes accountable to the extent that the reasoning behind it can be rendered intelligible to the person it affects. This paper contends that this consensus, however well-intentioned, rests on a category error. It treats accountability as a property of outputs—the legibility of a model's internal logic—rather than as a property of institutions: the durable channels through which a polity can contest, correct, and if necessary reverse the exercise of public power. Explainability asks whether a decision can be understood.

* Assistant Professor at Department of Law, Hsuan Chuang University, Hsinchu, Taiwan.

Accountability, in the older and more demanding sense the term carries in administrative law, asks whether a decision can be *undone*.

This distinction is not merely semantic. It reflects two different theories of where the danger of automated decision-making actually lies. The explainability paradigm locates the danger in opacity: the automated system is a black box, and the remedy is to open it. The reviewability paradigm advanced in this paper locates the danger elsewhere—in the concentration of the material and epistemic infrastructure of governance (data, compute, proprietary models) in private hands whose interests do not necessarily align with the public interest the administration is charged with executing. On this view, even a perfectly explainable algorithm remains a threat to administrative legitimacy if the institutional mechanisms for contesting and correcting its outputs have atrophied or were never built. Conversely, an opaque model embedded within a robust structure of continuous review, judicial oversight, and participatory correction may pose less of a legitimacy problem than a transparent one operating outside any such structure.

To develop this argument, the paper turns to a theorist largely absent from the algorithmic accountability literature: the nineteenth-century German jurist and social theorist Lorenz von Stein (1815–1890). Stein is remembered, where he is remembered at all in Anglophone scholarship, as a founder of comparative public administration and as an intellectual forerunner of the modern *Rechtsstaat*'s administrative apparatus. His more distinctive contribution, however, was a dialectical theory of the relationship between state (*Staat*) and society (*Gesellschaft*)—one in which administration (*Verwaltung*) occupies the structurally decisive position of mediating between the universal will of the state and the particular, often self-interested, claims of social classes. Stein's abiding worry, developed across his multi-volume *Verwaltungslehre*[1] and his earlier writings on socialism and the social question,[2] was that administration could be captured: that a dominant social class, through its command of economic resources, could hollow out the state's claim to represent the common good and redirect administrative power toward its own particular ends. Administration remains free, in Stein's framework, only for as long as it retains institutional mechanisms that allow it to sense and respond to the totality of social interests rather than the loudest or best-resourced among them.

This paper's central claim is that Stein's diagnosis, developed to explain the class conflicts of industrializing Europe, describes with uncanny precision the structural position that large technology firms now occupy vis-à-vis the administrative state in the age of AI. The firms that supply foundation models, cloud compute, and the data pipelines on which AI-ADM systems depend are not merely vendors; they are, in Stein's terms, a rising social power whose command of a scarce and decisive resource—computational and epistemic infrastructure—gives them the capacity to shape administrative outcomes in ways that elude ordinary channels of political and legal accountability. An accountability framework built solely around explainability treats this structural relationship as a technical footnote, a matter of interface design, rather than what it is: a question of who controls the means of administrative cognition, and whether the state has retained institutional capacity to correct that control when it drifts from the public interest.

This argument situates the paper within a broader research program concerned with recovering classical and nineteenth-century theories of administration—Stein's account of *Verwaltung* prominent among them—as resources for evaluating the legitimacy of contemporary regulatory and administrative

[1] LORENZ VON STEIN, DIE VERWALTUNGSLEHRE (8 vols., 1865–1868).

[2] LORENZ VON STEIN, GESCHICHTE DER SOZIALEN BEWEGUNG IN FRANKREICH VON 1789 BIS AUF UNSERE TAGE (1850).

arrangements that a purely proceduralist or technocratic vocabulary struggles to assess.[3] This recovery project runs parallel to, and draws methodological encouragement from, recent efforts to retrieve classical legal-theoretical resources for the assessment of contemporary public law more generally.[4] Where much of the algorithmic accountability literature treats the problem of AI-ADM as historically unprecedented, requiring correspondingly novel technical and legal instruments, this paper's method is deliberately genealogical: it treats the AI-ADM accountability problem as a contemporary instance of a much older and better-understood problem in the theory of administration, namely the risk that an ostensibly neutral administrative apparatus will be captured by whichever social force controls a scarce and decisive resource, and it looks to the administrative law traditions built to manage that older problem for guidance on how the newer one might be governed. This is not to deny that AI-ADM raises genuinely novel technical questions; it is to insist that the normative question of accountability—who may contest an exercise of public power, and through what institutional channel—is not one of them.

The paper proceeds in seven further parts. Part 2 surveys the explainability-centered discourse on algorithmic accountability and identifies its normative limitations, particularly its tendency to conflate transparency with legitimacy and its neglect of the institutional question of who may act on what is disclosed. Part 3 reconstructs Stein's dialectical theory of state, society, and administration, with particular attention to his concepts of *Selbstverwaltung* (self-administration) and the administrative state's vulnerability to class capture. Part 4 develops the paper's central analogy, arguing that concentrated technological capital reproduces the structural threat Stein associated with concentrated economic capital, and that explainability-based accountability regimes reproduce the narrowly technocratic conception of administration Stein rejected. Part 5 elaborates the alternative this paper proposes: a reviewability principle grounded in continuous, institutionalized feedback between administration and society, modeled on Stein's account of self-administration. Part 6 turns to Taiwan as a case study, examining the country's emerging AI governance architecture—including the newly enacted Artificial Intelligence Basic Act, the National Science and Technology Council's AI governance framework, and existing administrative law instruments—to assess how far current arrangements approximate, and how they might be reconstructed to better approximate, the reviewability model. Part 7 considers objections and limitations. Part 8 concludes.

2 THE EXPLAINABILITY PARADIGM AND ITS DISCONTENTS

The turn to explainability as the organizing principle of algorithmic accountability has a coherent intellectual history. Early critiques of automated decision-making in the 2010s focused on the "black box" problem:[5] machine learning models, particularly deep neural networks, produce outputs through computational processes too complex for a human to trace step by step, even when the underlying code is fully disclosed. Commentators observed that this opacity was qualitatively different from the opacity of earlier bureaucratic decision-making, where a human official could always, in principle, be asked to state reasons. The proposed remedy followed naturally from the diagnosis: if the problem is that models cannot explain

[3] *See generally* LORENZ VON STEIN, DIE VERWALTUNGSLEHRE (8 vols., 1865–1868).

[4] ADRIAN VERMEULE, COMMON GOOD CONSTITUTIONALISM: RECOVERING THE CLASSICAL LEGAL TRADITION (2022).

[5] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015).

themselves, the solution is to build models that can, or to build supplementary tools that approximate an explanation after the fact.

This diagnosis produced an enormous and genuinely valuable technical research program—counterfactual explanations, feature-attribution methods, model distillation into interpretable surrogates, and so on—alongside a parallel legal literature debating whether instruments such as the GDPR's Article 22 and Recital 71 in fact create an enforceable right to explanation, and if so, what kind of explanation would satisfy it.[6] Administrative law scholars extended the inquiry to public-sector ADM specifically, asking whether existing doctrines of reasoned decision-making, developed for human bureaucrats, could or should be adapted to algorithmic ones.

Yet even sympathetic observers have noted the limits of this approach. Three limitations bear directly on the argument of this paper.

First, explainability is frequently unavailable or, when available, unhelpful, in exactly the cases where accountability matters most. High-performing models are often the least amenable to faithful, human-legible explanation, so that jurisdictions face a persistent trade-off between the accuracy of an automated decision and the honesty of the explanation offered for it. Where this trade-off is resolved in favor of interpretable-but-weaker models, one accountability problem is exchanged for a different one: administrations now systematically deploy less accurate tools in order to satisfy an explainability requirement whose connection to the public interest the requirement was meant to serve is, at best, indirect.

Second, explainability, even where technically achievable, does not by itself confer any capacity to act. A citizen who receives a lucid account of why an algorithm denied their benefits application, revoked their license, or flagged their tax return for audit is no better positioned to have that decision reversed unless some further institutional mechanism exists through which the explanation can be contested. Explainability is, in this sense, a necessary but radically insufficient condition for accountability in the administrative-law sense of the term[7]—a sense that has always centered not on whether reasons were given, but on whether an independent body could review those reasons and substitute its own judgment, or compel the original decision-maker to reconsider. The explainability literature has tended to treat this second, institutional question as separate from, and often secondary to, the technical question of interpretability. This paper treats the ordering as reversed: institutional reviewability is the primary accountability variable, and explainability is, at most, an input that makes reviewability more effective once the reviewing institution exists.

Third, and most importantly for this paper's argument, the explainability paradigm has a specific and underappreciated ideological valence. It locates the site of accountability inside the algorithm itself—in its architecture, its features, its decision boundary—and correspondingly locates the remedy inside the technical apparatus that produced the algorithm: better documentation, better interface design, better disclosure. This has the effect, whether intended or not, of depoliticizing the governance question. It invites regulators to ask "can this system explain itself adequately?" rather than "who controls the infrastructure that makes this system possible, and does the public retain any leverage over them?" The former question can be answered, and often is answered, by the very firms that build and sell the systems in

[6] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 76–99 (2017).

[7] Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1–33 (2014); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249–313 (2008).

question, through voluntary disclosure frameworks, model cards, and algorithmic impact assessments that they themselves author.[8] The latter question cannot be answered by the firms at all; it can only be answered by public institutions willing to examine the structural position those firms occupy relative to the administrative state.

A further, closely related difficulty deserves separate mention. Much of the explainability literature implicitly assumes a bilateral relationship between the affected citizen and the administration that deploys the algorithm, on the model of the classical due process encounter between an individual and a bureaucrat. This assumption obscures the fact that the party best positioned to explain an AI-ADM system's behavior is frequently neither the citizen nor the administration, but the private vendor that designed, trained, and often continues to host and update the model. Administrations procuring AI systems under standard commercial licensing arrangements not infrequently lack full visibility into training data provenance, model weights, or the precise mechanism by which a given input produces a given output, particularly where the vendor treats these as trade secrets. In such arrangements, an explainability mandate addressed to "the algorithm" in the abstract risks becoming, in practice, a mandate addressed to whatever the vendor is willing to disclose, filtered through the vendor's own commercial incentives to protect proprietary methods and limit liability exposure. This triangulation—citizen, administration, vendor—is largely invisible in accountability frameworks that speak only of "the algorithm" and "the affected person," and it is precisely this triangulation that the Steinian framework developed below is designed to bring into view.

Comparative administrative law scholarship offers a further, complementary observation. Jurisdictions with strong traditions of reasoned administrative decision-making—Germany's *Begründungspflicht*, France's *motivation des actes administratifs*, Taiwan's own statutory requirement under Article 96 of the Administrative Procedure Act that administrative dispositions state their factual and legal basis—already possessed, prior to the emergence of AI-ADM, doctrinal resources considerably more demanding than anything the explainability literature has proposed as a novel remedy. The puzzle this raises, and one this paper returns to in Part 6, is why AI governance discourse has tended to treat explainability as a novel technical challenge requiring novel technical solutions, rather than as an instance of a much older administrative law problem—the adequacy of stated reasons—for which many legal systems already possess doctrine, albeit doctrine developed with human decision-makers in mind and in need of extension rather than wholesale replacement.

It is this third limitation that motivates the turn to Stein. Stein wrote at a moment when a strikingly analogous depoliticizing move was available and was, in fact, being made by contemporaries who treated the "social question" of industrial capitalism as a matter of amelioration—better wages, better working conditions, better relief for the poor—rather than as a structural question about who controlled the means of production and what that control implied for the state's capacity to govern in the name of the whole of society. Stein's intervention was to insist that amelioration without structural attention to control was self-defeating: that an administration which improved the treatment of subordinate classes without addressing the structural capacity of a dominant class to redirect state power toward its own ends had not solved the problem of legitimacy, but merely postponed its recognition. The parallel to contemporary AI governance, this paper argues, is exact. Improving the explanations automated systems provide

[8] Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1–56 (2019).

without addressing the structural capacity of the firms that build those systems to shape administrative outcomes is amelioration without structural correction.

3 STEIN'S DIALECTICAL THEORY OF STATE AND SOCIETY

Lorenz von Stein's political and administrative theory is best understood as an extended argument about the conditions under which a state can be said to represent something more than the interests of whichever social group currently controls it. Stein's starting point, developed most fully in his early study of French socialist and communist movements and elaborated across the volumes of his later *Verwaltungslehre*, is a distinction between state (*Staat*) and society (*Gesellschaft*). Society, for Stein, is the sphere of particular interests: the realm in which individuals and classes pursue their own material advantage, structured by the prevailing distribution of property and labor. The state, by contrast, is or ought to be the sphere of the universal will—the institutional expression of the interest of the whole community as such, standing above and independent of any particular class interest within society.

The difficulty, which Stein treats as the central problem of modern politics, is that the state does not exist in a vacuum above society; it is staffed, financed, and to a significant degree captured by particular social actors, and above all by whichever class currently holds the dominant position within the prevailing economic order. Stein's analysis of nineteenth-century industrial society identified the emerging propertied bourgeoisie as the class most capable of translating its economic dominance into political dominance, precisely because economic power confers disproportionate access to the instruments of governance: to legislatures, to the press, and—crucially for this paper's purposes—to the administrative apparatus itself. Where a dominant class successfully colonizes the state, Stein argues, the state ceases to perform its distinctive function. It no longer stands above particular interests; it becomes an instrument of one particular interest against the rest of society, while continuing to claim the legitimacy that only representation of the universal interest can confer. Stein terms this condition, in various formulations across his work, a degeneration of the state into an instrument of class rule—a diagnosis that anticipates, and indeed directly influenced, aspects of Marx's own analysis of the state, even as Stein's proposed remedy diverges sharply from Marx's.

Stein's remedy is not revolution but administration. He assigns to *Verwaltung*—the administrative apparatus of the state, as distinct from both the legislature and the monarch or executive head—the decisive role of preventing class capture, a conception later elaborated within the German administrative law tradition itself.[9] Administration, on Stein's account, is the institutional site at which the abstract universal will of the state is translated into concrete measures affecting particular persons and groups, and it is precisely because administration operates at this point of translation that it has the structural capacity either to correct for class bias or to entrench it. A well-constituted administration, continuously attentive to the actual condition of all social classes rather than only the dominant one, can act as a countervailing force: extending education, regulating labor conditions, redistributing opportunity, and in general using the instruments of public power to prevent any one class's particular interest from permanently capturing the state's claim to represent the whole. A poorly constituted administration—one that has itself been captured, whether through personnel, financing, or dependence on information supplied by the dominant

[9] ERNST FORSTHOFF, LEHRBUCH DES VERWALTUNGSRECHTS (10th ed. 1973).

class—ceases to perform this corrective function and instead becomes an additional instrument of capture.

The concept that operationalizes this corrective function in Stein's mature work is *Selbstverwaltung*, self-administration. Stein's use of this term is broader than its later, narrower legal sense of local-government autonomy. For Stein, self-administration names the structural condition under which administration remains continuously responsive to the actual, changing configuration of social interests, rather than freezing at a moment of maximal influence for whichever class happened to be dominant when a given administrative arrangement was established. Self-administration requires institutional mechanisms of feedback: channels through which changes in social condition can register within the administrative apparatus and prompt revision of administrative practice, before those changes harden into the kind of grievance that Stein associated with revolutionary rupture. Administration, in other words, must be reviewable and revisable by the society it governs—not in the sense that society directly commands administration on a case-by-case basis, which would simply invert the capture Stein feared, but in the sense that durable institutional mechanisms exist through which administrative drift away from the universal interest can be detected and corrected before it consolidates into permanent class rule.

Stein's account of *Selbstverwaltung* is best approached through the distinction he draws between three moments of state activity: the constitutional moment, in which the fundamental form of the state and the relationship among its organs is fixed; the legislative moment, in which the state's universal will is articulated in general rules; and the administrative moment, in which those general rules are applied to the particular, constantly changing circumstances of individual life. Stein's insight, distinctive among nineteenth-century state theorists, is that only the third of these moments is structurally capable of tracking the ongoing transformation of social conditions in real time. Constitutions are amended rarely and legislation revised episodically, but administration operates continuously, encountering daily the concrete situations—a family's poverty, a worker's injury, a merchant's failed enterprise—in which the abstract promise of the universal interest either is or is not made good. It follows, for Stein, that administration is not merely the executive tail of the legislative dog, mechanically implementing rules fixed elsewhere, but the organ through which the state's claim to represent society as a whole is continuously tested against social reality. This is why Stein devotes such disproportionate attention, relative to his contemporaries among state theorists, to the mundane machinery of administration—poor relief, factory inspection, public health, education administration—rather than to constitutional design narrowly conceived: it is in this machinery, not in the constitutional text, that the fate of the state's claim to universality is actually decided.

This account carries a further implication that bears directly on the present paper's argument. Because administration is, for Stein, the site at which the universal interest is continuously tested against social reality, the health of administration cannot be assessed by inspecting its formal legal authorization alone. A ministry may be entirely lawful in its formal constitution, entirely faithful to the letter of the legislation it implements, and still have degenerated, in Stein's sense, into an instrument of a particular class, if the informational and material channels through which it perceives and responds to social reality have themselves been captured. This is precisely the diagnostic move this paper transposes to AI-ADM: a public administration may comply fully with data protection law, procurement law, and whatever AI-specific disclosure statute is in force, and its use of automated decision-making may still have degenerated, in the relevant sense, if the informational and material channels through which the underlying AI system perceives the social field—training data, feature selection, model architecture, update cadence—have been

captured by a vendor whose commercial interests do not track the public interest the administration is charged with serving. Formal legality, on Stein's account as on this paper's extension of it, is a necessary but insufficient condition of administrative legitimacy; the sufficient condition additionally requires that the channels of perception themselves remain open to correction from outside the interest that currently dominates them.

Three features of this framework are worth isolating for the argument that follows. First, Stein's analysis is structural rather than moralistic: the danger he identifies does not depend on any individual administrator or capitalist acting in bad faith. Capture can occur, and typically does occur, through the ordinary operation of institutional dependency—an administration that relies on a dominant class for information, personnel, or resources will tend to reproduce that class's perspective even where every individual official involved believes themselves to be acting impartially. Second, the remedy Stein proposes is institutional rather than technical: it does not consist in better disclosure of what administration is doing, but in durable structures that allow administration's outputs to be contested and revised. Third, and most distinctively, Stein locates the danger of capture not in the formal separation of powers—the classical liberal concern that the executive might usurp legislative or judicial authority—but in the material dependency of administration on the resources controlled by a particular social class, however constitutionally proper the formal channels of administrative action might appear. It is this third feature, in particular, that gives Stein's framework its purchase on the AI governance question, since the risk of capture posed by concentrated technological infrastructure operates precisely through material and epistemic dependency rather than through any formal usurpation of administrative authority.

4 FROM CLASS CAPTURE TO DATA CAPTURE: REFRAMING ALGORITHMIC GOVERNANCE

The central move of this paper is to argue that the structural position Stein assigned to the capitalist class of industrializing Europe is today occupied, with respect to public administration's use of artificial intelligence, by the small number of firms that control the foundation models, cloud computing infrastructure, and large-scale data pipelines on which most AI-ADM systems depend. This is not a claim that technology firms are malicious in the way Stein sometimes suggested industrial capital could be; it is, in keeping with Stein's own structural method, a claim about institutional dependency that holds regardless of the intentions of any particular actor.

Consider the ordinary lifecycle of an AI-ADM system deployed within public administration—for instance, a system used to triage tax audits, allocate social welfare inspections, or support permitting decisions. Few public administrations possess in-house the computational infrastructure, the proprietary training data, or the specialized technical personnel required to build such systems from first principles. The overwhelming tendency, documented across jurisdictions with widely varying administrative traditions, is for public administration to procure these capabilities from a small number of firms capable of supplying them at scale: firms that operate the major cloud platforms, that have developed and trained the leading foundation models, and that increasingly bundle the technical, legal, and even ethical-compliance infrastructure (model documentation, bias audits, explainability tooling) that accompanies deployment.[10] Administration, on this arrangement, becomes structurally dependent on a small set of private actors not merely for a discrete input, as it might depend on a stationery supplier, but for the very capacity

[10] David S. Rubenstein, *Acquiring Ethical AI*, 73 FLA. L. REV. 747–819 (2021).

to exercise a core function of contemporary governance: the sorting, prioritizing, and adjudicating of claims upon the state.

This dependency reproduces, with striking fidelity, the structure Stein worried about. It is a dependency of *capacity*, not merely of *resources* in the ordinary sense: administration does not simply purchase a tool from these firms, the way a nineteenth-century ministry might purchase paper or ink; it purchases a cognitive apparatus—the very means by which it perceives and processes the claims of the population it governs. Where Stein worried that a dominant economic class could shape the administrative apparatus because that class controlled the material conditions on which the state's ordinary operation depended (financing, information networks, and eventually direct administrative personnel drawn from its own ranks), the contemporary technology sector occupies an analogous position because it controls the computational and epistemic conditions on which the state's algorithmic operation depends. The parallel extends to Stein's account of how capture typically proceeds: not through crude bribery or overt political capture, but through the ordinary, seemingly neutral operation of technical dependency, in which the vendor's design choices, training data selections, and default configurations come to shape administrative outcomes in ways that are difficult for the administration itself, let alone the public, fully to perceive or interrogate.

Against this backdrop, the limitations of the explainability paradigm identified in Part 2 take on a sharper political meaning. An accountability framework that asks only whether the algorithm can explain its own outputs leaves entirely untouched the antecedent question of who designed the algorithm's objective function, who selected and curated its training data, and who retains the practical capacity to modify the system once deployed. It is entirely possible—indeed, on current market structures, it is the ordinary case—for a system to satisfy every prevailing explainability standard while the actual power to alter that system's behavior, to update its weights, or to withdraw it from service rests with a private firm whose contractual relationship with the public administration was negotiated with limited public administrative bargaining power and even less public participation. Explainability, on this reading, functions analogously to the "amelioration without structural correction" that Stein criticized in his contemporaries' response to the social question: it treats the symptom (the citizen's inability to understand a given decision) while leaving the underlying structure of dependency (the administration's inability to control the conditions of its own decision-making capacity) entirely intact.

This structural dependency is compounded by an epistemic dimension that has no close analogue in Stein's own account and that, if anything, sharpens the concern. Nineteenth-century capital, however dominant, did not control the categories through which administration perceived the social field it governed; ministries retained their own statistical bureaus, their own inspectors, their own independently gathered census and survey data, even where they depended on capital for financing and even where dominant classes contested the interpretation of that data. Contemporary AI-ADM systems, by contrast, frequently supply not only the mechanism of decision but also the very features and categories through which an administrative problem is represented: which variables are treated as predictive of tax fraud, which patterns are treated as indicative of benefit ineligibility, which combinations of factors are treated as elevating a permit application's risk profile. Where these representational choices are embedded in a proprietary model whose training data and feature engineering the administration cannot independently reconstruct, the dependency Stein worried about extends from the material means of governance to its cognitive preconditions: administration comes to see the social field it governs through categories it did

not choose and often cannot fully inspect. This is, in effect, a deeper form of the capture Stein described, since it operates not merely on the resources available to implement a policy, but on the antecedent perception of what the policy problem consists in.

It is important to be precise about the scope of this claim. The argument is not that technology firms occupy a position identical in every respect to the industrial bourgeoisie Stein analyzed, nor that the remedy must therefore be identical. Stein's own framework was explicitly historical and comparative; he traced how the specific form taken by class capture, and the specific administrative remedies available to counteract it, varied across the different national and constitutional contexts he studied in England, France, and the German states. The claim advanced here is at the level of structure, not of surface detail: that wherever a scarce and decisive input to administrative capacity becomes concentrated in the hands of a narrow set of private actors, the risk Stein identified—administration ceasing to represent the universal interest and instead reproducing the particular interest of the class that controls the decisive input—recurs, and that the appropriate remedy is accordingly structural rather than technical. This reframing has direct implications for how accountability ought to be defined and pursued, developed in the following part.

5 SELBSTVERWALTUNG AND THE REVIEWABILITY PRINCIPLE

If the risk posed by AI-ADM is best understood, following Stein, as a risk of structural capture rather than a risk of insufficient technical transparency, the appropriate remedy follows Stein's own prescription: not better disclosure, but durable institutional mechanisms of self-administration—continuous, reciprocal feedback between administration and the society it governs, capable of detecting and correcting drift away from the universal interest before that drift consolidates into permanent capture. This paper terms the administrative-law operationalization of this idea the *reviewability principle*.

The reviewability principle can be stated in general terms as follows: an AI-ADM system deployed within public administration is accountable not to the extent that its outputs can be explained, but to the extent that its outputs, its training and deployment conditions, and the contractual and institutional relationships underlying its procurement remain subject to continuous, independent, and effective review by institutions capable of compelling correction. This formulation deliberately shifts the locus of accountability from the algorithm to the surrounding institutional architecture, in keeping with Stein's insistence that the health of administration is a function of its institutional relationship to society, not a function of the internal transparency of any single administrative act.

Four institutional components follow from this reformulation, each with a rough correlate in Stein's own account of self-administration and each translatable, as Part 6 will argue, into concrete instruments of administrative law.

The first component is *contestability at the point of decision*: affected persons must retain a meaningful, accessible avenue to challenge an individual automated decision, not merely to receive an explanation of it. This corresponds to Stein's insistence that self-administration requires channels through which particular grievances can register within the administrative apparatus. In administrative-law terms, this maps onto rights of appeal, rights to a human re-determination, and rights to have an automated decision suspended pending review—instruments considerably more demanding than a bare right to explanation,

since they entail an obligation on the part of administration to actually reconsider, not merely to account for, the decision reached.

The second component is *structural auditability independent of the vendor*: administration, or an independent body acting on its behalf, must retain the practical and legal capacity to audit an AI-ADM system's training data, decision logic, and performance across affected populations, without that capacity being mediated or gatekept by the private firm that supplied the system. This corresponds most directly to Stein's structural diagnosis: since the danger identified is one of dependency, the remedy must include institutional capacities that do not themselves depend on the cooperation of the party whose power is being checked. In practice, this requires procurement law to mandate audit rights, data portability, and, where feasible, in-house or third-party technical capacity within administration or its oversight bodies, sufficient to conduct meaningful review without relying exclusively on documentation the vendor itself authors.

The third component is *institutionalized, periodic re-evaluation independent of any individual complaint*: because Stein's concern is with structural drift that may not register as an individually cognizable grievance until it has already consolidated, self-administration requires mechanisms of review that operate on a scheduled or triggered basis, independent of whether any particular affected person has filed a challenge. This corresponds to instruments such as mandatory algorithmic impact assessments conducted at defined intervals, sunset clauses requiring active re-authorization of deployed systems, and standing oversight bodies charged with monitoring aggregate outcomes (disparate impact across demographic or regional lines, drift in system performance over time, and so on) rather than only individual disputes.

The fourth component is *judicial or quasi-judicial capacity to compel correction, not merely to require disclosure*: the reviewing institutions contemplated above must possess remedial powers adequate to Stein's demand that administration remain genuinely revisable, not merely observable. A court or oversight body that can require an administration to explain a system's operation, but cannot order the suspension, modification, or withdrawal of that system when it is found to depart from the public interest, replicates rather than solves the depoliticized, disclosure-centered model this paper criticizes. Genuine reviewability requires that the ultimate finding of a reviewing body carry binding, corrective force.

A brief illustration may make the interaction of these four components more concrete. Suppose a municipal administration deploys a machine-learning system to prioritize building-safety inspections, trained on historical inspection and violation data supplied by the vendor together with the system. Under a purely explainability-centered regime, accountability is satisfied once the vendor publishes a model card describing the features used (building age, prior violations, neighborhood density) and the administration provides, on request, a feature-attribution summary explaining why a given building was flagged. Under the reviewability principle, this disclosure would be a starting point rather than an endpoint. Contestability at the point of decision would require that a building owner who disputes an inspection priority ranking be entitled to request human reconsideration that genuinely revisits the ranking, not merely restates the algorithmic rationale. Structural auditability would require that the municipal administration, or an independent auditor acting on its behalf, be able to examine whether the historical violation data used to train the system itself reflects prior discriminatory enforcement patterns—a question the vendor's own model card, however detailed, has no particular incentive to surface, since it implicates the quality of the vendor's own training data rather than merely the technical fidelity of its model. Periodic independent re-evaluation would require that aggregate inspection outcomes be reviewed on a fixed schedule for disparate impact

across neighborhoods, independent of whether any individual building owner has filed a complaint, since the harm of a systematically skewed inspection regime may never register as a cognizable individual grievance even as it redistributes the burdens and protections of building-safety regulation in ways that track, rather than correct, prior patterns of neglect. Binding corrective remedy, finally, would require that findings from either the individual reconsideration process or the periodic structural review be capable of triggering an enforceable requirement that the system be retrained, reweighted, or withdrawn, rather than merely a public report noting the disparity for future policy consideration. Each component addresses a distinct failure mode that the others leave untouched, which is why the reviewability principle is specified as a conjunction of four elements rather than a single master criterion: a regime strong on contestability but weak on structural auditability will catch individually salient errors while leaving systemic bias in training data undetected; a regime strong on periodic review but weak on binding remedy will detect systemic bias without being able to correct it; and so on.

It bears emphasis that the reviewability principle does not dispense with explainability; it relocates it. Explainability remains valuable, on this account, primarily as an evidentiary input that makes the four components above more effective—an audit is easier to conduct, a periodic re-evaluation more meaningful, and a judicial remedy more precisely targeted, when the system under review can articulate its own reasoning. The paper's claim is one of priority and sequencing, not of substitution: explainability without institutionalized reviewability is inert, while reviewability substantially strengthens whatever explainability a system can offer. The following part examines how far Taiwan's emerging AI governance framework approximates this model, and what a Stein-inspired reconstruction of its administrative law instruments might look like.

6 TAIWAN'S AI GOVERNANCE FRAMEWORK: A CASE STUDY

Taiwan offers a particularly instructive case study for the reviewability model developed above, for three reasons. First, Taiwan's administrative law tradition, drawing heavily on German and Japanese administrative law scholarship, already possesses many of the doctrinal building blocks—judicial review of administrative disposition (行政處分), procedural rights under the Administrative Procedure Act (行政程序法), and a specialized administrative court system—on which a reviewability-centered AI governance regime could be constructed without wholesale institutional innovation.[11] Second, Taiwan's AI policy is at a formative stage: Taiwan has enacted its first dedicated AI statute, the Artificial Intelligence Basic Act, passed by the Legislative Yuan on 23 December 2025 and promulgated on 14 January 2026,[12] which designates the National Science and Technology Council (NSTC), successor to the former Ministry of Science and Technology, as the central competent authority for AI governance and calls for an internationally aligned AI risk classification framework. Because the Act is deliberately framed as a skeleton of

[11] Administrative Procedure Act [行政程序法] (Taiwan) (amended Jan. 20, 2021), https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030055; Administrative Appeal Act [訴願法] (Taiwan) (amended June 27, 2012), https://law.moj.gov.tw/Eng/LawClass/LawAll.aspx?pcode=A0030020; Administrative Litigation Act [行政訴訟法] (Taiwan), https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030154 (all last visited Sept. 17, 2026).

[12] Artificial Intelligence Basic Act [人工智慧基本法] (Taiwan) (promulgated Jan. 14, 2026), https://law.nstc.gov.tw/LawContent.aspx?id=GL000592; *The Artificial Intelligence Basic Act Is Promulgated and Enters into Force*, LEXOLOGY (Jan. 15, 2026), https://www.lexology.com/library/detail.aspx?g=6cc0ad64-a156-4bb1-9e1e-0bc3d961894f; *Taiwan's AI Basic Act: A Different Logic of Democratic AI Governance*, MEDIALAWS (July 6, 2026), https://www.medialaws.eu/taiwans-ai-basic-act-a-different-logic-of-democratic-ai-governance/.

seven governance principles rather than a set of binding obligations, and delegates risk-classification and enforcement detail to sector-specific competent authorities over a two-year implementation window, the choice between an explainability-centered and a reviewability-centered model of implementation remains genuinely open at the level that matters for this paper's argument, rather than already locked in by legacy regulation. Third, Taiwan's economic position within the global semiconductor and AI supply chain gives the dependency dynamic described in Part 4 a distinctive local inflection: Taiwanese public administration is simultaneously a downstream consumer of foundation models and cloud infrastructure supplied predominantly by non-domestic firms, and an upstream supplier, through its semiconductor industry, of the computational hardware on which those same firms depend—a structural position that complicates, without dissolving, the capture dynamic this paper has described.

Taiwan's evolving AI governance architecture has, to date, leaned heavily on instruments recognizable from the explainability paradigm. Policy documents and draft legislative frameworks emphasize risk-tiered classification of AI systems (echoing the EU AI Act's structure), disclosure obligations for high-risk applications, and the establishment of an AI safety or governance institute tasked with technical evaluation of models. These instruments are valuable and this paper does not counsel their abandonment. But considered against the four components of the reviewability principle set out in Part 5, current and proposed Taiwanese arrangements remain heavily weighted toward disclosure and technical evaluation, and comparatively underdeveloped with respect to contestability, independent auditability, periodic structural re-evaluation, and binding corrective remedy specific to public-sector AI-ADM.

Consider contestability at the point of decision. Taiwan's Administrative Procedure Act already establishes a general right to a statement of reasons for administrative dispositions and a general framework for administrative appeal (訴願) followed by administrative litigation before the administrative courts.[13] These general instruments extend, in principle, to dispositions informed or produced by automated systems, since Taiwanese administrative law does not currently exempt automated decisions from the ordinary requirements governing administrative dispositions. The practical difficulty is that these general instruments were designed for a world in which the reasons an administrator gives for a decision are also the reasons that actually produced it—a correspondence that AI-ADM systems, particularly those built on complex machine learning models procured from external vendors, cannot be assumed to preserve. A reviewability-consistent reconstruction would require Taiwan's administrative procedure framework to be supplemented with AI-specific provisions clarifying that a citizen's right to a statement of reasons under Article 96 of the Administrative Procedure Act extends to a right to know that a decision was substantially informed by an automated system, coupled with a right to request human re-determination that is not satisfied merely by an official's post hoc ratification of the algorithmic output.[14]

Consider next structural auditability independent of the vendor. Here the gap between current arrangements and the reviewability model is most pronounced. Public procurement of AI systems in Taiwan, as in most jurisdictions, proceeds through the Government Procurement Act, whose provisions were drafted with conventional goods and services in mind and do not, absent specific amendment or

---

[13] Administrative Appeal Act [訴願法] (Taiwan), https://law.moj.gov.tw/Eng/LawClass/LawAll.aspx?pcode=A0030020; Administrative Litigation Act [行政訴訟法] (Taiwan), art. 4, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030154 (both last visited Sept. 17, 2026).

[14] Administrative Procedure Act [行政程序法] (Taiwan), art. 96, https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030055 (last visited Sept. 17, 2026).

contractual innovation, mandate the kind of ongoing audit access, training-data disclosure, or model documentation rights that structural auditability requires.[15] A Stein-inspired reconstruction would treat procurement law itself as a central site of administrative-law reform: government contracts for AI-ADM systems should, as a matter of mandatory procurement policy rather than case-by-case negotiation, include standing audit rights exercisable by an independent oversight body (discussed further below), data and model portability provisions that prevent vendor lock-in from becoming a de facto veto over administrative correction, and contractual terms that subordinate vendor intellectual property claims to the administration's overriding need to demonstrate, to a reviewing court, the actual basis on which a contested decision was reached.

Consider periodic, independent re-evaluation. Taiwan's draft AI governance instruments have moved some distance toward this component through proposals for risk classification and periodic technical assessment of high-risk systems, plausibly to be housed within a National Institute of Cyber Security or an analogous technical body.[16] The reviewability model developed here would extend this instinct specifically to public-sector AI-ADM by proposing a standing administrative body—analogous in function, though not necessarily in form, to the Control Yuan's existing oversight jurisdiction over administrative maladministration—charged with conducting mandatory, scheduled reviews of deployed AI-ADM systems' aggregate outcomes, independent of individual complaints, with findings reportable to the Legislative Yuan and directly actionable by the administrative courts.

Consider, finally, binding corrective remedy. Taiwan's administrative courts already possess, under the Administrative Litigation Act, the power to annul unlawful administrative dispositions and, in appropriate cases, to order administration to make a specific decision or to reconsider within a specified framework. These remedial powers are, in principle, adequate to the reviewability model's fourth component; the reconstruction required here is less a matter of new remedial authority than of ensuring that the courts' existing remedial powers are not frustrated in practice by the technical opacity or vendor-controlled documentation that currently surrounds AI-ADM systems—which returns the analysis to the auditability component discussed above. In other words, Taiwan's existing administrative law doctrine supplies much of the remedial architecture the reviewability principle requires; what is missing is the evidentiary and structural infrastructure—independent audit capacity, procurement-law-mandated transparency to the reviewing court rather than merely to the administration—that would allow that existing architecture to function against AI-ADM systems as effectively as it functions against conventional administrative action.

It is also worth situating Taiwan's approach comparatively against the European Union's AI Act,[17] which has functioned as the de facto international template for much recent AI legislation, including elements of Taiwan's own newly enacted Artificial Intelligence Basic Act. The EU AI Act's risk-tiered structure—prohibiting certain uses outright, subjecting "high-risk" systems to conformity assessment, documentation, and human oversight requirements, and leaving lower-risk systems to voluntary codes of

---

[15] Government Procurement Act [政府採購法] (Taiwan) (promulgated May 27, 1998), https://law.moj.gov.tw/ENG/LawClass/LawAll.aspx?pcode=A0030057 (last visited Sept. 17, 2026).

[16] National Institute of Cyber Security, *About Us*, https://www.nics.nat.gov.tw/en/about/introduction/ (last visited Sept. 17, 2026); Alex Myslinski, *The Role of Taiwan's National Institute of Cyber Security*, TAIWAN BUSINESS TOPICS (Feb. 10, 2025), https://topics.amcham.com.tw/2025/02/the-role-of-taiwans-national-institute-of-cyber-security/.

[17] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, 2024 O.J. (L 1689).

conduct—is, in the vocabulary developed in this paper, a paradigmatically explainability- and disclosure-centered instrument. Conformity assessment asks whether a system meets a technical standard and is documented as doing so; it does not, of its own force, create the individualized contestability, vendor-independent auditability, periodic re-evaluation, or binding corrective remedy that the reviewability principle requires, although national implementing legislation in EU member states may supply some of these elements through pre-existing administrative law doctrine, much as this paper argues Taiwan's existing doctrine could be extended to do. The risk for Taiwan, in adopting elements of the EU AI Act's structure without a parallel, deliberate integration with its own administrative law tradition, is that the country ends up with a technically sophisticated risk-classification and disclosure regime running alongside, rather than through, the administrative law instruments—the Administrative Procedure Act, the Government Procurement Act, the Control Yuan, the administrative courts—that Stein's framework suggests are the more decisive site of accountability. A reviewability-conscious approach to implementing Taiwan's now-enacted Artificial Intelligence Basic Act would treat the EU-style risk classification called for under the Act's implementing framework as a floor for technical documentation, while making explicit, in the subordinate regulations that sector-specific competent authorities are to issue within the Act's two-year implementation window, that classification as high-risk triggers not only disclosure and conformity-assessment obligations but also the specific contestability, auditability, and remedial mechanisms set out in Part 5, cross-referenced directly to the relevant provisions of the Administrative Procedure Act, the Government Procurement Act, and the Administrative Litigation Act, so that the AI-specific statute functions as an integrating amendment to existing administrative law rather than a freestanding parallel track.

The institutional design of any standing oversight body also merits closer attention than existing proposals have generally given it. A body modeled loosely on the Control Yuan's investigatory jurisdiction, but purpose-built for AI-ADM oversight, would need, to satisfy the reviewability principle's third component, statutory authority to initiate review of a deployed system independent of any individual complaint, a technical staff or a standing arrangement with independent technical auditors not employed by or contracted through the system's vendor, and a statutory obligation to report findings in a form usable by the administrative courts in subsequent litigation under the Administrative Litigation Act. Absent this last feature, an oversight body's findings risk remaining advisory in the same sense that vendor-authored algorithmic impact assessments are currently advisory: informative, but without the binding corrective force the reviewability principle's fourth component demands. Taiwan's Control Yuan already possesses, under the Constitution and the Control Act, corrective measure (糾正) and impeachment powers directed at maladministration; extending an analogous, AI-specific corrective measure power—paired with a formal channel connecting Control Yuan findings to standing to sue in the administrative courts—would substantially close the gap between Taiwan's existing oversight architecture and the reviewability model without requiring a wholly new constitutional organ.[18]

Taken together, this case study suggests that the reconstruction Taiwan's AI governance framework requires is less a matter of importing entirely new doctrines than of extending, deliberately and self-consciously, existing administrative law instruments—the Administrative Procedure Act's reasons requirement, the Government Procurement Act's contractual architecture, the Control Yuan's oversight

[18] CONSTITUTION OF THE REPUBLIC OF CHINA art. 97; Control Act [監察法] (Taiwan); Control Yuan, *Impeachment*, https://www.cy.gov.tw/en/cp.aspx?n=241 (last visited Sept. 17, 2026).

jurisdiction, and the Administrative Litigation Act's remedial powers—into a domain (AI-ADM) that current drafting has left to a parallel, disclosure-centered track modeled on the EU AI Act. The Stein-inspired argument of this paper suggests that this parallel track, however useful as a technical risk-management framework, cannot substitute for the extension of ordinary administrative law's contestability and remedial machinery to automated decisions, precisely because it is that ordinary machinery, and not technical disclosure standards, that historically has performed the function Stein assigned to self-administration: keeping administrative power answerable to the whole of the society it governs, rather than to the particular interests, however well-documented, of whichever actor currently controls the means of administrative cognition.

7 OBJECTIONS AND LIMITATIONS

Four objections merit consideration. The first is that the analogy between technology firms and Stein's industrial bourgeoisie is overdrawn: technology firms, unlike a nineteenth-century propertied class, do not seek direct political representation and are, in many jurisdictions, subject to competition law, data protection law, and sector-specific regulation that Stein's industrial capitalists never faced. This objection has force but does not undermine the paper's structural claim. Stein's own analysis did not require that a capturing class seek direct political representation; the danger he identified operated equally, and in his view more insidiously, through administration's ordinary dependency on that class's resources, independent of any deliberate political strategy on the class's part. The existence of competition and data protection law is precisely analogous to the "amelioration" Stein criticized: valuable correctives to particular abuses that leave the underlying structural dependency intact, since neither competition law nor data protection law is designed to, or does, guarantee public administration independent capacity to audit and correct the AI systems it procures.

The second objection concerns feasibility: the reviewability principle's demand for independent, in-house or third-party technical audit capacity may exceed the fiscal and human-capital resources of many public administrations, particularly outside the small number of jurisdictions with substantial domestic technical sectors. This is a genuine limitation, and the paper does not claim that reviewability can be achieved costlessly. It suggests, however, that the appropriate response is regional or international pooling of independent audit capacity—an approach with precedent in other domains of technical regulatory capacity, such as pharmaceutical and nuclear safety regulation[19]—rather than acceptance of vendor-authored disclosure as a substitute for independent review. Taiwan's position within the global semiconductor supply chain, noted in Part 6, may in fact afford it comparative technical capacity to develop or participate in such pooled review infrastructure, a possibility beyond this paper's scope but worth flagging for future research.

The third objection is theoretical rather than practical: critics sympathetic to Stein's broader corpus may object that this paper's use of *Selbstverwaltung* strips the concept of its original institutional referent (nineteenth-century local self-government) and repurposes it as a general metaphor for institutional feedback, risking anachronism. The paper accepts that its use of Stein is reconstructive rather than exegetical:

[19] International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, *Mission*, https://www.ich.org (last visited Sept. 17, 2026); International Atomic Energy Agency, *Nuclear Harmonization and Standardization Initiative*, https://nucleus.iaea.org/sites/smr/SitePages/Nuclear-Harmonization-and-Standardization-Initiative.aspx (last visited Sept. 17, 2026).

it does not claim that Stein anticipated artificial intelligence, but that the structural logic of his diagnosis—administration's vulnerability to capture by whichever social force controls a scarce and decisive resource, and the remedy of institutionalized, continuous reviewability—retains explanatory and normative purchase when that scarce resource is computational and epistemic rather than industrial and financial. This reconstructive use of historical theory to illuminate a contemporary institutional problem follows a well-established mode of argument in comparative administrative law and public law theory more broadly, and the paper's claims should be assessed on the strength of the structural analogy, not on fidelity to every detail of Stein's nineteenth-century institutional referents.

A fourth objection concerns the risk of regulatory overreach: demanding vendor-independent auditability and binding corrective remedy for all AI-ADM systems, however minor their impact, could impose compliance burdens disproportionate to the interests at stake and could deter beneficial adoption of AI tools that improve administrative efficiency without materially affecting individual rights. This objection is well taken as a caution against indiscriminate application, but it does not tell against the reviewability principle as such; it tells in favor of calibrating the principle's four components to the stakes of the decision at issue, much as existing administrative law already calibrates procedural protection to the severity of the interest affected—compare, for instance, the more exacting procedural requirements Taiwanese administrative law attaches to dispositions affecting personal liberty or livelihood against the lighter requirements attaching to routine administrative convenience. A calibrated version of the reviewability principle would reserve the full complement of contestability, independent auditability, periodic re-evaluation, and binding remedy for AI-ADM systems materially affecting individual rights or the equitable distribution of public benefits and burdens—systems of exactly the kind the EU AI Act designates as "high-risk," and that the risk-classification framework mandated under Taiwan's newly enacted Artificial Intelligence Basic Act is intended to capture[20]—while permitting lighter-touch, primarily disclosure-based governance for lower-stakes administrative automation. This calibration is fully consistent with Stein's own method, which nowhere suggests that every administrative act requires the same intensity of self-administrative feedback, only that the intensity of feedback should track the potential for a given administrative function to consolidate capture by a particular interest against the interest of the whole.

8 Conclusion

This paper has argued that the accountability of AI-driven automated decision-making in public administration should be evaluated not by the interpretability of algorithmic outputs but by the robustness of the institutional channels through which state and society can jointly contest and correct administrative power exercised through AI systems. Drawing on Lorenz von Stein's dialectical theory of state and society, the paper has argued that the concentration of data, models, and computational infrastructure in the hands of a small number of technology firms reproduces the structural risk Stein associated with class capture of the administrative state, and that the explainability paradigm currently dominant in AI governance discourse reproduces the narrowly technocratic, disclosure-centered response to that risk that Stein explicitly rejected in favor of institutionalized self-administration. The paper has proposed a reviewability principle, comprising contestability at the point of decision, structural auditability independent of the vendor, periodic independent re-evaluation, and binding corrective remedy, as the administrative-law

[20] Artificial Intelligence Basic Act [人工智慧基本法] (Taiwan), art. 16; Regulation (EU) 2024/1689 (cited above at note 17).

operationalization of Stein's demand for continuous, institutionalized feedback between administration and society. Taiwan's emerging AI governance framework, examined as a case study, illustrates both the promise and the current limitations of this approach: substantial existing administrative law infrastructure—the Administrative Procedure Act, the Government Procurement Act, the Control Yuan's oversight jurisdiction, and the Administrative Litigation Act—supplies much of what a reviewability-centered regime requires, but current AI-specific policy has, to date, developed largely along a parallel, disclosure-centered track that has not yet been deliberately integrated with this existing architecture.

The broader implication for algorithmic accountability scholarship is that the field's heavy investment in the technical and legal apparatus of explainability, while not misguided, has been incomplete in a specific and remediable way: it has addressed the question of whether automated decisions can be understood while leaving comparatively underdeveloped the older and more demanding administrative-law question of whether they can be undone, and by whom. Recovering this second question, and the institutional resources of administrative law traditions—Taiwan's among them—that were built to answer it, offers a more secure foundation for algorithmic accountability than any refinement of explanation-generating technique is likely to provide on its own. Future work might extend this framework comparatively, examining whether jurisdictions with differently structured administrative law traditions and differently structured relationships to global technology infrastructure arrive at correspondingly different institutional solutions to the capture risk this paper has identified, while confronting the same underlying structural problem Stein diagnosed a century and a half ago in a very different technological register.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.2

CHAPTER 2

Governing by Algorithm: Constitutional Limits on Automated State Decisions in India

DR. MINI S[*] AND DR. JOBY BHASKER[**]

As Indian public administration becomes increasingly data-driven, algorithmic systems may score, rank, flag, recommend or effectively determine outcomes in welfare administration, public employment, taxation, education and access to essential services. The constitutional difficulty is not automation as such, but the displacement of intelligible and contestable public power by systems whose decisive contribution may be hidden behind a nominal human signature. This article asks when automated State decision-making should trigger a constitutional entitlement to human reconsideration, and what that reconsideration must entail. It uses doctrinal analysis of Articles 14 and 21, post-2021 Supreme Court jurisprudence on reasons, disclosure and natural justice, recent High Court decisions arising from digital administration, India's emerging AI-governance and data-protection framework, and a focused comparison with European Union and United Kingdom law. The article rejects an absolute "right to a human decision." Such a right rests too heavily on an assumption of human decisional superiority and sits uneasily with the context-sensitive character of Indian natural justice. Instead, it argues that materially adverse State action should be constitutionally contestable whenever an automated system materially contributes to the result. In that setting, meaningful human reconsideration requires notice of the system's role, disclosure of decision-relevant material, intelligible reasons, an opportunity to identify error and contextual facts, review by an officer with genuine authority to depart from the automated output, and a record adequate for judicial review. The proposed "material contribution–material adversity" test deliberately extends beyond decisions that are formally "solely automated." It preserves administrative use of beneficial automation while preventing human involvement from becoming a ceremonial rubber stamp. The argument therefore locates AI accountability within familiar constitutional commitments: non-arbitrariness, fair procedure, reasoned government and effective review.

Keywords: Automated decision-making, Article 14, Article 21, Natural justice, Human reconsideration.

1 INTRODUCTION

The constitutional problem created by automated government is often framed too dramatically. The question is not whether a machine may ever participate in public decision-making. Public administration has always relied on instruments that classify, aggregate and standardise information. Digital systems can reduce delay, expose inconsistent treatment, assist overburdened officials and make public services easier to administer at scale. The harder question arises when computational systems cease to be merely clerical and begin to structure the reasons for State action: a welfare claimant is flagged as ineligible, an applicant is ranked out of a public opportunity, a tax case is selected or assessed through a digital workflow, or access to an essential service turns on an automated risk or identity determination. At that point, efficiency collides with constitutional accountability. A citizen may confront an adverse result without knowing that

[*] Associate Professor at Government Law College, Ernakulam, Kerala, India.
[**] Assistant Professor at Government Law College, Ernakulam, Kerala, India.

automation mattered, what information drove the result, or whether any official was capable of reconsidering it.[1]

Indian law does not yet contain a general statute governing automated public decisions. The Digital Personal Data Protection Act, 2023 ("DPDP Act") is relevant because automated administration depends upon data processing and inference, but it does not create a GDPR-style right directed specifically at automated decisions. More importantly, its substantive processing duties and individual rights are being brought into force in stages. Under G.S.R. 843(E) of 13 November 2025, sections 3–17 and much of the core rights-and-obligations architecture fall within the group commencing eighteen months from publication; as of 30 August 2026, that group is not yet operative.[2] India's 2025 AI Governance Guidelines move closer to the present problem. They place people at the centre, call for meaningful control and human oversight, encourage grievance redress, transparency and algorithmic auditing, and recognise threats to livelihood and well-being as especially salient. Yet they remain governance guidance rather than a self-executing constitutional remedy.[3]

The gap is therefore neither a complete absence of law nor a simple demand for new AI legislation. Indian constitutional and administrative law already insist that coercive or materially adverse public decisions remain non-arbitrary, procedurally fair, reasoned and judicially reviewable. The Supreme Court's post-2021 jurisprudence has strengthened those requirements. T. Takano v. Securities and Exchange Board of India connected fair adjudication to disclosure of relevant material; State Bank of India v. Rajesh Agarwal treated natural justice as a substantive restraint on administrative arbitrariness; and Madhyamam Broadcasting Ltd. v. Union of India located reason-giving, disclosure and reviewability within a constitutional "culture of justification."[4] In 2026, however, State Bank of India v. Amit Iron Pvt. Ltd. supplied an equally important limit: natural justice does not invariably require an oral or personal hearing. Notice, disclosure, a meaningful opportunity to answer and a reasoned decision may satisfy fairness, depending upon statutory context and consequences.[5]

Existing Indian scholarship has already identified algorithmic administration as a problem for administrative law, and recent writing has pressed for a constitutional right to reasons in an emerging algorithmic State.[6] The unresolved question is more specific: what procedural entitlement follows once an

---

[1] *See* Divij Joshi, *Automated Administration: Administrative Law and Algorithmic Decision-Making in India*, *in* THE PHILOSOPHY AND LAW OF INFORMATION REGULATION IN INDIA (Sudhir Krishnaswamy & Divij Joshi eds., 2022); Aya Rizk & Ida Lindgren, *Automated Decision-Making in Public Administration: Changing the Decision Space Between Public Officials and Citizens*, 42 GOV'T INFO. Q. 102061 (2025), https://doi.org/10.1016/j.giq.2025.102061.

[2] Digital Personal Data Protection Act, 2023; Ministry of Electronics & Information Technology, Notification G.S.R. 843(E), Gazette of India, Extraordinary, pt. II sec. 3(i) (Nov. 13, 2025), https://egazette.gov.in/WriteReadData/2025/267647.pdf. The notification brings different provisions into force on different schedules; the principal processing duties and rights in sections 3-17 are in the eighteen-month group.

[3] MINISTRY OF ELECTRONICS & INFORMATION TECHNOLOGY, INDIA AI GOVERNANCE GUIDELINES: ENABLING SAFE AND TRUSTED AI INNOVATION 5, 30–33, 42–43 (2025), https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf. The Guidelines describe "People First" in terms of human-centric design, oversight and empowerment; recommend review and override of AI outputs at critical decision points; and support grievance redress, transparency and algorithmic auditing.

[4] T. Takano v. Sec. & Exch. Bd. of India, (2022) 8 SCC 162, ¶¶ 39–44; State Bank of India v. Rajesh Agarwal, (2023) 6 SCC 1, ¶¶ 71, 78–81; Madhyamam Broad. Ltd. v. Union of India, 2023 INSC 324, ¶¶ 52, 56–59, 102, 170.

[5] State Bank of India v. Amit Iron Pvt. Ltd., 2026 INSC 323, ¶¶ 79–81, 93–96, 108–09, 124. The Court relied inter alia on *Union of India v. Jyoti Prakash Mitter*, (1971) 1 SCC 396 (Constitution Bench), and *State of Maharashtra v. Lok Shikshan Sanstha*, (1971) 2 SCC 410 (Constitution Bench), for the contextual character of natural justice and the absence of a universal right to personal hearing.

[6] Joshi, *supra* note 1; Nandini Shukla, *The Constitutional Duty to Explain Automated Decisions: Towards a Right to Reasons in India's Emerging Algorithmic State*, 6 JUS CORPUS L.J. 346, 346–60 (2026), https://doi.org/10.66918/juscorpus.v6i4.2026.43. The present argument differs by locating the operative entitlement in reconsideration, using a material-contribution trigger rather than treating explanation alone as the remedy.

algorithm has materially shaped an adverse State decision? A general "right to a human decision" is both over-inclusive and under-theorised. It would protect trivial decisions merely because they are automated while saying too little about nominally human decisions that simply rubber-stamp an algorithmic score. This article develops a narrower doctrine. It argues that Articles 14 and 21 should guarantee meaningful human reconsideration where two conditions coincide: an automated system has materially contributed to State action, and that action has a materially adverse effect on rights, benefits, legal status, public opportunities or access to essential public services. The inquiry is doctrinal and comparative rather than empirical. It reads recent Indian public-law decisions together with emerging policy and data-protection developments, and uses the European Union and United Kingdom as contrasting regulatory models. The objective is not to constitutionalise technical design preferences. It is to preserve a basic proposition of public law: when the State acts seriously against a person, responsibility for the decision must remain intelligible, contestable and attributable.

2 AUTOMATED STATE ACTION AND THE PROBLEM OF THE NOMINAL HUMAN

Automated decision-making is not a binary category. A system may make a final determination without intervention; it may recommend an outcome that officials usually accept; it may assign a risk score or eligibility classification; it may prioritise cases for scrutiny; or it may filter evidence before an official sees the file. The legal importance of automation therefore cannot depend exclusively on who clicks the final button. A formally human decision may be functionally algorithmic when the officer lacks access to the underlying reasons, is institutionally expected to follow the score, or cannot practically revisit the model's assumptions. Conversely, a genuinely assistive tool may improve administration without displacing human judgment. Scholarship on public-sector automation accordingly emphasises that automation changes the "decision space" between officials and citizens, not merely the identity of the final decision-maker.[7]

India's faceless income-tax architecture illustrates the procedural stakes even though a faceless assessment should not automatically be described as an AI decision. Digital administration can separate the affected person from the decision-maker, channel communication through portals and standardise workflow. In D.B. Engineering Pvt. Ltd. v. National Faceless Assessment Centre, the Delhi High Court set aside an assessment where a hearing requested under the then-applicable statutory scheme was denied. The Bombay High Court reached a similar result in Sulzer Pumps India Pvt. Ltd. v. National Faceless Assessment Centre, and the Patna High Court in Vaibhav Singh v. National Faceless Assessment Centre treated failure to act on an e-portal request for video hearing as a clear violation of natural justice.[8] These cases do not establish a constitutional right to an oral hearing in digital administration. Their deeper significance is institutional: digitisation does not dilute the duty to provide the procedure that law requires, and a portal cannot become the point at which a legally relevant request disappears.

The same danger becomes sharper when an algorithm contributes substantive judgment. Automation bias may induce officials to treat computational output as objective, while proprietary design may prevent

[7] Rizk & Lindgren, *supra* note 1; Peter Parycek, Verena Schmid & Anna-Sophie Novak, *Artificial Intelligence (AI) and Automation in Administrative Procedures: Potentials, Limitations, and Framework Conditions*, 15 J. KNOWLEDGE ECON. 8390, 8390–415 (2024), https://doi.org/10.1007/s13132-023-01433-3.

[8] D.B. Eng'g Pvt. Ltd. v. Nat'l Faceless Assessment Ctr., W.P.(C) 11754/2021, 2022/DHC/004895, ¶¶ 4–7 (Del. HC Nov. 9, 2022); Sulzer Pumps India Pvt. Ltd. v. Nat'l Faceless Assessment Ctr., W.P.(L) 12310/2024, 2024:BHC-OS:6568-DB, ¶¶ 1–4 (Bom. HC Apr. 22, 2024); Vaibhav Singh v. Nat'l Faceless Assessment Ctr., C.W.J.C. 8331/2024, ¶¶ 1–5 (Pat. HC July 4, 2024).

them from explaining why the system produced a particular result. A "human-in-the-loop" label then obscures rather than cures the problem. The constitutional question should therefore turn on material contribution. If the automated output substantially influenced the adverse result—by supplying a decisive score, classification, inference, recommendation or evidentiary filter—the State should not escape procedural scrutiny merely because an official formally adopted it. This functional approach is consistent with the broader public-law concern with real decision-making power rather than administrative form.

3 CONSTITUTIONAL FOUNDATIONS OF CONTESTABILITY

Article 14 supplies the first foundation. Modern Indian equality doctrine does not tolerate State arbitrariness simply because the challenged action is technologically mediated. In Madhyamam Broadcasting, the Supreme Court described proportionality as part of a constitutional culture in which public power must be justified. It treated reasons as fundamental to fair administration because they expose the link between material, inference and conclusion, discipline the decision-maker, and permit effective judicial review. The Court also rejected procedures in which the affected party was denied material necessary to challenge the case against it, observing that non-disclosure can reduce review to an exercise conducted in opacity.[9] Algorithmic State action presents the same structural risk in a new form. A decision that cannot be connected to intelligible reasons is difficult to distinguish from an unexplained assertion of power.

T. Takano sharpens the disclosure principle. The Supreme Court held that material relevant to adjudication cannot be insulated from disclosure merely through the authority's assertion that it did not rely on it, where the material has a nexus with the proceedings and could have influenced the outcome.[10] Transposed carefully to automated administration, the principle does not require routine publication of source code. It does require access to decision-relevant information: the data or factual inputs attributed to the person, the principal factors or rules that mattered, and enough information about the system's role to permit a meaningful challenge. A State agency should not be able to answer a request for reasons with the circular statement that "the system generated the result."

Article 21 provides the second foundation. Since Maneka Gandhi v. Union of India, procedure affecting life or personal liberty must be fair, just and reasonable; Articles 14 and 21 together constitutionalise procedural fairness rather than confining it to statutory grace.[11] Rajesh Agarwal illustrates how this operates outside the conventional courtroom. Classification of a borrower's account as fraud carried serious civil consequences. The Court read audi alteram partem into the regulatory process, requiring notice, disclosure sufficient to answer the case, an opportunity to represent, and a reasoned order.[12] The doctrinal lesson is not that every serious administrative decision must reproduce a trial. It is that a process producing material civil consequences must offer a real opportunity to affect the outcome.

Amit Iron is therefore central rather than inconvenient to the present argument. The Court rejected the proposition that natural justice always includes a personal hearing and relied on Constitution Bench authority recognising the contextual character of audi alteram partem. A written opportunity accompanied by relevant material and followed by a reasoned order can be adequate. At the same time, the Court

[9] *Madhyamam Broad.*, 2023 INSC 324, ¶¶ 52, 56–59, 102, 170.
[10] *T. Takano*, (2022) 8 SCC 162, ¶¶ 39–44.
[11] Maneka Gandhi v. Union of India, (1978) 1 SCC 248, 281–84 (seven-judge bench); *see also* S.N. Mukherjee v. Union of India, (1990) 4 SCC 594, 612–13 (Constitution Bench) (reason-giving as a restraint on arbitrariness and an aid to review).
[12] *Rajesh Agarwal*, (2023) 6 SCC 1, ¶¶ 71, 78–81.

reaffirmed disclosure as the rule for a forensic audit report relevant to fraud classification and insisted that the decision-maker engage with the response.[13] Meaningful human reconsideration is compatible with that reasoning. It is not a demand for an interview with every claimant. It is a demand that a competent human officer be capable of understanding the contested basis of an automated outcome, considering the person's answer and changing the result when the answer warrants it.

The privacy dimension is complementary. Puttaswamy recognised privacy as a constitutional guarantee tied to dignity, autonomy and informational control; the Aadhaar judgment subjected large-scale State data architecture to legality and proportionality review.[14] Automated administration adds a further concern because decisions may be driven not only by data supplied by the citizen but by inferred attributes, correlations and profiles. Data accuracy and lawful processing are therefore prerequisites to fair decision-making, but privacy law cannot substitute for administrative justice. Even when data are lawfully processed, the inference drawn from them may be contestable, the model may use a poor proxy, or the individual case may contain context that the system cannot recognise. The DPDP framework, once its substantive provisions become operative, may strengthen data rights and fiduciary duties; it still does not answer who must reconsider an adverse State outcome or what reasons that reviewer must give.[15]

4 From a Right to a Human Decision to a Right to Meaningful Reconsideration

A categorical right to a human decision has intuitive appeal but a weak constitutional fit. Human decision-makers are neither inherently transparent nor invariably fair. They may be inconsistent, biased, hurried or incapable of processing complex data as accurately as well-designed tools. Aziz Huq's influential critique is therefore persuasive to an important extent: neither accuracy, participation nor reason-giving supplies a general justification for banning machine decisions as a class.[16] Stefan Schäferling reaches a more rights-protective position in the context of governmental automation, but the disagreement itself reveals why the Indian doctrine should not rest on a romantic premise that human judgment is always superior.[17]

The constitutional interest is better described as contestability backed by accountable human authority. Reconsideration matters after an adverse output because some forms of error are visible only from the affected person's perspective: an income entry is duplicated; a database incorrectly links identities; an employment gap has a lawful explanation; a welfare rule has been applied to an atypical household; a model treats a proxy as determinative although the statutory standard demands individual judgment. The right should therefore arise when automation materially contributes to materially adverse State action, not whenever software is used somewhere in the workflow.

Pragya Prasun v. Union of India offers a narrow but revealing bridge. In the context of inaccessible digital know-your-customer processes, the Supreme Court recognised digital access as part of Article 21

---

[13] *Amit Iron*, 2026 INSC 323, ¶¶ 79–81, 93–96, 108–09, 124.

[14] K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1, ¶¶ 297, 322–23 (nine-judge bench); K.S. Puttaswamy (Retd.) v. Union of India (Aadhaar), (2019) 1 SCC 1, ¶¶ 124–26 (five-judge bench).

[15] Digital Personal Data Protection Act, 2023, §§ 2, 3–17; Notification G.S.R. 843(E), *supra* note 2. *Compare* Regulation (EU) 2016/679, art. 22, 2016 O.J. (L 119) 1, 46–47, which expressly regulates certain solely automated decisions.

[16] Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 611–88 (2020). Huq argues against a general right to a human decision and instead develops a case for a "well-calibrated machine decision." *Id.* at 612.

[17] STEFAN SCHÄFERLING, GOVERNMENTAL AUTOMATED DECISION-MAKING AND HUMAN RIGHTS: RECONCILING LAW AND INTELLIGENT SYSTEMS 231–99 (2023), https://doi.org/10.1007/978-3-031-48125-3.

and directed institutions to create a human-review mechanism for rejected applications, with a designated officer capable of overriding automated rejection in appropriate cases.[18] The case arose from disability access and cannot simply be transformed into a general AI holding. Its institutional logic is nevertheless important: a safeguard is meaningful only if the human reviewer possesses authority to depart from the automated result. A reviewer who merely verifies that the system was followed is not reconsidering the decision.

This distinction also explains why "right to explanation" and "right to reconsideration" should not be collapsed. Explanation makes a decision intelligible; reconsideration makes it alterable. Reasons without remedial authority may reveal error but leave the individual trapped by it. Conversely, human review without intelligible reasons risks becoming an unstructured second opinion. Constitutionally adequate contestability requires both: information sufficient to identify the basis of the adverse result, and an institutional route through which a responsible officer can correct it.

5 A CONSTITUTIONAL FRAMEWORK FOR HIGH-IMPACT STATE AUTOMATION

The proposed framework begins with a trigger rather than a sectoral list. Two conditions should be satisfied. First, the automated system must have made a material contribution to the decision. "Material" should capture outputs that substantially shape the result—such as a score, ranking, flag, eligibility determination, predictive classification or recommendation that the official treats as presumptively authoritative. Second, the State action must be materially adverse. The clearest cases are denial, suspension or substantial reduction of welfare benefits; exclusion from public employment or education; tax or regulatory determinations carrying significant financial consequences; and denial of essential public services or legal status. Trivial personalisation, clerical automation and tools with no meaningful effect on the outcome should fall outside the doctrine.

Once the trigger is met, six safeguards follow from existing public-law principles. First, the person should receive notice that an automated system materially contributed to the decision. Hidden automation prevents an affected person from knowing what kind of error to challenge. Second, the State should disclose decision-relevant material. This ordinarily includes the material inputs attributed to the person and the principal factors, threshold or rule that drove the adverse outcome. Full source-code disclosure should not be the default; it can be both technically unhelpful and legitimately restricted. What matters is decision-useful disclosure.

Third, the initial decision or review notice should contain intelligible reasons. "Algorithmic score: high risk" is not a reason in the constitutional sense. The explanation must connect the legal criterion to the facts and identify the role played by automated output. Fourth, the affected person must be permitted to contest factual error, data quality, the applicability of the rule, exceptional circumstances and any legally impermissible proxy or inference. Depending on context, written representation may be sufficient. An oral or video hearing should be required where statute provides it, credibility is material, disability or communication needs make writing inadequate, or the complexity and consequences of the case make oral engagement necessary. This contextual position respects Amit Iron rather than evading it.

[18] Pragya Prasun v. Union of India, 2025 INSC 599, ¶¶ 17–18 & direction (xvi). The holding arose in the specific setting of accessible digital KYC; this article relies on the design of the remedy, not on a claim that *Pragya Prasun* itself establishes a general right to human review.

Fifth, reconsideration must be undertaken by a designated human officer with both competence and authority. The officer must have access to the information needed to evaluate the automated contribution, understand the legal standard independently of the model, and possess genuine power to depart from the recommendation. Where the agency itself cannot interrogate a vendor system sufficiently to meet these obligations, procurement is the constitutional problem: the State cannot outsource a public function on terms that make lawful review impossible. The 2026 Parliamentary Standing Committee on Personnel, Public Grievances, Law and Justice moved in precisely this institutional direction, recommending government-wide safeguards for AI-assisted administration, proper record-keeping, enterprise-level controls over providers and retention of final decision-making authority by designated human officers.[19]

Sixth, the agency must preserve a reviewable record. Courts need not audit model architecture in every dispute, but they must be able to determine what system was used, what version operated, what information entered the decision, what output emerged, how the human officer treated it and why the final conclusion followed. Logging and provenance are therefore not merely technical best practices; in high-impact administration they are conditions for effective judicial review. India's AI Governance Guidelines already endorse accountability, understandability, human oversight, grievance mechanisms, transparency and algorithmic auditing.[20] Constitutional doctrine can give these aspirations a more precise procedural consequence when State action becomes materially adverse.

The framework should contain proportionate exceptions. Emergency action may occasionally precede full process where delay would defeat a legitimate and urgent public purpose, but prompt post-decisional reconsideration should follow. Confidentiality, cybersecurity and trade secrets may justify tailored redaction or controlled disclosure, not a blanket refusal to provide decision-useful reasons. National-security settings may require a different calibration, as Madhyamam itself demonstrates, yet even there secrecy must be justified rather than treated as self-validating. The purpose of the framework is not maximum disclosure. It is minimum constitutional contestability.

6 Comparative Perspective: Lessons from the European Union and United Kingdom

European law offers two useful lessons and one caution. Article 22 of the General Data Protection Regulation protects individuals, subject to exceptions, against certain decisions based solely on automated processing that produce legal or similarly significant effects; where an exception applies, safeguards include human intervention, the opportunity to express a view and the ability to contest the decision.[21] The "solely automated" formulation, however, risks excessive formalism. In SCHUFA Holding (Scoring), the Court of Justice treated automated credit scoring as an automated individual decision where a third party

[19] Dep't-Related Parliamentary Standing Comm. on Personnel, Pub. Grievances, Law & Just., 160th Report: Demands for Grants (2026–27) Pertaining to the Department of Personnel and Training ¶¶ 3.20–3.23 (2026), https://sansad.in/getFile/rsnew/Committee_site/Committee_File/Press_ReleaseFile/18/216/890P_2026_3_12.pdf?source=rajyasabha (recommending record-keeping for AI-assisted actions and retention of final decision-making authority with designated human officers).

[20] India AI Governance Guidelines, *supra* note 3, at 29–33, 42–43.

[21] Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 22, 2016 O.J. (L 119) 1, 46–47 (General Data Protection Regulation).

accorded the score a determining role.[22] The functional insight is more valuable for India than the precise statutory test: law should examine what actually drove the outcome.

The EU Artificial Intelligence Act adds a different safeguard. Article 86 grants, for specified high-risk systems and adverse decisions, a right to obtain a clear and meaningful explanation of the role of the AI system and the main elements of the decision. The Regulation's general application date began on 2 August 2026, subject to its detailed transitional provisions.[23] In CK v. Magistrat der Stadt Wien (Dun & Bradstreet), decided under the GDPR, the Court of Justice further held that meaningful information about automated logic must enable the data subject to understand and challenge the decision; protection of trade secrets cannot convert the right into an empty formality.[24] These developments support decision-useful explanation rather than indiscriminate demands for source code.

The United Kingdom now provides a revealing counter-model. Section 80 of the Data (Use and Access) Act 2025 replaced the previous UK GDPR structure with a more permissive regime for significant solely automated decisions in many settings. Yet the reform retained safeguards requiring information, an opportunity to make representations, human intervention and contestation.[25] This matters because contestability is not simply a by-product of European hostility to automation. Even a legislature deliberately widening the space for automated decisions preserved an avenue for consequential decisions to return to human judgment.

International instruments reinforce the same orientation without supplying a ready-made Indian rule. The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law requires procedural safeguards where AI significantly affects human rights; its explanatory materials identify human oversight, including ex ante or ex post human review, as a possible safeguard where AI substantially informs or takes a rights-impacting decision.[26] UNESCO's Recommendation on the Ethics of Artificial Intelligence similarly links transparency, explainability and human oversight to human-rights-respecting AI governance.[27] India need not transplant these instruments. Their value is confirmatory: contemporary AI governance increasingly treats explanation and review as complementary, not competing, safeguards.

7 OBJECTIONS, LIMITS AND INSTITUTIONAL DESIGN

Three objections deserve weight. The first is administrative cost. A right to reconsideration could recreate the very delays automation is meant to reduce. The answer lies in the trigger. The proposed doctrine does not require human re-adjudication of every automated output. It attaches to materially adverse action

[22] Case C-634/21, SCHUFA Holding (Scoring), ECLI:EU:C:2023:957, ¶ 73 (Dec. 7, 2023).

[23] Regulation (EU) 2024/1689 of the European Parliament and of the Council, arts. 86, 111, 113, 2024 O.J. (L 1689) 1, 101, 126–28 (Artificial Intelligence Act). Article 113 sets 2 August 2026 as the general application date while Article 111 contains transition rules for certain systems already placed on the market or put into service.

[24] Case C-203/22, CK v. Magistrat der Stadt Wien (Dun & Bradstreet Austria), ECLI:EU:C:2025:117, ¶¶ 58–61, 67–76 (Feb. 27, 2025).

[25] Data (Use and Access) Act 2025, c. 18, § 80 (UK) (inserting UK GDPR arts. 22A–22D); Dep't for Sci., Innovation & Tech., *Data (Use and Access) Act 2025: Data Protection and Privacy Changes*, GOV.UK (June 27, 2025), https://www.gov.uk/guidance/data-use-and-access-act-2025-data-protection-and-privacy-changes.

[26] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law arts. 14–16, Sept. 5, 2024, C.E.T.S. No. 225; COUNCIL OF EUROPE, EXPLANATORY REPORT TO THE FRAMEWORK CONVENTION ¶ 103 (2024), https://rm.coe.int/1680afae67. India is not treated here as bound by the Convention; it is used as a comparative human-rights benchmark.

[27] UNESCO, RECOMMENDATION ON THE ETHICS OF ARTIFICIAL INTELLIGENCE ¶¶ 37–42, 85–91 (Nov. 23, 2021), https://www.unesco.org/en/artificial-intelligence/recommendation-ethics.

after automation has materially contributed to the result, and ordinarily becomes operative when the person contests that result. Agencies can triage review, standardise disclosure and use automation to assemble the record. What they cannot do is make consequential error practically irreversible.

The second objection is that human review may be performative. Automation bias can lead reviewers to defer to a model even when authorised to depart from it; poorly trained officers may be less competent than the system they review. That objection strengthens rather than defeats the proposed doctrine. "Meaningful" reconsideration must be evaluated institutionally: the reviewer needs training, adequate time, access to the relevant material and authority to reject the output. Aggregate reversal rates, recurring grounds of correction and system-level error should inform internal audit. Human review is not valuable because human intuition is sacred; it is valuable because public law requires an accountable point at which legally relevant reasons and individual context can change the State's position.

The third objection concerns secrecy and technical complexity. Some models are proprietary, some are difficult even for developers to interpret, and some public systems implicate cybersecurity or fraud-prevention concerns. A constitutional right framed as full technical transparency would be fragile for precisely these reasons. The better requirement is functional disclosure. The affected person needs information sufficient to contest the adverse application to her; the reviewing officer and, where necessary, the court may require deeper documentation under confidentiality protections. The CJEU's recent insistence that meaningful explanation need not amount to disclosure of source code but must still enable challenge provides a useful comparative illustration.[28]

The larger institutional lesson is that constitutional compliance must begin before deployment. Agencies procuring high-impact systems should contract for access to logs, documentation, version history, audit support and explanations adequate to discharge public-law duties. A vendor's confidentiality clause cannot determine the content of Article 14. Nor should an agency deploy a system whose outputs it is incapable of defending through reasons. Ex ante governance and ex post reconsideration serve different functions: audits identify systemic defects; reconsideration protects the person whose case has already gone wrong. Neither is a substitute for the other.

8 Conclusion

Automated administration does not require Indian constitutional law to invent a technological exception to familiar principles. Nor does it justify a categorical constitutional preference for humans over machines. The durable concern is the exercise of public power without a meaningful route from adverse output back to accountable judgment.

Articles 14 and 21 already provide the necessary doctrinal materials: non-arbitrariness, fair procedure, disclosure of relevant material, reasoned decision-making and effective judicial review. The post-2021 cases make those commitments unusually clear, while Amit Iron prevents them from hardening into an indiscriminate oral-hearing rule. Read together, they support a narrower proposition suited to automated government. Where an automated system materially contributes to materially adverse State action, the affected person should be entitled to know that automation mattered, understand the principal basis of the result, contest error and context, and obtain reconsideration from a human officer with genuine authority to depart from the system.

[28] *CK v. Magistrat der Stadt Wien*, ECLI:EU:C:2025:117, ¶¶ 58–61, 67–76.

The proposed material contribution–material adversity test also avoids a weakness visible in foreign data-protection law. Constitutional accountability should not disappear because a human has formally touched the file. What matters is whether the automated system exercised real decisional influence and whether the resulting State action seriously affected the person. India's emerging AI-governance policy already recognises human control, accountability, audit and grievance redress; Parliament's 2026 committee work likewise insists that designated human officers retain final authority in public administration. The constitutional task is to convert those governance instincts into a disciplined rule of contestability. Efficiency may justify automation. It does not justify an administrative dead end.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.3

CHAPTER 3

Constitutional Limits on Automated Governance: An Article 21 Perspective on Algorithmic Accountability in India

YASHWANT SC*

Artificial Intelligence plays a major role in modern-day, e.g., AI-based automated governance. Using machine learning and natural language processing, this helps in monitoring compliance, predicting risks and enforcing policies across enterprise systems. This study analyses the implications of increasing AI-driven surveillance and governance technologies on the Indian polity in context to Article 21 of the Constitution of India pertaining to right to life and personal liberties. In addition, this paper investigates the literature on protection of constitutional rights under Article 21 by examining case laws and applying, in context of human right violations pertaining from introduction of AI based technologies through constitutional principles on data available until now including privacy, dignity and other constitutionally protected rights. This research examines AI implementation challenges, looking specifically at black-box system implementations inside governmental organizations that have security, defense, protection responsibilities or are mandated to investigate social good (such as housing, employment, civil/affairs) objectives and enforce laws.

Increased usage of AI-based decision-making systems by governmental bodies or authorities for purposes of predictive policing, public decision-making, and welfare distribution increases the threat of transparency and accountability, which is necessary to uphold constitutional guarantees associated with Article 21 provisions in India. This research also helps to understand the relationship between AI-enabled governance and different interpretations of legal provisions and to recommend incorporating procedural mechanisms aimed at enhancing explainability and oversight required for realising these protections through analysis of relevant judgments issued by the Supreme Court in landmark cases such as Maneka Gandhi v. Union of India and Justice K.S. Puttaswamy v. Union of India relating to Right to Privacy and Dignity. This research paper provides a broad scope for addressing concerns arising out of emerging technological developments, failure to establish specific processes for ensuring algorithmic transparency and accountability which could potentially lead to unconstitutional governance. It also provides effective suggestions for the creation of a rights-based regulatory framework requiring incorporation of both technical and procedural safeguards into AI systems used in the Indian State.

1 INTRODUCTION

The adoption of artificial intelligence (AI) together with algorithmic-based decision-making systems has created new legal challenges which now confront the Indian Constitution, which protects essential human rights through its core principles. The Constitution establishes a strong rights protection system, yet its effectiveness needs evaluation because the digital age brings new challenges which require better understanding of artificial intelligence technologies. The "conflict of artificial intelligence with Indian constitutionalism" is a critical issue which requires resolution through establishing boundaries which protect constitutional rights while allowing scientific advancements to proceed. The rapid adoption of AI technology

* Student at SASTRA Deemed University, Thanjavur, Tamil Nadu, India.

across multiple sectors requires assessment of its impact on existing rights and the need for establishing new rights or different legal interpretations.

Article 21 of India's Constitution, which states that no individual shall be deprived of life or personal liberty except in accordance with legal procedures, has seen a significant extension in court interpretation. The Supreme Court originally used an A.K. Gopalan v. State of Madras decision to define Article 21 rights because the court believed that legal requirements represented all rights protection.[1] The judicial authority to review cases was restricted because courts handled fundamental rights as separate matters from other judicial proceedings.

The Supreme Court departed from this narrow view in Maneka Gandhi v. Union of India, which required that any procedure depriving a person of life or liberty must also be fair, just, and reasonable.[2] The ruling established substantive due process and formed an essential link between Articles 14, 19, and 21, which together protect individual freedoms.

The Article 21 declaration developed into a legal mechanism which provided essential rights that people need to maintain their dignified existence.[3] The Supreme Court expanded its definition of life rights to include essential rights for people who need to survive through work and medical care and educational access and safe housing and unpolluted surroundings. The Supreme Court case Justice K.S. Puttaswamy v. Union of India[4] marked a major breakthrough when it recognized privacy rights as fundamental rights which every person possesses through their existence and personal freedom. The evolution of Article 21 shows its ability to adapt to new challenges especially those connected to digital governance and artificial intelligence.

2 RISE OF ALGORITHMIC GOVERNANCE IN INDIA: IMPLICATIONS FOR ARTICLE 21 AND PRINCIPLES OF NATURAL JUSTICE

The increasing implementation of artificial intelligence (AI) and data-driven technologies in government operations has led to the emergence of algorithmic governance systems throughout India. The government now employs automated systems to manage processes which include welfare distribution systems and digital identity verification and predictive policing and taxation and judicial administration. The technologies deliver operational benefits through higher efficiency and faster performance and unbiased results but they can threaten constitutional rights where they disrupt Article 21 protections and natural justice standards. Article 21 establishes that all life and personal liberty restrictions must comply with fair and just and reasonable procedures according to the Maneka Gandhi v. Union of India decision. Automated decision-making systems used in algorithmic governance may not meet this requirement where they function as black box systems which keep their decision-making processes hidden from users. The absence of system transparency creates problems because essential services like welfare assistance and identity identification depend on these systems to determine user eligibility for vital services. The Supreme Court confirmed the Aadhaar system in K.S. Puttaswamy (Aadhaar) v. Union of India[5] while it recognized the system's problems with user exclusion and data security issues. Biometric authentication failures that

[1] A.K. Gopalan v. State of Madras, AIR 1950 SC 27.
[2] Maneka Gandhi v. Union of India, (1978) 1 SCC 248.
[3] INDIA CONST. art. 21.
[4] Justice K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1.
[5] Justice K.S. Puttaswamy v. Union of India, (2019) 1 SCC 1.

result in welfare benefit denials show how algorithmic techniques can create obstacles to Article 21 rights that protect life and dignity. This suggests that efficiency obtained through technology cannot replace constitutionally protected rights.

The Article 21 fair procedure requirements depend on two natural justice principles which include audi alteram partem and nemo judex in causa sua. The Court in A.K. Kraipak v. Union of India[6] established that administrative action must also follow the principles of natural justice, since there are no clear boundaries between administrative and quasi-judicial powers. The principles which should guide algorithmic governance present challenges for their actual enforcement. The automated decision-making processes create a situation where people cannot understand the results or challenge them which violates their right to a hearing. The system of algorithmic bias leads to decisions that create discrimination which results in a loss of both equal rights and fair treatment. The Supreme Court declared in E.P. Royappa v. State of Tamil Nadu that arbitrariness violates the principle of equality. AI systems produce unfair results because they mirror or enhance existing societal biases which create an unintended impact on the Article 21 dignity protection.[7]

The establishment of privacy as a fundamental right through Justice K.S. Puttaswamy v. Union of India makes algorithmic governance more difficult to implement. Data collection and processing serve as the core foundation for AI systems because these systems require data to operate while people remain unaware and unprotected. The Court's emphasis on informational self-determination highlights the importance of regulating the state's use of personal data in automated systems.[8]

The current execution of algorithmic governance may lead to major constitutional problems despite its ability to create transformative changes. The three main issues which AI systems present to Article 21 and natural justice principles are their hidden internal workings, their unidentified system biases, and their absence of defined operational safeguards. The development of a constitutional framework which guarantees transparency and accountability alongside the right to explanation in AI-based decision-making needs to happen because it connects technical governance with fundamental rights.

3 Challenges

AI systems require large amounts of data to train and make choices. Data security and privacy are crucial. The protection of sensitive citizen information needs to be maintained because it presents a major security risk. Personal information must be protected from misuse as well as unauthorized access and data breaches. Previous data contains biases that AI systems will use to make decisions. AI decision-making processes need to achieve both equitable outcomes and transparent results. AI creates ethical problems through its responsibility issues and its capacity to develop biased systems which produce unexpected results. The public and government agencies show doubt towards decisions made with AI technology. AI systems gain public trust through effective communication and understandable system operation and open information sharing. The concept of AI as a "black box" needs to be rejected for successful implementation. E-governance needs people to access the internet. AI-powered services face deployment difficulties in regions with limited connectivity. E-governance aims to provide equal access to AI services

[6] A.K. Kraipak v. Union of India, (1969) 2 SCC 262.
[7] E.P. Royappa v. State of Tamil Nadu, (1974) 4 SCC 3.
[8] Justice K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1.

while reducing the existing digital technology divide. The development of professional skills together with capacity building needs to be established as an essential requirement. The training program for government staff needs to include AI principles together with applications and management training. The process of creating a workforce that can handle AI systems presents multiple difficulties. The operation of data centers and computational resources together with AI infrastructure construction leads to environmental damage. The need for technological progress must reach equal standing with environmental protection efforts. It remains difficult to reduce the carbon footprint produced by AI systems. The combination of AI and traditional governance methods presents substantial operational difficulties. AI technology may not work properly with existing outdated systems. The process of achieving compatibility and handling transitions between systems proves to be a demanding task.

The process of creating legal and regulatory systems for government AI usage presents considerable difficulties. The process requires maintaining an equilibrium between innovative solutions and safety measures. The responsibilities of duty and liability together with intellectual property rights for AI systems must be managed in a proper manner. The implementation of AI technologies to serve extensive user bases incurs high expenses. Governments face two primary challenges which include maintaining cost-effectiveness and adhering to budget limitations. Governments need to evaluate the advantages of artificial intelligence together with its related expenses. The artificial intelligence landscape requires governments to change their operational methods. Governments face difficulties in keeping up with technological progress while using AI to its full potential. E-governance presents problems that involve ethical decision-making. AI systems create ethical challenges which include determining the appropriate decision-making process and assessing their effect on employment and the ethical ramifications of automated task execution. The implementation of AI in e-governance requires establishing a balance between ethical standards and technological progress.

4 TOWARDS A CONSTITUTIONAL GOVERNANCE FRAMEWORK FOR ARTIFICIAL INTELLIGENCE IN INDIA

The fast introduction of artificial intelligence (AI) into government operations requires the establishment of a constitutional governance system which protects essential rights while enabling technological progress. In India, such a framework must be based on Article 21 which demands that any state action affecting life or personal liberty be fair, just, and reasonable as outlined in Maneka Gandhi v. Union of India. Government agencies which use algorithmic decision-making systems need to deal with three main challenges which arise because these systems operate with hidden processes and extensive data and work without human oversight. The implementation of AI systems requires a rights-based regulatory structure which ensures their operation according to constitutional principles.

The principle of procedural fairness functions as an essential element which Article 21 establishes to create a constitutional framework that governs artificial intelligence. The established definition of fair procedures requires three elements which include notice and hearing and reasoned decision-making. The algorithmic systems of today function as black boxes because they produce results without revealing their operational mechanisms. This situation prevents people from understanding the reasons behind decisions which might breach their legal rights. The framework requires a "right to explanation" provision which needs to inform people about decision-making processes and their underlying reasons. The requirement

establishes ties to Article 14 non-arbitrariness principles while it also reinforces Article 21 due process rights.

The demand for system fairness creates a direct link to the need for organizations to achieve system transparency together with system accountability. Public authorities that utilize AI technology need to reveal information about both their system design and operational procedures and system objectives. AI implementation in high-risk domains such as welfare distribution and criminal justice and surveillance should require mandatory algorithmic impact evaluations. The assessments will discover potential dangers which include bias and exclusion and rights violations. The E.P. Royappa v. State of Tamil Nadu case enables a constitutional analysis of AI systems which operate without transparency and accountability by linking arbitrary decisions to social injustice.

The framework's essential element protects personal information rights which Justice K.S. Puttaswamy v. Union of India recognized as a fundamental human right. AI systems depend on their ability to obtain and process and evaluate extensive personal data, which may be acquired without proper consent or adequate security measures. The constitutional governance framework needs to establish the existence of all Puttaswamy requirements which include legal compliance and essentiality and proportionality standards. The system needs to implement restrictions which include verifying data collection needs and defining specific collection purposes and applying limits on data collection and implementing robust protection systems. People need complete authority over their personal information which includes their rights to access their data and make changes and delete data.

The principles of natural justice require reevaluation within the framework of algorithmic governance. The right to be heard (audi alteram partem) mandates that those impacted by automated judgments have the ability to challenge them before an independent body. This system requires the development of public complaint resolution systems which include appellate courts that can investigate algorithmic outcomes. The rule against prejudice (nemo judex in causa sua) requires that AI systems be created and tested to avoid discriminatory results. Organizations need to perform regular algorithmic audits which focus on identifying and eliminating bias that specifically targets vulnerable and marginalized communities.

Constitutional AI governance requires institutional supervision as a further fundamental element. The public sector requires independent regulatory agencies or specialized authorities which need to monitor AI systems. The organizations will develop standards which they will use to conduct audits and check compliance with established standards. Parliamentary monitoring is also necessary to maintain democratic accountability in the implementation of AI technologies. The judicial review process needs to exist as a safeguard against unconstitutional AI applications which enables courts to assess whether algorithmic systems fulfill fairness and rationality and proportionality requirements.

A constitutional framework needs to include both legal protections and ethical principles which safeguard human dignity and personal freedom and promote inclusive practices. According to the Supreme Court, dignity functions as a vital element which protects human existence according to Article 21.[9] AI systems which create welfare program exclusion and enable invasive surveillance and establish dehumanized decision-making systems violate this social principle. Human judgment needs to stay in control of essential decisions through a human-in-the-loop system which prevents automated systems from operating without human supervision.

[9] *Article 21 of the Constitution of India: Right to Life and Personal Liberty*, ACADEMIKE (Mar. 18, 2026), https://www.lawctopus.com/academike/article-21-of-the-constitution-of-india-right-to-life-and-personal-liberty.

The existing ideas need legislative action to establish them as a proper regulatory framework through their formal codification. India has made progress in creating data protection laws and policies but it currently lacks a complete legal framework dedicated to artificial intelligence. The law must define government responsibilities to protect individual rights and establish methods for legal enforcement together with effective remedies. The framework should support innovation while protecting constitutional rights during technological development.

Government AI implementation requires a complete constitutional system which protects essential human rights while allowing for technological progress. The digital age requires a framework based on Article 21 which upholds fairness and transparency and privacy protection and accountability standards to maintain the legal order. The Indian government can achieve efficient governance through AI system design by integrating constitutional values which will create a system that protects human rights and operates fairly and compassionately.[10]

5 STATUTORY DIMENSIONS OF CONSTITUTIONAL AI GOVERNANCE: THE DPDP ACT AND THE IT ACT

The complete constitutional governance system for artificial intelligence (AI) in India requires establishment of a robust legal framework. The constitutional principles from Article 21 serve as the main normative base for rights protection while the Digital Personal Data Protection Act of 2023 and Information Technology Act of 2000 establish statutory requirements that implement those rights through digital and algorithmic governance systems.

The Digital Personal Data Protection Act, 2023 (DPDP Act) serves as the primary law in India that governs how organizations handle personal information.[11] The law provides crucial support for AI governance because it establishes rules that control data processing and restricts information usage to specific purposes while applying the fundamental right to privacy recognized in Justice K.S. Puttaswamy v. Union of India. The DPDP Act mandates that "data fiduciaries" must process data exclusively for verified valid objectives which require obtaining consent from the data principal because AI systems depend on extensive databases that frequently include confidential personal data.[12]

The DPDP Act enhances information privacy protections but faces challenges when applied to AI governance. The Act provides state agencies with comprehensive exemptions that they can use to protect their sovereignty and maintain public order and safeguard national security interests. The broad exemptions create a situation where authorities can bypass necessary security measures that Article 21 mandates especially when they deploy AI technology for monitoring or predictive policing. The Act lacks explicit rules for automated decision-making and does not establish a "right to explanation" which prevents it from achieving fair procedural treatment in algorithmic governance.

The Information Technology Act of 2000 (IT Act), which originated before artificial intelligence became popular, stands as India's primary digital law framework. The AI systems operated by private companies face indirect effects from Section 43A which demands proper security measures for handling

[10] NITI AAYOG, RESPONSIBLE AI #AIFORALL: APPROACH DOCUMENT FOR INDIA, PART 1: PRINCIPLES FOR RESPONSIBLE AI (Feb. 2021), https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf.

[11] Digital Personal Data Protection Act, 2023.

[12] Centre for Law & Policy Research, *Justice K.S. Puttaswamy v. Union of India (Privacy)*, TRANS LAW (Dec. 7, 2021), https://translaw.clpr.org.in/case-law/justice-k-s-puttaswamy-anr-vs-union-of-india-ors-privacy/.

sensitive personal data and Section 79 which establishes rules for intermediary accountability.[13] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules of 2021 establish due diligence requirements which AI-powered platforms must follow when they conduct content moderation and automated decision-making processes.[14]

The IT Act requires constitutional interpretation because Article 21 establishes due process standards as explained in Maneka Gandhi v. Union of India. The Act permits executive actions which include content restriction and surveillance operations only when they satisfy the three requirements of justice and rationality and proportionality. The Supreme Court ruling in Shreya Singhal v. Union of India declared Section 66A of the IT Act invalid because it contained ambiguities and arbitrary elements which created obstacles to the implementation of digital regulations that needed to comply with constitutional protections.[15]

The DPDP Act and the IT Act both contain vital elements yet they fail to provide complete solutions for the unique challenges which AI technologies introduce. The two laws fail to provide complete solutions for problems which include algorithmic bias and transparency deficiencies and automated decision-making and responsibility for AI-related damages. The result is a regulatory gap in which constitutional rights risk remaining theoretical unless courts and lawmakers give them effective enforcement.

The establishment of a constitutional governance framework for AI requires expansion through integration of existing statutes and their additional components. This requires the development of specific regulations which will govern algorithmic accountability while establishing required impact assessments and audit procedures and automatic decision-making rights. The legal system becomes more cohesive when the DPDP Act's data protection standards and the IT Act's digital governance framework are synchronized.

The existing digital technology regulations need more extensive legal frameworks because the current laws do not fully address the constitutional issues created by artificial intelligence. The government must develop stronger regulations that give effect to Article 21 principles, which include justice and openness but also require accountable AI use in government.

6 Conclusion

The Article 21 provision exists as a "living" law. The right to privacy and clean environmental rights have both become part of the law, so the Right against Automated Arbitrariness must now be established as a new legal right. Dignity should never be sacrificed because organizations want to operate more efficiently. India must transition from a "Rule by Code" to a "Rule of Law in Code." India should implement its first phase of legislation within the next three years to regulate high-risk areas that encompass finance and healthcare and employment and surveillance and vital infrastructure protection. India should create specific regulations through sector-based rules and IT Act and DPDP Act delegated legislation and AI deployment regulations for major AI organizations instead of establishing one unified AI law. The court

[13] Information Technology Act, 2000.
[14] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, G.S.R. 139(E) (Feb. 25, 2021).
[15] Shreya Singhal v. Union of India, (2015) 5 SCC 1.

system should establish accountability through its decisions, which will create legal precedents for algorithm-based decisions and bias and automated harm.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.4

CHAPTER 4

Judicial Disengagement and the Automated Welfare State

NOURIA RAFI*

The Indian State has come to rely heavily on automated systems for welfare benefit allocation, fundamentally shifting the onus of proof onto marginalised citizens when an algorithmic exclusion occurs. On August 13, 2026, the Supreme Court of India in Narendra Kumar Goswami v. Union of India refused to decide on the constitutional questions associated with automated welfare exclusions. Observations from the Bench, as reported, treated algorithmic governance as a highly technical issue beyond judicial expertise and one belonging to the policy domain; the Court declined to intervene and directed the Centre to consider the petition as a representation.

This paper argues that this judicial disengagement is difficult to reconcile with the Court's own jurisprudence. Where legislative vacuums threatened fundamental rights, the Court has intervened on more than one occasion, often issuing binding interim safeguards until the legislature fills the gap. A contrasting approach to this disengagement is offered by the Dutch SyRI judgment. There, a district court reviewed an automated welfare system using its existing procedural framework, showing that judicial review of such systems does not require specialised technical expertise.

Ultimately, what technological due process (under Article 21) requires is not a new doctrine. It lies in the disciplined application of the Puttaswamy proportionality test. To protect the rule of law, the judiciary must ask whether the state's use of an automated system that threatens fundamental rights is legally authorised, necessary, and proportionate.

Keywords: algorithmic governance, automated systems, judicial disengagement, proportionality, technological due process.

1 INTRODUCTION

The State[1] in India is now increasingly relying on automated systems for the administration of welfare policies. The administrative tasks that were earlier carried out end-to-end by humans now have an algorithm to share the burden. These algorithmic systems are being used to cross-check databases, to flag duplications across the recorded entries, to flag potential frauds and build risk profiles, and to require biometric matching before welfare benefits can be accessed, among other things. These are not mere software upgrades; rather, these represent a systematic automation in governance. Decisions as basic as whether a family will receive food rations, whether wages will be paid to a labourer, or whether an elderly citizen will receive her pension, which until now required human adjudication, are being left to be decided by an algorithm.

To break it down, the automated systems are the platform or point of access to avail any welfare benefit; algorithmic processing is the procedural work done by the AI to process a person's application to verify the information it contains against government datasets; and outputs are the decisions given

* Independent Researcher at Aligarh Muslim University, Aligarh, Uttar Pradesh, India.

[1] Under Article 12 of the Constitution of India, "the State" includes the Government of India and the Government of each State. The systems examined in this paper operate at both levels.

post such processing accepting or rejecting such applications. This structure has altered the entire ecosystem of welfare delivery, and has at times detrimentally affected the fundamental rights of citizens.

Now this alteration has also given rise to a constitutional problem in India. It is settled law under Article 21 that any procedure depriving any person of her right to life and personal liberty must be just, fair and reasonable. It is also well understood that the right to live with dignity rests on the shoulders of policies of a welfare state and their equitable benefit delivery. However, the established jurisprudence presumes that a procedure governing access to such entitlements is open to inspection. When a citizen is wrongfully excluded by algorithmic processing, the exclusion is a mathematical output rather than a reasoned administrative decision. Such a procedure is not aligned with the constitutional standards of transparency in decision making.

The Supreme Court of India was recently asked to address this exact issue in *Narendra Kumar Goswami v. Union of India*[2] on August 13, 2026. The Court, however, declined to decide the petition; observations from the Bench, as reported, indicated two grounds. First, the Court considered the issue of algorithmic governance as a technical one requiring special expertise. Second, the Court considered the issue as belonging to the policy domain of the executive and the legislature. This paper argues that neither ground sits comfortably with the Court's own historical jurisprudence. It also argues that the constitutional standard required to regulate these systems is the one that the Court already possesses.

It is to be noted that the original petition in *Goswami* covered four distinct areas of state automation, but this paper restricts its scope to analyse the issue associated with welfare administration. The remaining domains raise distinct questions that require a different constitutional treatment.

Methodologically, this paper conducts a doctrinal analysis, supported by a single comparative reference point. The paper proceeds in six parts. It starts by tracing the jurisprudential shift under Article 21 and the gap it reveals regarding automation. It then sets out the factual matrix of *Goswami*, what was asked and what was refused. It then tests the grounds of refusal against the Court's own pattern of judicial intervention in cases of legislative vacuum. It then turns to automated welfare systems presently in operation, and the exclusions they have produced in the absence of any legislative or judicial supervision. Next, it makes a comparative analysis contrasting *Goswami* with a Dutch case arising from a similar factual matrix. Finally, it argues that the Court need not devise new constitutional safeguards because the existing proportionality test is a sufficient mechanism to hold the State accountable within such an ecosystem.

2 FROM A. K. GOPALAN TO MANEKA GANDHI, AND THE CONTEMPORARY ALGORITHMIC GAP

2.1 The Shift from Gopalan to Maneka Gandhi

In 1950, the Supreme Court took a strictly textual approach to read "the procedure established by law" under Article 21. As long as the procedure was prescribed by a validly enacted law, the Court would not

[2] Narendra Kumar Goswami v. Union of India, Writ Petition (Civil) No. 837 of 2026 (SC Aug. 13, 2026) (order), https://api.sci.gov.in/supremecourt/2026/36405/36405_2026_1_19_73291_Order_13-Aug-2026.pdf. The written order records that the petitioner had already submitted a representation on the issues raised, and disposes of the petition without expressing any opinion on the merits. The observations made orally from the Bench during the hearing, including the two grounds discussed in this paper, are drawn from contemporaneous reporting of the proceedings: Amisha Shrivastava, *Supreme Court Declines Plea Seeking Regulation of AI Use, Asks Centre to Consider Representation*, LIVELAW (Aug. 13, 2026), https://www.livelaw.in/top-stories/supreme-court-declines-plea-seeking-regulation-of-ai-use-asks-centre-to-consider-representation-545700.

enquire into its fairness.[3] Decades later, in 1978, this position was fundamentally changed. The Court ruled that any procedure depriving a person of life or personal liberty must be just, fair, and reasonable, rather than arbitrary or oppressive.[4] This jurisprudential shift has been thoroughly studied but what matters the most today are the standards that *Maneka Gandhi* proposed. An entire procedure can actually be examined by the courts to assess its fairness; however, this approach presumes the presence of an identifiable human decision maker, reasons for that decision, and a factual record. Thus, the test of fairness is applied by examining how a particular decision was reached. Yet all these legitimate assumptions give out when procedure is automated.

2.2 What the Fairness Standard Assumes

When an algorithmic system gives decisions, it is simply a technologically generated output. Such an output cannot be fairly assessed using the same approach. There is no identified decision-maker who can be held accountable for erroneous outputs; the output is rarely supplemented by a statement of reasons; and there is no accessible record against which the decision may be examined. The term "technological due process" has been used by legal scholars to describe this problem. Danielle Keats Citron observed that automated administrative systems collapse individual adjudication into rigid rulemaking while observing the procedural safeguards of neither.[5]

The Indian law currently lacks any statutory framework to address this specific procedural gap. On the one occasion on which the issue was called for judicial intervention, the Supreme Court refused to engage with it. The following sections trace the consequences of that judicial disengagement.

3 THE GOSWAMI CASE: WHAT WAS ASKED AND WHAT WAS REFUSED

3.1 The Petition and Its Prayers

On August 13, 2026, a three-judge bench of the Supreme Court of India, comprising Chief Justice Surya Kant, Justice Joymalya Bagchi, and Justice V. Mohana, heard a public interest petition filed by advocate Narendra Kumar Goswami.[6] The petitioner prayed before the Court to lay down constitutional safeguards to govern the state's deployment of "high-risk" artificial intelligence systems across four critical domains: welfare, policing, surveillance, and content moderation. Within the specific sphere of welfare, the petition described a range of statutory entitlements which reveal immediate risk of algorithmic exclusion. These entitlements generally relate to food rations, public health benefits, daily wages, old-age pension benefits, scholarships, and targeted subsidies.[7] For the purpose of conceptual comprehensibility, this paper takes up only the welfare aspect of the petition.

It is to be noted that the petition was framed as a claim of constitutional rights against the automation of sovereign power, rather than a mere wishlist for technology policy and AI regulation. It sought a judicial declaration that deploying artificial intelligence in public governance without a governing statutory

[3] A.K. Gopalan v. State of Madras, AIR 1950 SC 27.
[4] Maneka Gandhi v. Union of India, (1978) 1 SCC 248.
[5] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2007).
[6] *Supra* note 2.
[7] *Narendra Kumar Goswami AI Regulation Case: Supreme Court Declines, Asks Centre to Consider Representation*, SUPREMETODAY (Aug. 14, 2026), https://supremetoday.ai/narendra-kumar-goswami-ai-regulation-case-supreme-court-declines-asks-centre-to-consider-representation-20260814045.

framework, without mandatory human oversight, and without a guaranteed right to an explanation is violative of fundamental rights guaranteed under Articles 14, 19(1)(a), 19(1)(g), and 21 of the Constitution.[8] In the alternative, it sought a declaration that any State AI system producing civil consequences must satisfy minimum constitutional safeguards. The petition articulated these minimum standards as the strict requirement of legality, transparency, non-arbitrariness, proportionality, human oversight and effective remedy.[9]

3.2 Two Grounds of Refusal

The bench questioned what tangible relief it could grant in matters of such nature. According to contemporaneous reporting, Chief Justice Kant acknowledged that the petition was very comprehensive but indicated two reasons for the Court's refusal to take up the matter: first, that algorithmic governance is a highly technical issue for which the judiciary is not the expert, and second, that the regulation of such technology falls squarely within the policy domain of the executive and legislature. Accordingly, the Court asked the Centre to consider the representation the petitioner had already submitted, permitted him to supplement it with a copy of the writ petition, and disposed of the matter without expressing any opinion on the merits. Notably, the Court, before concluding, observed that the use of artificial intelligence within the judiciary itself was already regulated by the Court's internal guidelines.[10]

This refusal by the Apex Court rests on two pillars of judicial restraint. The first is institutional propriety, based on the premise that such technological regulation belongs to the executive and legislature. The second is institutional incompetence, based on the premise that algorithmic systems are too technically complex for judicial evaluation. Both are established grounds for declining constitutional intervention. Yet, in the field of fundamental rights, both grounds remain to be tested. The first ground must be tested against the Court's own extensive record of intervening to protect fundamental rights where the Parliament has left a vacuum. The second ground must be tested by asking whether the Court in fact requires technical expertise to apply existing constitutional safeguards to an automated state procedure. It is precisely these questions that this paper now explores.

4 WHEN THE COURT HAS FILLED VACUUMS BEFORE

The first ground of institutional propriety can be evaluated against the Court's own established jurisprudence of intervening to protect fundamental rights where legislative vacuum occurs. This approach is illustrated in *Vishaka v. State of Rajasthan, where the Court dealt* with the pressing issue of widespread workplace harassment, but there was no legislative regulation. Rather than directing the petitioners to approach the executive, the Court itself framed guidelines under Article 32. It declared that these guidelines would be binding until a law is made by the Parliament. The Court drew the substance of these safeguards from an international document: the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW), even though it was not enacted as a domestic law.[11]

---

[8] *Supra* note 2.

[9] *Narendra Kumar Goswami AI Regulation Case: Supreme Court Declines, Asks Centre to Consider Representation*, *supra* note 7.

[10] *Supra* note 2.

[11] Vishaka v. State of Rajasthan, (1997) 6 SCC 241.

4.1 Understanding the Structural Parallel

Here, a structural parallel can be drawn with *Goswami*. Both cases exhibit an apparent legislative vacuum, ongoing harm, an affected class of citizens, and a prayer seeking the Court's intervention or procedural rules rather than asking for a permanent statutory code. This parallel is not merely structural. The petition in *Goswami* expressly prayed that any guidelines framed by the Court operate as law under Articles 141 and 142 until Parliament enacts legislation,[12] and sought the constitution of a High-Powered Expert Committee, including Supreme Court nominees, to frame them.[13] The petitioner, in other words, requested the Court to do precisely what it had done in *Vishaka*. However, there is a stark difference in their factual context. Where *Vishaka* arose from a specific criminal incident with an identifiable victim, *Goswami* sought anticipatory relief in the welfare domain with no individual complainant. Yet, both face the same institutional problem of an unregulated space affecting fundamental rights. In one case, the Court stepped in with interim safeguards, but in the other, the matter was sent to the executive.

Vishaka is not the only instance where the Court stepped up when the legislature could not. The Supreme Court has repeatedly crafted safeguards when the existing legal framework could no longer adequately counter emerging harms. In *M.C. Mehta v. Union of India*, the Court upon realising the inadequacy of the common-law rule in *Rylands v. Fletcher* formulated the doctrine of absolute liability.[14] Similarly, the proportionality test which the Court now applies to rights claims was not made by the Parliament. It was judicially constructed in Modern Dental College and Research Centre v. State of Madhya Pradesh and applied by the Court in *K.S. Puttaswamy v. Union of India* to address the problems of modern data practices.[15]

4.2 Was the Restraint the Right Move?

This does not mean that the Court should intervene every time there is a legislative void. Judicial overreach is frowned upon and courts cannot take over legislative functions. Judicial restraint has its place. However, the reasons given by the Court in *Goswami* present an inconsistency with its own past actions. Creating a workable framework, albeit on an interim basis, to protect fundamental rights was the exact step the Court took in the case of Vishaka. Technical complexity did not stop the Court there. Nor has the lack of familiarity with technical subject matter prevented it from articulating an extensive privacy framework in *Puttaswamy,* or from regulating hazardous industry in *M.C. Mehta.* The real issue here is not whether the Court should take over, but whether its claim of institutional incompetence holds up against its own established record.

5 WHAT RUNS WHILE THE COURTS WAIT

The Court in *Goswami* treated the core issue of algorithmic governance as a policy issue to be considered by the executive in due course. But this treatment assumes that there is time. There is not. Automated

[12] The Constitution of India, art. 141, provides that the law declared by the Supreme Court shall be binding on all courts within the territory of India. Article 142 empowers the Court to pass such decree or order as is necessary for doing complete justice in any cause or matter pending before it. Together, these provisions furnish the basis on which the Court has framed binding interim guidelines in the absence of legislation.

[13] *Narendra Kumar Goswami AI Regulation Case: Supreme Court Declines, Asks Centre to Consider Representation*, *supra* note 7.

[14] M.C. Mehta v. Union of India, (1987) 1 SCC 395.

[15] Modern Dental Coll. & Rsch. Ctr. v. State of Madhya Pradesh, (2016) 7 SCC 353; K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1.

systems are already making exclusionary decisions across Indian welfare administration, and the harms are documented rather than anticipated. However, this problem is not India-specific. A global transition has been observed and what the United Nations has termed "digital welfare states" has made the tension between automated government decision-making and rule-of-law principles a subject of international academic attention.[16]

5.1 Automation in Indian Welfare Delivery

The operation of these automated systems runs across multiple welfare sectors through standardised patterns. In Telangana, the Samagra Vedika platform has, since 2016, cross-matched citizens across up to thirty government databases to build a comprehensive 360-degree profile and flag eligibility, resulting in wrongful exclusions, including a widow denied benefits because the system attributed a vehicle to her deceased husband.[17] In Haryana, the Parivar Pehchan Patra works as a primary platform for state welfare schemes, where documented failures include residents wrongly recorded as deceased and losing pensions, alongside auto-populated income figures determining eligibility without input from citizens.[18] A third instance arises under the Targeted Public Distribution System (Control) Amendment Order, 2025, which mandates periodic e-KYC for all households covered by the National Food Security Act and provides for deletion of those found ineligible.[19] Where authentication is not completed, names face removal, and authentication fails at the point of sale where fingerprints do not match. This causes serious problems for the elderly and manual labourers. Another important example of automated deletions was seen under the since repealed MGNREGA framework (before its replacement by VB-G RAM G in December 2025) where roughly 27 lakh workers were reported removed from the database during the shift to Aadhaar-based payment and app-based attendance verification.[20]

5.2 The Inversion of Burden of Proof

What these systems have in common is not that automation in administration has failed, but that it has placed the citizen on the wrong side of the failure. The system-generated output (automated decision) is presumed to be correct. The burden lies entirely on the wronged citizen to first discover the exclusion, then to identify its cause and finally disprove it. All of this without access to the reasoning behind the decision, and against data the citizen has never seen. This is precisely the issue identified by the technological due process literature almost two decades ago. As demonstrated by Danielle Keats Citron, automated administrative systems erode the foundational protections of a hearing. They dilute the

---

[16] Sonja Bekker, *Fundamental Rights in Digital Welfare States: The Case of SyRI in the Netherlands*, *in* 50 NETHERLANDS YEARBOOK OF INTERNATIONAL LAW 2019 289 (Otto Spijkers, Wouter G. Werner & Ramses A. Wessel eds., 2020); Monika Zalnieriute, Lyria Bennett Moses & George Williams, *The Rule of Law and Automation of Government Decision-Making*, 82 MOD. L. REV. 425 (2019).

[17] Tapasya, Kumar Sambhav & Divij Joshi, *How an Algorithm Denied Food to Thousands of Poor in India's Telangana*, AL JAZEERA (Jan. 24, 2024), https://www.aljazeera.com/economy/2024/1/24/how-an-algorithm-denied-food-to-thousands-of-poor-in-indias-telangana.

[18] Kumar Sambhav, Tapasya & Divij Joshi, *In India, an Algorithm Declares Them Dead; They Have to Prove They're Alive*, AL JAZEERA (Jan. 25, 2024), https://www.aljazeera.com/economy/2024/1/25/in-india-an-algorithm-declares-them-dead-they-have-to-prove-theyre.

[19] Targeted Public Distribution System (Control) Amendment Order, 2025, notified by the Ministry of Consumer Affairs, Food and Public Distribution on July 22, 2025, amending the Targeted Public Distribution System (Control) Order, 2015; *see also* Sameet Panda & Sweta Dash, *Infrastructures of Exclusion: How e-KYC Impacts Access to Food*, INDIA DEVELOPMENT REVIEW (Aug. 12, 2025), https://idronline.org/article/rights/infrastructures-of-exclusion-how-e-kyc-impacts-access-to-food/.

[20] *27 Lakh MGNREGA Workers Deleted Between October 10 and November 14: Report*, SCROLL.IN (Nov. 17, 2025), https://scroll.in/latest/1088613/27-lakh-mgnrega-workers-deleted-between-october-10-and-november-14-report. The figures originate from Lib Tech India and were first reported by The Hindu. MGNREGA was subsequently replaced by the Viksit Bharat–Guarantee for Rozgar and Ajeevika Mission (Gramin) framework in December 2025.

requirement of notice and, most importantly, promote the idea among adjudicating officers that machine outputs are correct.[21] Both tendencies are visible in the Indian examples above. In Telangana, officials were required to consult the algorithm's output before recording their own decision, but in practice they generally deferred to it.[22]

It was these automated systems that the petition in *Goswami* asked the Union to disclose. But the Court's response leaves these systems running, without the articulation of any safeguards against which they might be tested. Judicial restraint may have merit in theory, but the practical cost is paid by the most vulnerable citizens.

6 A COURT THAT DID ENGAGE: SYRI

6.1 The SyRI System and the Judgment of the District Court of The Hague

The Netherlands operated Systeem Risico Indicatie (SyRI), a statutory mechanism that linked data held across various government departments to generate automated risk scores identifying individuals considered likely to commit social security fraud.[23] SyRI closely resembles the Indian welfare systems discussed above. Both involve cross-departmental data linkage, risk profiling of welfare recipients, lack of prior notice, and outputs generated by algorithmic logic.

A coalition of civil society organisations challenged the legislation that authorised SyRI in *NJCM c.s. v. The State of the Netherlands* before the District Court of The Hague.[24] The Court decided against the legislation, holding it incompatible with Article 8[25] of the European Convention on Human Rights. It was further held that the system failed the proportionality assessment under Article 8 for two reasons: the absence of transparency about the system's logic, and the inadequacy of safeguards for those profiled.[26] The Court also drew on principles from the General Data Protection Regulation as an interpretative aid, and received an *amicus curiae* submission from the UN Special Rapporteur on extreme poverty and human rights.[27]

6.2 Ordinary Procedural Tools with Non-Technical Expertise

It is noteworthy that this Hague court possessed no greater technical expertise in automated systems than the Supreme Court of India. It did not go into the details of the system's underlying code, nor did it require the help of technical experts. What this shows is that reviewing an automated system does not require specialised technical expertise.[28] The reach of this judgment should not, however, be overstated. As scholars note, the decision was limited to the circumstances of the case and says relatively little about

---

[21] Citron, *supra* note 5, at 1255–60.

[22] Tapasya, Sambhav & Joshi, *supra* note 17.

[23] Marvin van Bekkum & Frederik Zuiderveen Borgesius, *Digital Welfare Fraud Detection and the Dutch SyRI Judgment*, 23 EUR. J. SOC. SEC. 323 (2021).

[24] NJCM c.s. v. State of the Netherlands, ECLI:NL:RBDHA:2020:865 (District Court of The Hague Feb. 5, 2020).

[25] Convention for the Protection of Human Rights and Fundamental Freedoms (European Convention on Human Rights) art. 8, 1950. Article 8(1) guarantees the right to respect for private and family life, home and correspondence. Article 8(2) permits interference by a public authority only where it is in accordance with law and necessary in a democratic society for one of the specified aims, including the economic well-being of the country and the prevention of disorder or crime.

[26] van Bekkum & Zuiderveen Borgesius, *supra* note 23.

[27] *Supra* note 16.

[28] *See generally* Jennifer Cobbe, *Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making*, 39 LEGAL STUD. 636 (2019).

automated fraud detection in general.[29] Its immediate practical effects on administrative practice were limited. Furthermore, the specific European rights framework is different from the existing Indian framework. Yet the point drawn here is narrower. The question before a constitutional court is not how the algorithm mathematically works, but whether its use by the state can be justified. What this Dutch decision demonstrates is the willingness of a court to take up the matter for review rather than treating it as beyond judicial competence. This willingness is in sharp contrast to the judicial disengagement shown by the Supreme Court of India in *Goswami*.

6.3 Limits of SyRI and the Legislative Comparison

Europe also has a legislative safeguard in the form of Article 22 of the GDPR, which India lacks. That provision governs decisions based solely on automated processing.[30] However, its scope has been questioned by scholars. It remains contested whether Article 22 establishes a genuine right to an explanation or merely provides a limited right to be informed about the logic involved. This ambiguity, surrounding even the finest available statutory model, strengthens the case for judicially articulated standards, particularly in the Indian case where there is no equivalent provision available in the Digital Personal Data Protection Act, 2023. In the absence of both legislative and constitutional safeguards, the automation of state systems runs a risk of systemic harms. This is well documented in the case of Australia's automated Robodebt scheme, where automated income averaging continued unchecked until it was challenged in the Federal Court, with a Royal Commission subsequently finding the scheme to have been unlawful.[31]

7 THE TEST ALREADY IN HAND

The proportionality test formulated in *Modern Dental College and Research Centre v. State of Madhya Pradesh* and adopted in Puttaswamy was not designed merely to protect data privacy. It is a general constitutional safeguard that the Court now routinely applies to evaluate any state action limiting fundamental rights. The test relies on four sequential questions: whether the measure has a legitimate aim, whether there is a rational nexus between this state measure and that aim, whether a less restrictive alternative exists, and whether the infringement of the right is balanced against the public benefit gained.[32] But this proportionality-based analysis is not unique to Indian jurisprudence. It has become a standard technique across modern constitutional systems for weighing the public interest served against the State's interference with basic rights.[33]

---

[29] van Bekkum & Zuiderveen Borgesius, *supra* note 23.

[30] General Data Protection Regulation (Regulation (EU) 2016/679), art. 22. The primary text of Article 22(1) reads: "The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." For the debate on the limits of its scope, *see* Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7 INT'L DATA PRIV. L. 76 (2017).

[31] *See* Prygodicz v. Commonwealth of Australia (No 2) [2021] FCA 634; REPORT OF THE ROYAL COMMISSION INTO THE ROBODEBT SCHEME (July 2023).

[32] Modern Dental Coll. & Rsch. Ctr. v. State of Madhya Pradesh, (2016) 7 SCC 353; K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1, ¶ 325.

[33] *See* Alec Stone Sweet & Jud Mathews, *Proportionality Balancing and Global Constitutionalism*, 47 COLUM. J. TRANSNAT'L L. 72 (2008); *see also* Vicki C. Jackson, *Constitutional Law in an Age of Proportionality*, 124 YALE L.J. 3094 (2015).

7.1 Applying the Test to an Automated Exclusion

When this test is applied to automated welfare exclusions, its utility becomes clear. Where an automated system deletes a citizen from a ration database or cancels a pension, the court can evaluate the state action using the four prongs of proportionality:

Legality: is the automation in such systems authorised by a statute? And if in fact a valid statute exists, does it explicitly authorise the deletion of names or cancellation of welfare entitlements via algorithmic processing?

Legitimate aim and rational nexus: is the algorithmic processing serving a legitimate purpose, such as efficient, cost-effective delivery of welfare benefits without leakage? Is there a rational nexus between the state's use of such automation and the aim it pursues?

Necessity: was the use of such a measure necessary or was there a less restrictive measure available to achieve the state's aim?[34]

Balancing: does the administrative gain in efficiency or cost saving justify the extent of harm caused to citizens?

None of these four prongs requires the court to inquire into the technical workings of code or machine learning. Asking questions through each prong can simplify the approach to such complex problems. The claim of institutional incompetence therefore cannot survive. The Court by declining to apply an established standard showed judicial disengagement over an issue which it is fully equipped to handle.

7.2 The Limits of Human Oversight

At the same time, the Court should avoid treating the 'human in the loop' in such systems as an adequate safeguard. In practice, this works as an empty formality. Recent scholarship warns that a human supervisor or an official usually rubber-stamps the same machine output rather than reviewing it on a case-by-case basis.[35] While applying the proportionality test, the court must take this into account before accepting the State's argument that a human remains in the loop. *Vishaka* supplies the form: binding interim safeguards holding the field until Parliament acts. *Puttaswamy* supplies the content: a standard to put inside them. The Court held both and reached for neither.

8 Conclusion

The growing reliance on algorithmic systems in state governance has revealed an accountability gap. This gap has dual dimensions, one legislative and the other judicial. The legislative gap lies in the absence of any statutory provision in the Indian law equivalent to Article 22 of the GDPR. This paper has examined the second dimension of this gap: the judicial reluctance as displayed in *Goswami*. The Court possessed both a compelling precedent for its interim intervention and a doctrine to supply its content, yet it chose to disengage citing grounds of policy consideration. In the absence of a proper legislative framework, the Court's intervention matters more, not less.

---

[34] A less restrictive alternative could be one which could have flagged the anomaly in its processing, creating a requirement for human intervention or marked the anomalous step instead of enforcing an automatic removal without a notice.

[35] Riikka Koulu, *Proceduralizing Control and Discretion: Human Oversight in Artificial Intelligence Policy*, 27 MAASTRICHT J. EUR. & COMP. L. 720, 725–27 (2020).

The analysis in this paper proposes something narrower. It does not seek to create new fundamental rights to deal with rising state automation. Nor does it call for judicial innovation of new safeguards to deal with it. What it proposes is the application of an existing constitutional test to the given set of problems arising from state action that is escaping scrutiny altogether. However, there is a crucial precondition to applying this test: basic disclosure of the existence of automation in state systems. A constitutional court cannot assess the proportionality of an automated system when its use and operation remain undisclosed. The disclosure affidavit requested in *Goswami* was not a remedy in itself, rather a precondition to get the actual remedy.

The approach proposed here follows the logic the Court itself has applied every time it issued interim guidelines pending legislation. The judicial guardrails hold the ground until the legislative vacuum is filled. But currently the cost of waiting for Parliament to act is being borne by people already unjustly affected by the automated system. These people are also least equipped to deal with their exclusions by litigation. For them, the interval between judicial disengagement and legislative action is not an abstraction. It is a matter of daily survival.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.5

CHAPTER 5

The Right to Digital Access as an Emerging Facet of Article 21

BHARATHA VASAN G[*] AND HEMASURUTHI A.S.[**]

In India, access to the internet and digital infrastructure has moved from being a matter of convenience to becoming a necessary condition for exercising basic civil, economic and social rights. Education, banking, healthcare, employment and participation in governance are now carried out largely through digital platforms, yet no settled constitutional principle treats digital access as a distinct fundamental right. This paper carries out a doctrinal study of whether the right to digital access can be read as an evolving part of Article 21 of the Constitution of India, against the backdrop of the wider growth of the right to personal liberty from a narrow procedural safeguard into a substantive right covering dignity, livelihood and effective participation in public life. It traces the reasoning by which courts have gradually widened Article 21 to respond to new socio-economic conditions and asks whether digital connectivity fits within this line of interpretation as a facilitating right, one that enables the exercise of other constitutional rights rather than standing as a right in itself. It also studies whether existing statutory and regulatory measures are adequate to meet gaps in connectivity, affordability and urban-rural disparity, and whether these measures amount to real rights protection or remain matters of policy discretion. Comparable decisions from other countries are examined to identify workable doctrinal tests for recognising digital access as a constitutionally protected interest. The paper closes with an interpretive framework for reading digital access into Article 21, while cautioning against an unqualified extension that could weaken the rigour needed when courts identify new fundamental rights.

Keywords: Article 21 - Digital Access - Right to the Internet - Constitutional Interpretation - Digital Divide.

1 INTRODUCTION

The internet has changed the way people in India live, work, learn and talk with the government. A student watches lessons on a phone screen, a farmer looks up the price of grain using an app and a worker gets paid through a bank account that is connected to a digital identity and a person applies for a ration card or a court date without leaving the house. Something that was once a helpful tool has become, for many reasons, something that people need. The Constitution of India does not talk about the internet, online platforms or a right to be connected. The word digital is not in the document that was written in the 1940s.

That gap between what the Constitution says and what people do every day is not new. Article 21 of the Constitution protects the right to life and personal liberty. Article 21 of the Constitution used to guard against unfair detention. But now Article 21 of the Constitution covers many more things such as livelihood, health, shelter, clean air, privacy and learning. Courts added those things to Article 21 of the Constitution, because without them people could not live a life of dignity. I wonder whether digital access

[*] Student at The Tamil Nadu Dr. Ambedkar Law University, School of Excellence in Law, Chennai, Tamil Nadu, India.
[**] Student at The Tamil Nadu Dr. Ambedkar Law University, School of Excellence in Law, Chennai, Tamil Nadu, India.

deserves this treatment. Can courts or Parliament treat the internet and digital public services as a part of Article 21 of the Constitution just as they added livelihood or education to Article 21 of the Constitution?

This paper is doctrinal in nature and uses a method of looking at case laws, statutes and materials from other legal systems instead of gathering empirical data. Part 2 follows how Article 21 evolved from a procedural protection into a substantive right. Part 3 explains why digital access should be seen as a facilitating right and not an entitlement. Part 4 reviews case laws that have already dealt with internet access. Part 5 analyzes the policy framework in India that governs digital connectivity and highlights missing elements. Part 6 considers developments in other countries. Part 7 suggests a test for recognizing digital access under Article 21 and Part 8 proposes reforms before the paper ends.

2 FROM PROCEDURE TO SUBSTANCE: THE GROWTH OF ARTICLE 21

When the Constitution was adopted, the Supreme Court interpreted Article 21 in a specific way. In the case of A.K. Gopalan v. State of Madras[1] the Court said that a law taking away a person's liberty only needed to follow a procedure set by the legislature. The Court did not care whether that procedure was fair or just. At that time Article 21 was seen as a guarantee completely separate from Articles 14 and 19.

This view changed later in the case of Maneka Gandhi v. Union of India.[2] In this judgment the Court said that the procedure used to take life or personal liberty must be fair, just and reasonable. It also said that Articles 14, 19 and 21 are not separate parts of the Constitution but are linked together. These three articles form one connected system of rights protection called the golden triangle. After Maneka Gandhi's case, the words "life" and "personal liberty" were no longer seen as being only about physical survival. They came to include everything that makes life meaningful and worth living with dignity.

Once this door opened the Court began to interpret Article 21 in a way that included a number of rights. These rights were seen as extensions of the right to life. In the case of Francis Coralie Mullin v. Administrator, Union Territory of Delhi[3] the Court said that the right to life is not just about staying alive. It also includes living with dignity. This means people have the right to things like good food, proper clothes, safe shelter and the ability to express themselves freely.

Then in Olga Tellis v. Bombay Municipal Corporation[4] the Court took this further. It said that the right to life includes the right to livelihood. The reason was simple: if someone loses their ability to earn a living, it means they are losing the basis of life itself. Without a means to support themselves a person cannot truly live.

Over time the Court continued to expand what Article 21 covers. It added the right to a healthy environment.[5] It recognized the right to shelter as essential.[6] It also included the right to fast medical

[1] A.K. Gopalan v. State of Madras, AIR 1950 SC 27. The literal "procedure established by law" standard adopted here was effectively displaced by *Maneka Gandhi v. Union of India*, (1978) 1 SCC 248. *See generally* GRANVILLE AUSTIN, WORKING A DEMOCRATIC CONSTITUTION: A HISTORY OF THE INDIAN EXPERIENCE (1999).

[2] On the "golden triangle" formed by Articles 14, 19 and 21, *see* UPENDRA BAXI, COURAGE, CRAFT AND CONTENTION: THE INDIAN SUPREME COURT IN THE EIGHTIES (1985).

[3] Francis Coralie Mullin v. Adm'r, Union Territory of Delhi, (1981) 1 SCC 608.

[4] Olga Tellis v. Bombay Mun. Corp., (1985) 3 SCC 545.

[5] Subhash Kumar v. State of Bihar, (1991) 1 SCC 598; M.C. Mehta v. Union of India, (1987) 1 SCC 395 (the Ganga pollution case).

[6] Chameli Singh v. State of U.P., (1996) 2 SCC 549.

treatment when someone is in an emergency.[7] Each of these decisions built on the idea that the right to life is not just about existence. It is about living with dignity, security and the ability to thrive.

I think education is very close to the question. Before the Eighty-Sixth Amendment the Court in Unni Krishnan, J.P. v. State of Andhra Pradesh[8] held that the right to free and compulsory education for children up to fourteen years comes from Article 21 together with the Directive Principle in Article 45. Later Parliament gave this judicial recognition a textual basis by inserting Article 21A into the Constitution in 2002 and by making the Right of Children to Free and Compulsory Education Act in 2009. The sequence is instructive: a social and economic need was first seen by courts as hidden in Article 21 and later turned into an explicit constitutional and statutory promise.

The right to privacy followed a similar path. In the case of Justice K.S. Puttaswamy v. Union of India[9] a nine-judge bench ruled that privacy is a part of Article 21. The court described privacy as a right that supports freedom and allows people to make personal choices about their own lives. This judgment matters for the paper because it shows that the Court is ready to recognize a right that does not appear clearly in Part III of the Constitution. The Court does so when it believes the right is necessary for liberty and dignity in today's world.

3 Digital Access as a Facilitating Right

Not every value recognized under Article 21 stands alone. Some rights, like the right against torture, hold importance on their own.[10] They matter simply because they are essential. Other rights matter more for what they enable. Livelihood matters because it helps a person provide food and shelter for a family. Education matters because it leads to jobs, allows self-expression and supports informed citizenship. Digital access falls into this group. Few people care about an internet connection just for the connection itself. They care about what they can do with it.

Today that list of things is long. School and college classes moved online during the COVID-19 pandemic and many continue in that form. Welfare payments under schemes such as PM-KISAN and old-age pensions are transferred directly into bank accounts and tracked through digital ledgers. The public distribution system across States has moved to digital ration cards linked to biometric authentication. Health services increasingly rely on telemedicine in areas with few doctors. Courts have adopted e-filing and virtual hearings. Even participation in governance from filing a Right to Information application to registering a grievance is migrating online. Digital access, in other words, has become the pipe through which many other Article 21 and Article 19 rights now flow.[11]

---

[7] Paschim Banga Khet Mazdoor Samity v. State of West Bengal, (1996) 4 SCC 37; Paramanand Katara v. Union of India, (1989) 4 SCC 286.

[8] Unni Krishnan, J.P. v. State of Andhra Pradesh, (1993) 1 SCC 645. The Constitution (Eighty-sixth Amendment) Act, 2002 inserted Article 21A; it was brought into force with effect from 1 April 2010, alongside the Right of Children to Free and Compulsory Education Act, 2009.

[9] Justice K.S. Puttaswamy v. Union of India, (2017) 10 SCC 1. The nine-judge bench in *Puttaswamy* overruled the contrary observations in *M.P. Sharma v. Satish Chandra*, AIR 1954 SC 300, and *Kharak Singh v. State of U.P.*, AIR 1963 SC 1295, to the extent that those decisions held there is no fundamental right to privacy under the Constitution.

[10] D.K. Basu v. State of West Bengal, (1997) 1 SCC 416, recognising procedural safeguards against custodial torture as part of Article 21.

[11] On the digitisation of Indian welfare delivery generally, *see* ANITA GURUMURTHY & DEEPTI BHARTHUR, VOICE OR CHATTER? DECONSTRUCTING THE MYTH OF DIGITAL DEMOCRACY (2016).

This instrumental character explains why digital access should at least for now be seen as a secondary right instead of a standalone one. A standalone right would mean courts have to say that internet access is valuable by itself no matter how it is used. That kind of claim is hard to defend. On the other hand, a facilitating right only asks whether denying access actually blocks another right that is already protected by the Constitution. That could be the right to education, the right to work, the right to health, the right to privacy or the right to free speech. This is the logic that led the Kerala High Court as explained in the next section to connect internet access to the right to education and the right to privacy, rather than treating it as a separate right on its own.

4 JUDICIAL RECOGNITION OF DIGITAL ACCESS IN INDIA

Indian courts have not yet declared a freestanding right to access under Article 21. However two decisions have pushed the law forward considerably in that direction.

The first is Faheema Shirin R.K. v. State of Kerala[12] decided by a judge of the Kerala High Court. A student living in a college hostel challenged a rule restricting the use of phones during certain hours arguing that the rule cut off the student's access to online study material. The Court held that the restriction violated the student's right to privacy and the student's right to education both traced to Article 21 and the student's freedom of speech and expression under Article 19(1)(a). In reaching this conclusion, the Court relied on a resolution of the United Nations Human Rights Council recognising access to the internet as a freedom and a tool for realising the right to education and invoked the principle, drawn from Vishaka v. State of Rajasthan,[13] that international conventions and norms may be read into the rights guaranteed by Part III when domestic law is silent on the point. Faheema Shirin is by far the clearest judicial statement in India that access to the internet is not simply a policy convenience but touches upon rights guaranteed by the Constitution.

The second case is Anuradha Bhasin v. Union of India.[14] This case came about due to the long-term suspension of internet and mobile services in Jammu and Kashmir after the reorganisation of the State in August 2019. The Supreme Court said that the freedom of speech and expression under Article 19(1)(a) and the freedom to carry on any trade, business, profession or occupation under Article 19(1)(g) include activities done over the internet. The Court did not say there is a right to internet access. It did say that any order to cut off internet services must pass the tests of necessity and proportionality. Such orders must be temporary and must be made public so that they can be challenged in court. This judgment is important because it sees the internet as a tool through which existing constitutional rights are used. It applies standards of reasonableness to restrictions on the internet. It does not treat internet shutdowns as something outside the reach of review.

Read together these two decisions show a pattern rather than a settled doctrine. I see that courts in India have so far protected digital access indirectly by treating it as the way to exercise rights such as education, privacy, speech and trade. Neither decision however forces the State to give internet access to

[12] Faheema Shirin R.K. v. State of Kerala, W.P.(C) No. 19716 of 2019 (Ker. HC Sept. 19, 2019). *Faheema Shirin* was decided by a single judge of the Kerala High Court and, as of writing, has not been reviewed by a Division Bench or the Supreme Court.

[13] *Vishaka v. State of Rajasthan*, AIR 1997 SC 3011, itself drew on the Convention on the Elimination of All Forms of Discrimination Against Women (CEDAW) (1979) to fill a legislative vacuum on workplace sexual harassment.

[14] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637. *See also* Found. for Media Pros. v. Union Territory of Jammu & Kashmir, (2020) 5 SCC 746, applying the *Anuradha Bhasin* standard to later internet-restriction orders in Jammu and Kashmir.

people who do not have it; both focus on lifting limits that already exist. This leaves a gap for many people, discussed in the next part, who face costs and lack of infrastructure instead of limits imposed by the State.

5 STATUTORY AND REGULATORY FRAMEWORK IN INDIA

India's approach to connectivity has mostly been shaped by laws and government policies rather than by a constitutional right. The Indian Telegraph Act, 1885 was the law to set up the Universal Service Obligation Fund.[15] This fund was later renamed as the Digital Bharat Nidhi under the Telecommunications Act, 2023. The Telecommunications Act, 2023 (Act No. 44 of 2023) replaced the Indian Telegraph Act, 1885 and established the Digital Bharat Nidhi in place of the old fund. This fund collects money from telecom companies and uses it to support the expansion of telecom and broadband services in rural and remote areas. These areas are often unprofitable for private companies to serve on their own.

The Telecom Regulatory Authority of India Act, 1997 set up a regulator.[16] This regulator can set tariffs and monitor how well telecom services are performing. The Information Technology Act, 2000 gives the framework for electronic transactions, digital signatures and dealing with cybercrime.[17] The Information Technology Act, 2000 does not guarantee that every person has the right to access digital services.

On the policy side, the Digital India programme launched in 2015 set out three aims which are digital infrastructure as a core utility for every citizen, governance and services on demand and digital empowerment of citizens.[18] The BharatNet project seeks to connect every gram panchayat in the country with fibre and the PM-WANI scheme allows small shopkeepers and other local providers to set up public Wi-Fi hotspots without a licence in an effort to bring down the cost of last-mile connectivity.[19] These schemes are useful and, in some places, effective, yet they remain matters of executive discretion. Their pace, funding and reach can be altered by decisions of the government of the day and a citizen denied connectivity under these schemes has no clear individual entitlement that a court can enforce, unlike for instance a child denied admission under the Right to Education Act, 2009.

The conflict between rules and basic rights is most clear when it comes to giving help through the internet. The Aadhaar law from 2016 makes it so that people need to use their fingerprint or face to get food, gas, old-age money and other help. In a court case called Justice K.S. Puttaswamy v. Union of India the Court upheld most of the Aadhaar Act, while striking down certain provisions.[20] The Court also said that no one should be left out of basic help like food just because the digital system isn't working. This warning is important for the present discussion. It shows that when the internet controls help, leaving people out because of bad connections can break a basic right even without a new rule being made.

---

[15] The Universal Service Obligation Fund was established through the Indian Telegraph (Amendment) Act, 2003, which inserted Sections 9-B and 9-C into the Indian Telegraph Act, 1885.

[16] The Telecom Regulatory Authority of India Act, 1997.

[17] The Information Technology Act, 2000.

[18] MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY, GOVERNMENT OF INDIA, DIGITAL INDIA PROGRAMME (2015).

[19] *See* DEPARTMENT OF TELECOMMUNICATIONS, GOVERNMENT OF INDIA, GUIDELINES FOR PUBLIC WI-FI ACCESS NETWORK INTERFACE (PM-WANI) (2020); Bharat Broadband Network Limited, *BharatNet Project* (overview).

[20] Justice K.S. Puttaswamy v. Union of India, (2019) 1 SCC 1 (the Aadhaar judgment), upholding the Aadhaar Act, 2016 with certain provisions struck down.

Despite these efforts large gaps still exist. Data from the Telecom Regulatory Authority of India and the National Family Health Survey show again and again that women use the internet less than men.[21] There is also a divide between urban and rural areas when it comes to internet access and cost. These differences are not just numbers on a page. They decide who can get the benefits that come from services like education, jobs and health support. When policies rely on subsidies and choices without a solid foundation of rights the people who need help the most may never get it. The system could leave the vulnerable behind for good.

6 INTERNATIONAL AND COMPARATIVE PERSPECTIVE

India is not the only country facing this question because international and comparative materials also give useful guidance. Article 19 of the Universal Declaration of Human Rights protects the freedom to seek, receive and share information through any media no matter the borders and this language easily extends to the internet.[22] Article 19(2) of the International Covenant on Civil and Political Rights, to which India is a party, repeats this guarantee in a binding treaty.[23] Building on these provisions the United Nations Human Rights Council in its 2016 resolution about the promotion, protection and enjoyment of rights on the internet said that the rights people have offline must also be protected online. The United Nations Human Rights Council also condemned measures that purposely block access to or the spread of information online.[24] This resolution was the one the Kerala High Court mentioned in the case of Faheema Shirin.

I think France gives an example of how courts can protect rights through normal constitutional review. In its ruling on the HADOPI law the Conseil Constitutionnel struck down a rule that allowed an administrative body to cut a person's internet access as punishment for suspected copyright theft, all without a court. The Conseil said that because the internet is now a part of democratic life and idea sharing, the right to use online communication services is part of the freedom to communicate and express oneself as protected by Article 11 of the 1789 Declaration of the Rights of Man. The Conseil added that only a judge may restrict that freedom.[25]

Costa Rica provides an example of a court taking a bold step and viewing internet access as essential for full participation in contemporary life. In its 2010 decision the Sala Constitucional, the country's constitutional chamber, declared that access to information and communication technology, internet access included, is a fundamental tool that lets citizens exercise their basic rights. The decision also asked the telecommunications regulator to set a functional internet access speed and to review that speed regularly as technology improves.[26] By treating internet access not as a side effect of other rights but as a

---

[21] INTERNATIONAL INSTITUTE FOR POPULATION SCIENCES, NATIONAL FAMILY HEALTH SURVEY-5, 2019–21 (2021); TELECOM REGULATORY AUTHORITY OF INDIA, INDIAN TELECOM SERVICES PERFORMANCE INDICATOR REPORTS.

[22] Universal Declaration of Human Rights art. 19, G.A. Res. 217A (III), U.N. Doc. A/810 (1948).

[23] International Covenant on Civil and Political Rights art. 19(2), 1966, 999 U.N.T.S. 171.

[24] U.N. Human Rights Council, *The Promotion, Protection and Enjoyment of Human Rights on the Internet*, U.N. Doc. A/HRC/32/L.20 (June 27, 2016). This built on the earlier report of the UN Special Rapporteur on freedom of opinion and expression: Frank La Rue, *Report on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/17/27 (May 16, 2011), among the first international documents to frame internet access chiefly as instrumental to other rights.

[25] Conseil Constitutionnel, Decision No. 2009-580 DC (June 10, 2009) (Loi favorisant la diffusion et la protection de la création sur internet). For commentary on the decision, *see* Nicola Lucchi, *Access to Network Services and Protection of Constitutional Rights: Recognizing the Essential Role of Internet Access for the Freedom of Expression*, 19 CARDOZO J. INT'L & COMP. L. 645 (2011), discussing the Conseil's reliance on Article 11 of the 1789 Declaration of the Rights of Man.

[26] Sala Constitucional de la Corte Suprema de Justicia de Costa Rica, Resolución No. 2010-012790 (2010).

public service that the State must keep at a high enough level this ruling places internet access closer to a standalone right than do the Indian or French approaches.

A different route is visible in Estonia and in the European Union. The guarantee of connectivity came through legislation instead of constitutional adjudication. Estonia amended its Telecommunications Act in 2000 to declare internet access a requirement for daily life. The amendment also directed libraries to provide free internet access, an early and often cited example of treating connectivity as a public utility.[27] The European Union's Universal Service Directive requires member States to ensure that all consumers have access to an internet connection of a quality sufficient for ordinary use. The Directive backs this requirement with consumer protection remedies, not with a right.[28] These examples show that a right of access can be secured either by reading it into existing guarantees as France and Costa Rica have done or by building it directly into statute as a universal service standard as Estonia and the European Union have done. India as Part 5 showed currently relies on the statutory route through the Digital Bharat Nidhi and related schemes but India lacks the enforceable minimum standards found in the European model.

7 Towards a Doctrinal Test for Recognising Digital Access under Article 21

This paper argues that digital access should be considered under Article 21 like other unlisted rights that Indian courts have recognized. The idea is to look at access through four related questions. The first one is a necessity test. Taking inspiration from the case of Francis Coralie Mullin the question is whether a person can live with human dignity without the facility in question. In today's world, where so much of education, support and work depends on being able to use tools there is a strong argument that for many people not having digital access already fails to meet this standard.

Second, a facilitative-instrumental test asks whether the claimed right mainly works as a means to exercise another right already recognised in the Constitution, such as education, livelihood, health, privacy or free speech, instead of standing as a value on its own. Part 3 explained that digital access meets this test more naturally than it would meet a claim to an independent free-standing status.

Third, a textual anchoring test asks whether the claimed right can be tied to language already present in the Constitution, including the Directive Principles of State Policy. Article 21A guarantees education and Articles 38, 39 and 41 direct the State to secure a means of livelihood and to make effective provision for work and public assistance. Digital access can be tied to each of these provisions as the medium through which they are increasingly delivered in the same way that the right to education was tied to Article 45 before the Eighty-Sixth Amendment gave it independent constitutional text.

Fourth, an international-law test asks whether international law supports the claimed right, as Article 51(c) of the Constitution tells the State to honour international law and the Vishaka principle lets courts read international conventions and resolutions into Part III rights when domestic law says nothing. The 2016 Human Rights Council resolution, which I talked about in Part 6 along with Article 19 of the International Covenant on Civil and Political Rights, gives Indian courts legal material to use, just like the Kerala High Court used it in Faheema Shirin.

---

[27] For a sceptical response to treating internet access as a right in itself, *see* Vint Cerf, *Internet Access Is Not a Human Right*, N.Y. TIMES (Jan. 5, 2012).

[28] Directive 2002/22/EC of the European Parliament and of the Council of 7 March 2002 on universal service and users' rights relating to electronic communications networks and services (Universal Service Directive), as amended by Directive 2009/136/EC.

Applying these four tests together suggests that digital access under Article 21 should be recognized in stages, not all at once. At the first stage, courts should continue to see digital access as a derivative right. This means it should be linked to an underlying right like education, privacy, livelihood or speech as the Kerala High Court did. It should not be declared as a broad freestanding right to internet access for every purpose. This approach respects the caution shown by the Supreme Court in the Puttaswamy case. That case made it clear that not every claim based on convenience or personal preference should be turned into a right. It also keeps recognition tied to real concrete facts rather than general policy ideas.

At the second stage, Parliament should turn this judicial recognition into a clear enforceable legal right. This would be similar to how the Right of Children to Free and Compulsory Education Act 2009 gave a practical form to the right recognized in the Unni Krishnan case. It makes the right more than a legal idea; it becomes something people can actually rely on.

At the third stage, any restriction the State imposes on digital access that already exists such as an internet shutdown must still go through the test of necessity and proportionality. This test was laid out in the Anuradha Bhasin case. It gives courts a workable standard to review whether such restrictions are justified.

8 SUGGESTIONS FOR REFORM

Based on the analysis this paper offers the following suggestions, aimed at closing the gap between the constitutional promise of a dignified life and the practical reality of unequal digital access.

First, Parliament should consider a provision whether in the Telecommunications Act, 2023 or in a separate enactment recognising a basic affordable minimum level of internet access as a legal entitlement for every household. This entitlement should be calibrated to India's capacity and modelled loosely on the European Union's universal service standard rather than on more expansive judicial declarations. This would give enforceable content to the derivative right discussed in Part 7 in the same way the Right to Education Act gave content to the right recognised in Unni Krishnan.

Second, the Digital Bharat Nidhi should publish reports at the panchayat level showing how much money has been collected and what connectivity has been delivered. These reports should include time-bound targets for bringing last-mile connections to every village. Now the way things are set up it is very hard for citizens or courts to know if the fund is being used properly. Public reporting would help spot gaps and make it possible to fix them. This could include public interest litigation when targets are missed again and again.

Third, the Telecom Regulatory Authority of India should be given the power to set enforceable standards for internet speed and reliability in rural and underserved areas. These standards should be backed by a consumer remedy, similar to what is used in Costa Rica and under the European Union's Universal Service Directive. This would be much better than relying on general quality-of-service guidelines, which don't give people any real right to take action when things go wrong.

Fourth, welfare schemes that depend on biometric authentication should include a statutory guarantee going beyond the caution expressed in the Aadhaar judgment that no person will be denied food, pension or other basic welfare solely because of a failure of network, device or biometric matching with a simple offline fallback built into the scheme rather than left to administrative discretion.

Fifth, given the gap in internet usage between men and women recorded in the National Family Health Survey, digital literacy programmes under Digital India should set monitored targets for women, older persons and persons with disabilities rather than treating digital literacy as a general undifferentiated goal.

Sixth, the Temporary Suspension of Telecom Services (Public Emergency or Public Safety) Rules, 2017 needs to be changed.[29] The changes should make it necessary to get permission from a court or a similar legal body before stopping internet services for more than a fixed time. This time must be clearly set out. This will give support to the idea of fairness that was already discussed in the case of Anuradha Bhasin. It will also help stop periods where the internet is shut down even after that court decision.

9 CONCLUSION

Article 21 has never stayed the same. It began as a protection against unjust detention in the case of A.K. Gopalan. Over time through landmark rulings like Maneka Gandhi and others that followed it expanded into a guarantee for a life of dignity. This includes rights such as earning a livelihood, staying healthy, having shelter, living in a healthy environment, enjoying privacy and getting an education. Digital access fits into this evolution. It is not important for its own sake. Today more and more of what people need such as education, jobs, welfare services and participation in public life, happens online. For many digital access has become just as essential to a life as the rights already protected under Article 21.

Indian courts have taken the first steps toward this recognition. The Kerala High Court, in Faheema Shirin, linked internet access to the right to privacy and the right to education. The Supreme Court, in Anuradha Bhasin, held that restrictions on internet access must follow the rules of necessity and proportionality. Neither of these decisions makes the State legally required to provide internet access. The laws and policies around Digital Bharat Nidhi, Digital India and similar schemes are still left to the government's choice not to people's rights. Studies from France, Costa Rica, Estonia and the European Union show that this shortfall can be fixed either by courts interpreting the law carefully or by lawmakers making a promise or, usually, by doing both.

This paper suggests that Indian courts should keep recognizing access as a supporting right linked to rights that are already recognized instead of immediately declaring it as a broad independent right. It also recommends that Parliament follow the route taken with the right to education turning judicial recognition into a clear and enforceable law. This step-by-step method would bring the promise of Article 21 into the digital age without losing the careful legal reasoning that has always guided new fundamental rights in India.

* * *

[29] The 2017 Rules were framed under Section 5(2) of the Indian Telegraph Act, 1885, and were the regulatory framework under judicial review in *Anuradha Bhasin v. Union of India*, (2020) 3 SCC 637.

PART II

Regulating Artificial Intelligence: Fairness, Accountability and Liability

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.6

CHAPTER 6

Governance by Design, Accountability by Default? India's Techno-Legal Turn in AI Regulation and the Problem of Contestability

GAUTAMI SETH[*] AND SURABHI KUMARI[**]

India's AI Governance Guidelines (2025) take a light-touch, principle-based path. Instead of a dedicated statute on the model of the EU's Artificial Intelligence Act, they rely on existing law, chiefly the Information Technology Act, 2000 and the Digital Personal Data Protection Act, 2023, and on a "techno-legal" approach that builds legal and policy requirements into the technical architecture of AI systems by design. This chapter argues that the approach is presented as a technical choice but carries constitutional consequences. By moving rule-making from legislatures and courts to code and to the compliance functions of AI developers, it reallocates governing authority, and its reliance on voluntary frameworks shaped by dominant providers risks leaving legal requirements symbolic.

Through a doctrinal and comparative reading of the Guidelines, the IT Amendment Rules 2026, the DPDP Act and the RBI's FREE-AI report against the EU AI Act and Article 22 of the GDPR, the chapter shows that a person affected by an automated decision in India must assemble a remedy from instruments not designed for the purpose: the DPDP Act offers grievance redressal, but no right to contest an automated decision or to receive reasons for it. The chapter concludes that India's principles need clear, legally enforceable rules behind them, with contestability and remedy as a baseline, and that the techno-legal framework, as presently designed, falls short of it.

Keywords: AI governance; techno-legal regulation; governance by design; algorithmic accountability; contestability; due process; self-regulation; India

1 INTRODUCTION

1.1 Overview of the AI Regulation Landscape

Manuel Castells wrote, 'We know that technology does not determine society: it is society'.[1] Increasingly, privately owned web-based platforms control our access to public services, security, education, the public sphere, health services and our very relationship with the countries we live in. As society is 'datafied', public services are delivered through public-private partnerships.[2] AI's potential has grown rapidly, and governance struggles to keep pace. As a result, AI's rapid evolution continues to create new legal and policy challenges. The growing use of AI also requires processing large volumes of personal and sensitive information, including healthcare data, medical records, and financial records. This raises major concerns

[*] Manager, New Initiatives & Research at India Legal Research Foundation, Noida, Uttar Pradesh, India.
[**] Legal Research and Content Associate at India Legal Research Foundation, Noida, Uttar Pradesh, India.
[1] *The Network Society: From Knowledge to Policy*, *in* THE NETWORK SOCIETY: FROM KNOWLEDGE TO POLICY 2 (Manuel Castells & Gustavo Cardoso eds., 2006).
[2] Chinmayi Arun, *AI and the Global South: Designing for Other Worlds*, *in* THE OXFORD HANDBOOK OF ETHICS OF AI 749 (Markus D. Dubber, Frank Pasquale & Sunit Das eds., 2020).

pertaining to privacy, data security, unauthorised access, and accountability. AI could affect rights to freedom of expression, privacy, social security, and protection against discrimination. It may also violate States' obligation to guarantee these rights without discrimination.[3]

1.1.1 Importance of the Topic in the Context of India and Global Trends

AI regulation has developed into a prominent legal and policy issue, given the prospective impact of AI systems on privacy, equality, human autonomy, employment, healthcare, and other fundamental rights. The European Union's Artificial Intelligence Act[4] is a legally binding, risk-based regulatory framework, while India's AI Governance Guidelines[5] adopt a more flexible, principles-based approach that promotes responsible AI development while boosting technological development and innovation. Using the EU's AI Act as a reference, the Guidelines set out a "light-touch," principle-based framework structured around seven guiding "sutras" that relies on existing laws. The "techno-legal" approach intentionally integrates legal instruments and policy regulations directly into the technical architecture of artificial intelligence (AI) systems. This "governance by design" methodology, within a "LAW PLUS" paradigm, is significantly different from the traditional broad legislative regimes. This research analyses the viability of a 'techno-legal' approach, which is 'soft' in nature and appears to be a technical arrangement that significantly reallocates governing authority.

1.1.2 Purpose of the Paper: Critiquing India's Techno-Legal Approach to AI Governance

The paper argues that India's shift toward a techno-legal framework is presented as a technical choice, but it raises questions about the legitimacy of state decision-making. By moving rule-making from courts and legislatures to code and AI developers' compliance functions, it risks impairing essential procedural guarantees. A measured regulatory approach is therefore essential to foster technological innovation while protecting individual privacy and the security of sensitive personal information. Such an approach is fundamental to defending the integrity of democratic rights in an increasingly technological domain.

2 RESEARCH QUESTIONS

1. What is the "techno-legal" model of AI governance in India, and how does its "governance by design" approach differ from traditional command-and-control regulation and the EU's risk-tiered model?
2. How are the legal arguments associated with the technical architecture, and where can remedies for automated decisions be found?

3 DEFINING THE TECHNO-LEGAL MODEL

The techno-legal approach to AI governance can be defined as the amalgamation of legal instruments, rule-based conditioning, regulatory monitoring, and technical enforcement processes embedded in the

[3] *Id.* at 12.

[4] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689) 1.

[5] MINISTRY OF ELECTRONICS AND INFORMATION TECHNOLOGY [MEITY], INDIA AI GOVERNANCE GUIDELINES (Nov. 5, 2025), https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf.

technical architecture by design. In this research article, the authors analyse the intertwined AI governance models, which vary in scope and application as well as the presupposition of the structure of compliance. Some reflect a "governance by audit" paradigm, emphasising post-deployment documentation and oversight. In contrast, others adopt a "governance by design" approach, embedding legal and ethical obligations into the development lifecycle itself.

This paper also distinguishes India's AI Governance Guidelines from the EU's risk-tiered model. Although the Indian guidelines recognise the importance of risk assessment, they do not establish a detailed statutory risk-classification system like the EU AI Act. The EU AI risk assessment law distinguishes AI on the basis of risk creation. Chapter II, Article 5 deals with prohibited AI practices, while Chapter III, Article 6 identifies high-risk AI systems. High-risk systems are then subject to stricter legal requirements. The main problem with India's present AI governance approach is not that it lacks principles of responsible and safe AI. The issue is that clear, legally enforceable rules do not support these principles.

India's present principles-based approach does not provide a complete answer to every such situation. It recommends a graded liability system based on the function performed, the level of risk, and whether proper care was taken. It also recommends clearer responsibilities for different actors in the AI value chain. This paper explores the complications of AI governance and regulatory systems in India, highlighting that ambiguous accountability may give rise to considerable data threats in the country.

4 Methodology

The proposed research adopts a doctrinal and comparative approach, using normative analysis while deliberately excluding empirical fieldwork. Through a comprehensive and technical analysis of India's primary policy and regulatory framework, i.e. the India AI Governance Guidelines, Office of the Principal Scientific Adviser white papers, the IT Amendment Rules 2026, the DPDP Act 2023 and the RBI FREE-AI report (2025), it aims to understand whether the 'governance by design' framework delivers ample security and does not risk the breakdown of essential procedural guarantees. The research article compares it with the European Union's AI Act and the provisions on automated decision-making in the General Data Protection Regulation (GDPR). This analysis showcases the necessary components of a contestability-protective alternative within the context of emerging AI governance.

5 Reallocation of Responsibility in AI Governance – Analysing EU and Indian AI Regulation

5.1 EU AI Act – Binding Legislation

The EU AI Act is detailed legislation to monitor and control the detrimental effects of Artificial Intelligence. The EU AI Act prohibits certain uses of artificial intelligence (AI). Some AI systems are prohibited if they have been created to manipulate an individual's decision-making, creativity and/or exploit vulnerabilities of individuals classified by their social behaviour and characteristics. Furthermore, the Act forbids the processing of facial recognition data and categorisation of individuals through biometric information, except when utilised for legitimate law enforcement objectives, such as locating missing persons or preventing acts of terrorism.[6] Chapter III, Article 6 outlines the classification of high-risk AI systems used

[6] Artificial Intelligence Act, art. 5, 2024 O.J. (L 1689).

as a product safety feature or as products themselves (Annex I). An Annex III system is exempt from high-risk classification only when it does not pose a significant risk to health, safety, or fundamental rights and meets the required criteria. The EU Commission is responsible for issuing guidelines and examples of high-risk and non-high-risk AI systems and further protecting the fundamental rights.[7] Therefore, under Article 27, public bodies and private organisations engaged with public services must disclose how the system could impact people's fundamental rights by outlining risks associated with the system, and the further steps to be implemented in a high-risk situation. This assessment must be done for the first use of the system, but can be updated if necessary. Unless exempt, they must report their findings to the market surveillance authority using the template provided by the AI Office.[8] The transparency of the high-risk AI system has been considered vital and must include clear information and instructions, identifying the system's capabilities and limitations, and any potential risks. The record-keeping of such data is important for the working mechanism of such systems.[9] Article 50 states that companies must inform users when they interact with an AI system or use AI for emotion recognition or biometric categorisation, or create or alter content; the company must disclose this, unless it's for legal purposes or the content is artistic or satirical. The EU AI Act provides India and other such nations with a useful example of how broad principles such as safety, transparency, accountability and equity can be converted into specific legal obligations.[10]

5.2 India's AI Governance Guidelines (2025) – 'Light-Touch' and 'Seven Sutras'

In November 2025, the Ministry of Electronics and Information Technology (MeitY) unveiled guidelines for AI governance in India.[11] The guidelines are based on seven voluntary principles adapted from the RBI's FREE-AI Committee report[12] and apply across all sectors: Trust; People First; Innovation over Restraint; Fairness & Equity; Accountability; Understandable by Design; and Safety, Resilience & Sustainability. India's future leadership depends on the ability to set an example in governing AI with technical foresight, which guarantees safety and inclusivity. India's Guidelines describe themselves as a technology-agnostic, techno-legal framework built on voluntary measures, as set out in the Preface and reinforced by the "Innovation over Restraint" sutra.[13] They further provide wider principles for responsible and safe AI, emphasising transparency and clarity. These principles are mere guidelines rather than strict sanctions like the EU's AI Act. They depend on existing legal mechanisms, which lack penalties and a compliance system.

5.3 Existing Legal Framework and Instruments in India

5.3.1 Analysing DPDP Act and EU AI Act

The EU AI Act regulates the development and deployment of artificial intelligence systems, emphasising risk-based classifications that dictate how AI is designed, tested, and utilised. The Act prioritises AI

[7] *Id.* art. 6.
[8] *Id.* art. 27.
[9] *Id.* art. 13.
[10] *Id.* art. 50.
[11] INDIA AI GOVERNANCE GUIDELINES, *supra* note 5.
[12] RESERVE BANK OF INDIA, COMMITTEE ON FRAMEWORK FOR RESPONSIBLE AND ETHICAL ENABLEMENT OF ARTIFICIAL INTELLIGENCE, FRAMEWORK FOR RESPONSIBLE AND ETHICAL ENABLEMENT OF ARTIFICIAL INTELLIGENCE (FREE-AI) (Aug. 13, 2025), https://www.rbi.org.in/Scripts/PublicationReportDetails.aspx?ID=1306.
[13] INDIA AI GOVERNANCE GUIDELINES, *supra* note 5, at 12.

systems and their technical management rather than designing a product-safety framework on personal data involvement. The Act, compliance with which is overseen by the European AI Office and national authorities, applies to any AI system affecting the EU market, regardless of where it is developed. However, India's DPDP Act seems similar to GDPR[14] when it is concerned with the collection, storage and processing of personal data focused on privacy, but limited to digital personal data. DPDP essentially focuses on personal rights and consent, mandating regulation over personal data through consent, access, and grievance redressal. The Data Protection Board of India (DPBI)[15] enforces it, with appeals to the Telecom Disputes Settlement and Appellate Tribunal (TDSAT), and it applies to data processing outside India that targets individuals within the country. DPDP Act is based on consensual processing of lawful data and its violations invite the imposition of penalties; however, the EU AI Act lacks a consent regime and instead imposes notice and disclosure obligations for transparency. Its penalties follow a percentage-of-revenue model with minimum thresholds, tiered by severity of its violations. And it raises the question whether such frameworks focus on personal data protection or on transparent risk-management regulation.

5.3.2 Analysing the IT Act, 2000 and the EU AI Act

The IT Act, 2000[16] was adopted to monitor digital governance, e-commerce and cybercrime. India has deviated from the global phenomenon of creating a separate AI Act – instead, it has chosen to adapt existing laws through reinterpretation and targeted amendments. Section 79 of the IT Act, which is the core provision, grants conditional immunity to intermediaries by exempting them from liability for third-party content, contingent on their passive role in data transmission. The IT Act's major drawback is the definition of 'intermediary' in Section 2(w)[17], which might not sufficiently encompass AI systems that generate or modify content autonomously. The protections under the section are limited when applied to AI systems, raising doubts about liability for AI-generated outputs. The fines and punishments similar to the criminal law system under the Act are also restricted. The EU AI Act offers a multi-layered, approach-driven mechanism for AI regulation, with a risk-assessment model and strict enforcement obligations. The Indian IT Act functions within a wider digital governance framework that has yet to decisively address the complications caused by AI technologies.

5.3.3 Analysing the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026

The clearest instantiation of the techno-legal method is found not in the Guidelines themselves but in the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, notified on 10 February 2026 and brought into force on 20 February 2026.[18] The Amendment introduces a statutory category of 'synthetically generated information' (SGI) — content that appears reasonably authentic but is algorithmically created, modified, or altered — and folds it into the

[14] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L 119) 1.
[15] Digital Personal Data Protection Act, 2023, § 18.
[16] Information Technology Act, 2000.
[17] *Id.* § 2(1)(w).
[18] Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, Gazette of India, G.S.R. 120(E) (Feb. 10, 2026), amending the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

existing intermediary due-diligence regime.[19] Intermediaries must ensure that such content is conspicuously labelled, that provenance metadata is embedded and preserved, and that users declare synthetic material at the point of upload; larger intermediaries that themselves enable the creation of synthetic content bear the heavier obligation of deploying automated detection.[20] Crucially, compliance is enforced through the conditionality of safe harbour under Section 79 of the IT Act rather than through any independent adjudicatory route: an intermediary that fails to meet these design obligations risks forfeiting the very immunity on which its operation depends. The Amendment is therefore a typical example of governance by design in which a legal obligation is translated into 'the labelling, watermarking, and filtering' layers of the platform, rather than into an adjudicable duty owed to an affected person. Yet it is here that the reallocation of authority this paper identifies becomes visible. The obligation runs between the State and the intermediary, and its currency is takedown speed — roughly three hours for unlawful content acted upon under a court order or government notification, and two hours for non-consensual intimate imagery.[21] The individual whose likeness, voice, or data is processed is the object of protection but not a participant in it: she acquires no right to notice of an automated determination, no reasons, and no forum in which to contest either the classification of her content or the decision to act, or not act, upon it. What the technical architecture delivers efficiently — detection, labelling, removal — it delivers in place of, not alongside, the procedural guarantees that contestability requires.

5.4 India's Law Plus Model

The Reserve Bank of India (RBI) FREE-AI Committee[22] undertook the assessment of AI adoption in financial institutions and identified challenges to innovation and integration. The Committee asserted that there is a rise in AI's transformative potential and associated risks, and it is important to review global governance approaches in financial services. The Committee recommended establishing a progressive framework that promotes ethical AI adoption in India's financial sector. The Office of the Principal Scientific Adviser[23] paper endorses a techno-legal model that ensures AI technologies comply with national technical, legal, and ethical standards. This governance approach emphasises transparency, explainability, provability, and enabling conditions aligned with India's constitutional values.

According to the India AI Governance Guidelines,[24] the goal is to encourage voluntary compliance rather than impose strict regulations. This balanced model unites mandatory legal baselines with voluntary guidelines to promote innovation—known as LAW PLUS.[25] By embedding legal obligations into the design and development phases of AI systems, this proactive approach ensures ongoing compliance. Therefore, clarity in AI development can increase investment, create new startups, and strengthen global leadership. The "Responsible AI by Design" principle will support large-scale adoption and encourage confidence in AI technologies.

But achieving enhanced inclusive AI requires systematised change across the full AI life cycle, including remedying structural imbalances in who develops, defines, owns, and governs AI systems. It also

[19] IT Amendment Rules, r. 2(1)(wa).
[20] *Id.* pt. II, sec. 3(2).
[21] *Id.* pt. II, sec. 2(b).
[22] FREE AI REPORT, *supra* note 12.
[23] Office of the Principal Scientific Adviser (OPSA), *Strengthening AI Governance Through Techno-Legal Framework* (Gov't of India, White Paper No. 2, Jan. 23, 2026).
[24] INDIA AI GOVERNANCE GUIDELINES, *supra* note 5.
[25] *Id.*

requires investments in AI capacity, infrastructure, and skills across countries and regions, as well as more representative data and benchmarks.[26]

6 Critical Review of India's Shift Towards Techno-Legal AI Governance

6.1 Assessment of the Effectiveness of the Techno-Legal Framework

Technologists design the basic infrastructure features that create and implement information policy defaults.[27] What Reidenberg highlighted three decades ago is still relevant in today's modern era – "technological developments outpace the rate of legal evolution. Consequently, today's regulations may easily pertain to yesterday's technologies. Second, today's technology may limit the ability of government to regulate."[28] Koops and Leenes, when discussing data governance, reiterate that 'hard privacy by design' is challenging for determining legal regulations of data. Although simple and very specific rules may be suitable for hardcoding in IT systems, one cannot rely on building in legal requirements in the design of data-processing systems in such a way that compliance with 'the' data protection requirements and 'the protection of the rights of the data subject' is ensured.[29] As Schartum notes, 'privacy by design orientated policy combined with privacy by design impeding legislation embodies a serious policy dilemma that makes it unrealistic to assume that privacy by design can be developed to its intended potential.'[30] How can one effectively regulate the governance of AI? The techno-legal framework has its own challenges. A recurring and important theme in scholarly and policy discussions focuses on the need to secure 'algorithmic accountability'. But to understand what securing meaningful algorithmic accountability might require, it becomes important to gain a deeper understanding of algorithmic power.[31] Thus, the most important question arises: How should accountability then be accomplished? Regulations could be derived from what Citron and Pasquale have referred to as "technological due process"— procedures ensuring that predictive algorithms live up to some standard of review and revision to ensure their fairness and accuracy.[32]

In a discussion paper released in June 2018, NITI Aayog states the overarching goal for a national AI strategy as one which will "leverage AI for economic growth, social development and inclusive growth, and finally as a 'Garage' for emerging and developing economies." NITI Aayog's role goes beyond recommending a policy approach; it also includes implementation and deployment.[33] The techno-legal framework under India's AI governance[34] is devoid of tangible compliance infrastructures, wherein legal requirements risk becoming symbolic, reinforcing inequality rather than accountability.[35] Voluntary AI

---

[26] Independent International Scientific Panel on Artificial Intelligence, Preliminary Report: Scientific Assessment of the Capabilities, Opportunities, and Risks of Artificial Intelligence (2026), https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report (last visited July 18, 2026).

[27] Joel R. Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, 76 Tex. L. Rev. 553, 20 (1997–1998).

[28] *Id.* at 35.

[29] Bert-Jaap Koops & Ronald Leenes, *Privacy Regulation Cannot Be Hardcoded. A Critical Comment on the 'Privacy by Design' Provision in Data-Protection Law*, 28 Int'l Rev. L. Comput. & Tech. 159 (2014).

[30] *Id.*

[31] Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 Regul. & Governance 505, 29 (2018).

[32] Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Wash. L. Rev. 1 (2014).

[33] NITI Aayog, Nat'l Inst. for Transforming India, National Strategy for Artificial Intelligence: #AIforAll (2018).

[34] India AI Governance Guidelines, *supra* note 5.

[35] W. Holmes Finch & Marya Butt, *Gaps in AI-Compliant Complementary Governance Frameworks' Suitability (for Low-Capacity Actors), and Structural Asymmetries in the Compliance Ecosystem—A Systematic Review*, 5 J. Cybersecurity & Privacy 101 (2025).

frameworks shaped by dominant providers shift obligations and, instead of assuming ethical maturity, prioritise regulatory capture by those providers. For an AI governance structure to be effective, the framework thus must be grounded in legal obligations.

6.2 Identification of Gaps in Contestability and Remedy

The Organisation for Economic Co-operation and Development (OECD), an intergovernmental economic organisation focused on stimulating world trade, includes a right to contest in its recommendations on AI.[36] The right to contest decisions is central to due process. Indeed, other familiar due process protections—for example, transparency, notice, and the right to an impartial arbiter—serve to strengthen contestation rights.[37] A utilitarian argument for a right to contest AI decisions can take three forms:[38]

1. May lead to increased acceptance of AI systems among individuals.
2. In major cases, for example, welfare benefits or child custody, process protections and contestations could prevent serious harm.
3. Even utilitarians might support individual process rights if AI decisions encroach on personal privacy.

In an ideal scenario, an impartial entity would be responsible for adjudicating disputes arising from AI decisions. This may be possible only in certain scenarios; for example, in the criminal justice system, defendants should have the opportunity to contest AI-driven decisions meaningfully before a judge. The GDPR incorporates both systemic governance measures and various individual rights for data subjects: transparency, notice, access, a right to object to processing, and, for those subject to automated decision-making, the right to contest certain decisions.[39] Furthermore, when AI is utilised in regulatory environments, constitutional due process may necessitate adjudication by a neutral party.[40] The increasing use of machine learning systems to make crucial decisions raises concerns about whether models can be fair and non-discriminatory. This seems intuitive and straightforward as a goal, but a number of complications arise in the context of fairness in AI systems.[41] In other words, the contestability of AI recommendations increases the need for a check mechanism. The primary purpose of the EU AI Act and the GDPR is to protect fundamental rights, but they originate from different sources — regulation of product safety versus privacy protection. This difference leads to inconsistencies in implementation, particularly in lawful data processing, consent, and impact assessments. Key differences are: firstly, the AI Act's requirement for human oversight in high-risk AI systems (Article 14) versus the GDPR's restrictions on automated decision-making (Article 22). Secondly, GDPR's principles of data minimisation and purpose limitation may conflict with the data needs of AI models. And finally, the AI Act's vague definitions of fairness and non-discrimination further complicate its alignment with GDPR's more detailed individual rights.[42] Scholars have often argued that instead of depending exclusively on static, audit-based mechanisms, these models should advocate for viewing compliance as an ongoing process—a dynamic system

[36] Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1963 (2021).
[37] *Id.* at 1974.
[38] *Id.*
[39] *Id.*
[40] *Id.* at 2039.
[41] Vidushi Marda, *Artificial Intelligence Policy in India: A Framework for Engaging the Limits of Data-Driven Decision-Making*, 376 PHIL. TRANS. R. SOC. A 20180087, 9 (2018).
[42] Emmanuel Kwasi Abrokwa, *Does the GDPR Impede AI Progress? A Critical Analysis*, SSRN (May 29, 2026), https://ssrn.com/abstract=7119642.

that adapts to technological advancements and the evolving capacities of organisations. It is apparent that effective AI governance for regulation may necessitate compliance infrastructures that develop alongside institutional capabilities, rather than requiring immediate and comprehensive implementation.

We need to create a meaningful right to contest AI, but it must be carefully designed to succeed. There is no single approach; legislators can start with different archetypes and refine them to avoid pitfalls. Contestation is influenced by various factors such as notice, reasoning, substantive law, and incentive structures. A successful contestation mechanism must consider not only the algorithm but the entire decision-making system—human, machine, and organisational—along with the legal framework.

A remedy for an automated decision within the Indian framework requires assembling one from instruments not designed for the purpose, and the exercise exposes the gap the paper diagnoses. The Digital Personal Data Protection Act, 2023 confers a right of grievance redressal against a data fiduciary and channels unresolved complaints to the Data Protection Board of India, with an appeal to the Telecom Disputes Settlement and Appellate Tribunal.[43] But the Act contains no analogue to Article 22 of the GDPR: it creates neither a right against solely automated decision-making nor a right to an explanation of one, so the data principal may complain about how her data was handled but cannot demand reasons for, or contest the logic of, the decision that data produced.[44] The intermediary route is narrower still. Section 79 of the IT Act immunises the intermediary rather than empowering the affected person, and its safe harbour presupposes a passive conduit — an uneasy fit for systems that generate or determine outcomes autonomously.[45] The residual avenues are general and ill-suited: grievance redressal under the Consumer Protection Act, 2019 treats the harm as a defect in a service rather than as a contested exercise of decisional power, and the constitutional writ jurisdiction under Articles 32 and 226 reaches state action but not, ordinarily, the private developers and deployers to whom the techno-legal model delegates the operative choices.[46] Set against the contestability baseline of notice, reasons, and adjudication before a neutral arbiter,[47] India's architecture offers redress for data and for content while leaving the automated decision itself substantially unreviewable. The EU's pairing of a human-oversight obligation for high-risk systems (Article 14 of the AI Act) with an individual right in respect of automated decisions (Article 22 of the GDPR) is not advanced here as a template to be copied; it simply marks the distance between a framework that anticipates contestation and one that, by embedding governance in design and leaving remedy to pre-existing and general law, does not.

6.3 How Can India Remedy the Issues under Its AI Governance?

AI has increased large-scale production of targeted content, which is often misleading and undermines the integrity of information and weakens public trust, social cohesion, and democratic discourse. Cybercriminals use AI for cyberattacks, disproportionately impacting vulnerable groups. The growing gap between advancing AI capabilities and effective risk management could lead to serious consequences. In Article 22, the GDPR dictates that for certain automated decisions, affected individuals must be provided "at least the right to obtain human intervention . . . to express his or her point of view and to contest the

[43] Digital Personal Data Protection Act, 2023, §§ 13, 18, 29.
[44] Regulation (EU) 2016/679, art. 22; Digital Personal Data Protection Act, 2023 (containing no equivalent right against automated decision-making). *See also* Kaminski & Urban, *supra* note 36.
[45] Information Technology Act, 2000, §§ 2(w), 79.
[46] Consumer Protection Act, 2019; INDIA CONST. arts. 32, 226.
[47] Kaminski & Urban, *supra* note 36, at 1974.

decision."[48] The right to contest goes beyond rectifying incorrect data that concludes a decision. The right to contest must be held to higher standards of not just correction, but regulation of personal rights. The HEW Report[49] published over 50 years ago contemplated the conception of 'data privacy' and data disclosure and concluded that affected individuals should have some agency in deciding "the nature and extent of such disclosure." These findings are still relevant today, in the context of AI – individual rights and the right to contestability will lead to better governance of modern information systems. As the Independent International Scientific Panel on AI report[50] highlights, growing disorder in global governance is observed, with some countries having introduced AI-specific legislation with fundamentally contradictory rules and compliance costs. Jurisdictions exhibit divergent regulatory philosophies, with no unified risk-management mechanism, no comparable evaluation standards, and limited cross-jurisdictional coordination, risking a fragmented regulatory landscape.[51] Hence, the inadequate, highly autonomous AI systems have necessitated the evolution of reliable methods for retaining control. Decisions that affect people's lives implicate dignity. Categorising individuals arguably objectifies them; affording a right to contest that categorisation restores at least some form of dignity.[52] An example of such contestation could look like – "An individual shall have a right to contest, as under the due process clause." Lawmakers should ensure that the implementation anticipates and closes foreseeable loopholes. To achieve comprehensive and inclusive AI governance, India must incorporate systemic changes not only to tackle structural imbalances (related to developers, rulemakers, and owners) but also to build capacity, invest in AI, and share infrastructure and knowledge across regions.

7 CONCLUSION

Pasquale correctly identifies that transparency alone does not always produce either redress or public trust in the face of institutionalised power or money.[53] After analysing multiple sources, the authors conclude that AI compliance exists along a non-linear range rather than a binary one. It has to exist on a spectrum from stringent legal effectiveness to aspirational ethics-based governance. India is yet to adopt a comprehensive legal framework for AI governance; the 'techno-legal' approach is highly ambiguous and requires essential reviews and checks. Amendments to the existing legislation could bridge the gap and foster innovation; however, a 'whole-of-government approach' could create a better resolution to the current shortcomings. Regulators must implement robust model governance mechanisms covering the entire AI model lifecycle, including model design, development, deployment, and decommissioning. Regulators must also invest in training and institutional capacity-building initiatives to ensure that they possess an adequate understanding of AI technologies and to ensure that the regulatory and supervisory frameworks match the evolving landscape of AI, including associated risks and ethical considerations. As discussed in detail, the right to contest AI should apply to high-risk sectors, and it must not be limited to decisions made solely by AI but also to human decisions that rely on AI. Many find AI systems opaque

[48] Regulation (EU) 2016/679, art. 22.
[49] SECRETARY'S ADVISORY COMMITTEE ON AUTOMATED PERSONAL DATA SYSTEMS, U.S. DEP'T OF HEALTH, EDUC. & WELFARE, RECORDS, COMPUTERS AND THE RIGHTS OF CITIZENS, Pub. No. (OS) 73-94 (1973).
[50] INDEPENDENT INTERNATIONAL SCIENTIFIC PANEL REPORT, *supra* note 26.
[51] *Id.*
[52] Christopher McCrudden, *Human Dignity and Judicial Interpretation of Human Rights*, 19 EUR. J. INT'L L. 655 (2008).
[53] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 8, 16 (2015).

and worry that autonomous decisions made by these systems will be inexplicable and have unintended consequences. They are concerned about the unethical sourcing of data and that these systems could be used for harmful activities. Therefore, there is an impending need for immutable audit trails and transparency in automated systems, which allow individuals to understand and challenge inaccuracies, and individuals under the 'due process clause' become entitled to reasonably calculated notice when governmental systems make adverse decisions. Fostering trust by maintaining transparency and safety could be achieved solely via ethical AI adoption that respects rights and upholds fairness. A technology, no matter how powerful, will only be adopted if people trust it. Overall, AI regulations need to be flexible, coordinated, and constantly updated to advance innovation while protecting human rights, consumer protection, and ethical values.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.7

CHAPTER 7

Algorithmic Bias in Automated Decision-Making: An Empirical Fairness Audit and Comparative Legal Analysis Across Healthcare, Labour, Criminal Justice, and Finance

ASMITA SRIVASTAVA[*] AND ASMITA SINGH[**]

Artificial intelligence systems increasingly mediate access to healthcare, credit, employment, and liberty, and a growing empirical literature documents substantial performance and outcome disparities across demographic groups in such systems—for instance, facial-analysis benchmarks reporting error rates as high as 34.7% for darker-skinned women against 0.8% for lighter-skinned men.[1]

This paper combines an original empirical fairness audit with a comparative legal analysis of the principal regulatory regimes governing automated decision-making: the European Union's Artificial Intelligence Act, the General Data Protection Regulation, and India's Digital Personal Data Protection Act, 2023. Four datasets spanning healthcare (diabetic readmission), labour and income (ACSIncome), criminal justice (COMPAS recidivism), and finance (HMDA mortgage lending) were audited for base-rate disparities, modeled using both a glass-box logistic regression classifier and a black-box XGBoost classifier, and evaluated using demographic-parity and equalized-odds metrics. Two hypotheses were tested: that explainable-AI techniques (SHAP) improve identification of discriminatory proxy variables (supported), and that more recent training data reduces bias relative to older data (only mixed support). An empirically grounded Algorithmic Risk Matrix classifies each use case into four risk tiers, and a post-processing mitigation experiment demonstrates that identified disparities are partially, but not fully, correctable. The legal analysis situates these findings within the EU's risk-tiered framework, the GDPR's individual right against solely automated decision-making, and India's largely consent-based regime, which conspicuously lacks an equivalent right. The paper argues that none of the three regimes currently operationalizes 'high risk' or 'significant risk' using quantitative fairness thresholds of the kind produced here, and that empirical audits of this type could supply the missing evidentiary bridge between statistical disparity and legal obligation.

Keywords: Algorithmic Fairness; AI Governance; Automated Decision-Making; Explainable AI; Comparative Data-Protection Law.

1 INTRODUCTION

Automated decision-making systems trained on historical data risk reproducing, and in some cases amplifying, the disparities already embedded in that data. This is now a live legal problem as much as a technical one: regulators in the European Union, and to a lesser extent in India, have begun to impose obligations on the developers and deployers of such systems, yet the statutory language used to trigger those obligations—'high risk,' 'significant risk,' 'material effect'—is rarely defined with reference to any measurable statistical threshold. This paper attempts to close part of that gap. Part 2 situates the inquiry

[*] Student at Bennett University, Greater Noida, Uttar Pradesh, India.
[**] Student at Bennett University, Greater Noida, Uttar Pradesh, India.
[1] Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACHINE LEARNING RSCH. 1, 1–15 (2018).

within the EU AI Act, the GDPR, and India's Digital Personal Data Protection Act, 2023 ("DPDP Act"), and identifies where each regime's language could, in principle, be operationalized using standard fairness metrics. Parts 3 through 8 report an original empirical audit of four public datasets spanning healthcare, labour, criminal justice, and finance, testing (i) whether explainable-AI tooling meaningfully aids discrimination detection, (ii) whether data recency mitigates bias, (iii) a structured, empirically grounded risk-classification framework, and (iv) whether identified disparities can be technically corrected through post-processing. Part 9 draws the technical and legal threads together, and Part 10 concludes.

2 LEGAL AND REGULATORY FRAMEWORK

2.1 The EU Artificial Intelligence Act

The EU Artificial Intelligence Act[2] entered into force on 1 August 2024 and adopts a risk-tiered structure: systems are classified as prohibited, high-risk, limited-risk, or minimal-risk, with obligations scaling accordingly. A system is "high-risk" if it falls within a use-case category enumerated in Annex III—including biometric identification, employment and worker management, access to essential private and public services (expressly including credit scoring and insurance risk-pricing), law enforcement, and the administration of justice[3]—unless the provider documents that the specific system does not pose a significant risk to health, safety, or fundamental rights. Three of this paper's four domains—healthcare, employment/income classification, and criminal justice—map directly onto Annex III categories, and mortgage lending falls within the essential-services category through credit scoring.[4] The audited domains overlap with several Annex III categories, although the legal classification depends on the particular purpose and deployment context of the AI system rather than on the sector alone. Annex III expressly covers specified uses relating to employment, access to essential private services including creditworthiness evaluation, certain law-enforcement applications, and specified healthcare-related uses.[5]

High-risk providers must maintain a risk-management system throughout the system's lifecycle, ensure that training data is subject to "appropriate data governance and management practices," and enable human oversight, but the Act does not itself specify a quantitative fairness metric or threshold that a provider must satisfy—the determination of "significant risk" and the adequacy of bias-mitigation measures are left to the provider's own documented assessment, reviewable by national authorities.

The legal significance of the empirical audit is particularly apparent when read against Articles 9, 10, 14 and 27 of the AI Act.[6] Article 9 requires providers of high-risk AI systems to establish a continuous and iterative risk-management system covering the system's lifecycle, including the identification, estimation and mitigation of risks to health, safety and fundamental rights. Article 10 further requires appropriate data-governance and management practices for training, validation and testing datasets, while Article 14 requires effective human oversight.[7] The relevance to the present audit is therefore not that demographic-parity or equalized-odds thresholds are themselves statutory thresholds, but that an empirical

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, 2024 O.J. (L 1689) [hereinafter AI Act].
[3] Regulation (EU) 2024/1689, art. 6(2) & annex III.
[4] *Id.*
[5] *Id.* annex III.
[6] *Id.* arts. 6, 9–10, 14, 27.
[7] *Id.* arts. 9–10, 14, 27.

fairness audit can provide evidence relevant to the identification, measurement and mitigation of risks contemplated by these provisions. Article 27 strengthens this connection by requiring specified deployers of high-risk systems to conduct fundamental-rights impact assessments addressing affected groups, specific risks of harm, human oversight and remedial measures.

2.2 The General Data Protection Regulation

The GDPR[8] predates the AI Act and approaches automated decision-making through an individual-rights lens rather than a systemic risk-tiering. Article 22 gives a data subject the right not to be subject to a decision based solely on automated processing, including profiling, that produces legal or similarly significant effects—language that squarely covers credit, employment, and insurance decisions of the kind audited in this study. Article 5's fairness, accuracy, and data-minimisation principles, together with the Article 35 requirement of a data-protection impact assessment for processing likely to result in high risk, supply additional footholds, but as with the AI Act, none of these provisions specifies a statistical fairness test.[9] Scholars have argued that Article 22's disclosure and explanation obligations are difficult to operationalize without exactly the kind of feature-attribution analysis performed in Part 5 of this paper, since a data subject cannot meaningfully contest a decision without knowing which inputs drove it and whether those inputs behaved differently across protected groups.

2.3 India's Digital Personal Data Protection Act, 2023

India's Digital Personal Data Protection Act, 2023,[10] received presidential assent on 11 August 2023 and is being brought into force in phases; the implementing Digital Personal Data Protection Rules, 2025 were notified on 13-14 November 2025, with core substantive obligations phased in through May 2027. The DPDP Act is built almost entirely around consent: a Data Fiduciary must generally obtain a Data Principal's free, specific, informed consent before processing personal data, subject to enumerated "legitimate uses."[11] Critically, the DPDP Act contains no provision equivalent to GDPR Article 22: it does not grant an individual a freestanding right to contest a solely automated decision that produces legal or significant effects, and it imposes no risk-tiered obligations analogous to the EU AI Act's Annex III framework. Section 12 gives Data Principals a right to correction and erasure of personal data, but that right is oriented toward data accuracy rather than the fairness of an automated inference drawn from accurate data. This means the disparities documented empirically in this paper for HMDA and ACSIncome-type use cases—both of which have direct Indian analogues in mortgage and credit-scoring and in labour-market classification—would, under the DPDP Act as currently enacted, generate no individual procedural right against the automated decision itself, even though an equivalent decision made in the EU would trigger Article 22 and, where the deployment falls within Annex III, the AI Act's high-risk obligations.[12]

The absence of an express automated-decision provision in the DPDP Act does not mean that algorithmic decision-making is constitutionally unregulated in India. Depending on the actor and decision context, automated systems may implicate constitutional guarantees of equality, non-discrimination, dignity and privacy.

[8] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data (General Data Protection Regulation), 2016 O.J. (L 119) 1.
[9] *Id.* arts. 5, 22, 35.
[10] Digital Personal Data Protection Act, 2023 [hereinafter DPDP Act].
[11] DPDP Rules, 2025, Gazette of India, pt. II sec. 3(i) (Nov. 14, 2025); Digital Personal Data Protection Act, 2023, §§ 4–7.
[12] *See* Digital Personal Data Protection Act, 2023 (containing no automated-decision-making provision analogous to GDPR art. 22).

Article 14 of the Constitution guarantees equality before the law and equal protection of the laws, while Article 15 prohibits discrimination by the State on specified grounds.13 The Supreme Court's recognition of privacy as a constitutionally protected right in *Justice K.S. Puttaswamy (Retd.) v. Union of India*14 further established privacy, dignity and decisional autonomy as constitutionally protected interests under Article 21.

The constitutional framework therefore provides an important additional layer to the present comparison. The DPDP Act does not create a general right equivalent to GDPR Article 22 to contest a solely automated decision producing legal or similarly significant effects. Nevertheless, the legal consequences of algorithmic bias in India cannot be assessed exclusively through the DPDP Act, particularly where an automated system is deployed by the State or affects constitutionally protected interests. The precise legal consequences would depend upon the nature of the decision, the actor involved, the applicable statutory framework and the facts establishing the alleged infringement.

2.4 Comparative Synthesis

Read together, the three regimes sit on a spectrum from ex-ante systemic regulation (the AI Act's risk tiers, which regulate the system regardless of any individual complaint) to ex-post individual contestation (GDPR Article 22, which is triggered only when a particular data subject invokes it) to a consent-based regime that addresses neither systemic risk nor individual contestation of automated inferences (the DPDP Act). None of the three defines "high risk," "significant effect," or an equivalent trigger by reference to a measurable fairness statistic such as a demographic-parity or equalized-odds difference. The empirical findings in Parts 3 through 8 are offered as a candidate for exactly that missing quantitative layer: a worked example of how likelihood, severity, and extent of harm can be derived from an auditable pipeline rather than left to qualitative judgment alone.

3 DATA AND METHODOLOGY

Four datasets were selected to correspond to the domains above: healthcare (Diabetic 130-US Hospitals readmission data, 1999-2008), labour/income (ACSIncome, the Folktables successor to the UCI Adult dataset, pulled for 2014 and 2022), criminal justice (COMPAS two-year recidivism data), and finance (HMDA mortgage-lending data, pulled for 2019 and 2024). ACSIncome and HMDA were each pulled for two years specifically to test the recency hypothesis (H2), since they are the only two datasets with genuine year-over-year structure; the Diabetes dataset carries only an aggregate 1999-2008 range at the dataset level rather than a per-record date, so H2 could not be tested on it.

For each dataset, the pipeline proceeded in three stages. First, a protected-attribute audit computed base-rate gaps in the target variable across protected-group categories prior to any modeling, to establish raw, data-level disparity independent of any classifier. Second, baseline models—a glass-box logistic regression model and a black-box XGBoost model—were trained on identical features with class-weighting to correct for label imbalance; an unweighted Diabetes model, for example, achieved 89% accuracy but only 1% recall on the minority readmission class, rendering any fairness metric computed on it meaningless. Third, the full Fairlearn metric suite was computed per protected group, including accuracy, false-positive/negative rates, demographic-parity difference/ratio, and equalized-odds difference/ratio.

13 INDIA CONST. arts. 14–15.
14 Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

3.1 Empirical Risk Classification Versus Legal Risk Classification

The proposed Algorithmic Risk Matrix is not intended to reproduce or replace the legal classification established by the AI Act. The AI Act uses legally defined risk categories and specified use cases, whereas the present matrix applies a quantitative research framework to empirical model outcomes. Its function is therefore evidentiary and prioritising rather than determinative: it provides a reproducible method for identifying systems that may warrant heightened legal and technical scrutiny.

This distinction is essential because the numerical thresholds used in the present study—0.20 and 0.45 for equalized-odds difference—are researcher-defined empirical thresholds and are not statutory thresholds under the AI Act. A system classified as "Critical" under this matrix does not thereby become legally "high-risk" under the AI Act. Conversely, a system may fall within a legally defined high-risk category even where a particular fairness metric appears comparatively modest. The proposed framework should therefore be understood as a supplementary evidentiary tool that can support, but cannot substitute for, the legally prescribed risk-assessment process.

The framework nevertheless corresponds conceptually with Article 9 of the AI Act, which requires the identification, estimation and mitigation of risks throughout the lifecycle of high-risk AI systems.[15] It also has relevance to Article 27, which requires specified deployers to assess affected groups, specific risks and mitigation measures through a fundamental-rights impact assessment.[16]

Supplementary Table 1. Datasets, domains, and protected attributes examined in this study.

Domain	**Dataset**	**Protected attributes examined**	**Years**
Healthcare	Diabetic 130-US Hospitals (readmission)	Race, gender, age	1999-2008 (aggregate range only)
Labour / income	ACSIncome (Folktables)	Race, sex	2014 and 2022
Criminal justice	COMPAS (2-year recidivism)	Race, sex	Single cross-section
Finance	HMDA (mortgage lending)	Race, sex, ethnicity	2019 and 2024

Note: the Diabetic 130-US Hospitals dataset carries no per-record date or year field, only an aggregate 1999-2008 range at the dataset level, so the recency hypothesis (H2, Part 6) could not be tested on it with a genuine temporal split; only ACSIncome and HMDA provide real recency tests.

4 Empirical Findings: Base-Rate and Model-Level Disparities

Base-rate gaps in the raw data, prior to any modeling, were substantial across every domain and were consistently largest along the race axis except in the Diabetes dataset, where age dominated. HMDA's racial gap in loan-approval rate was 50.7 percentage points in 2019, narrowing to 44.4 points in 2024. COMPAS's racial gap in two-year recidivism (27.4 points between the highest and lowest race categories; 51.4% for Black defendants versus 39.4% for White defendants) closely matches the widely cited investigative findings that first brought the tool's racial disparities to public attention, which serves here as a validation of the audit pipeline before it is applied to the newer datasets.

[15] Regulation (EU) 2024/1689, art. 9, 2024 O.J. (L 1689) 1.
[16] *Id.* art. 27.

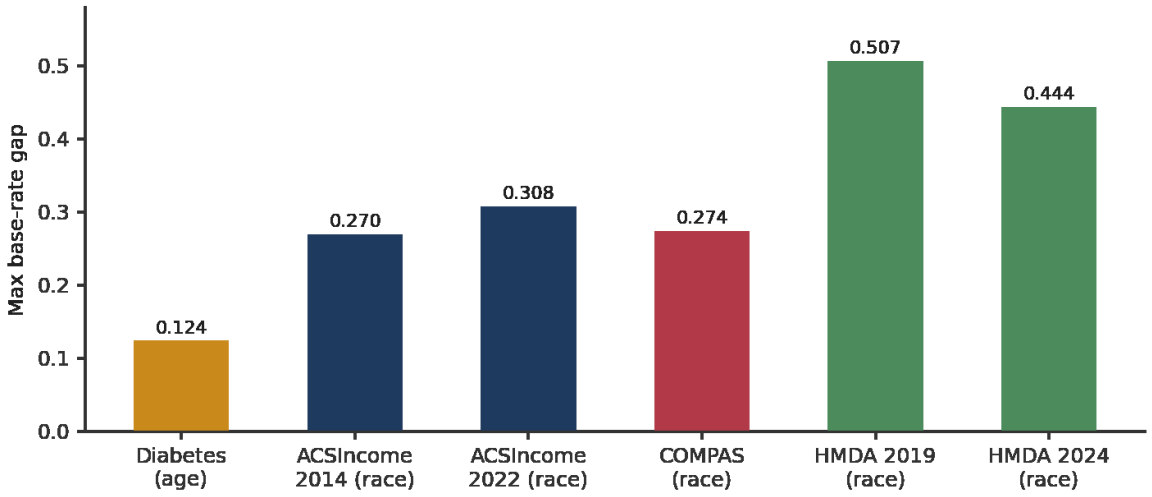

FIGURE 1. *Largest raw outcome disparity observed for each dataset on its dominant protected attribute, computed prior to any model training.*[17]

4.1 Legal Significance of the Base-Rate Disparities

The presence of a statistical disparity does not, by itself, establish unlawful discrimination. Its legal significance depends upon the governing jurisdiction, the protected characteristic, the decision context and the relationship between the challenged practice and the observed disparity. Statistical disparity can nevertheless constitute important evidence in discrimination analysis. In *Griggs v. Duke Power Co.*[18], the U.S. Supreme Court held that a facially neutral employment practice may violate Title VII where it disproportionately excludes a protected group and cannot be justified by business necessity. Title VII subsequently codified the disparate-impact framework, including the employer's burden to demonstrate that a challenged practice is job related and consistent with business necessity once the statutory conditions are met.[19] Accordingly, the disparities observed in Figure 1 should be treated as potential evidentiary signals rather than findings of legal liability. The purpose of the base-rate analysis is to establish whether unequal outcomes exist before model training; determining whether those disparities are legally attributable to discriminatory criteria, historical inequality, legitimate predictive factors or other causes requires a separate legal and causal inquiry.

At the model level, XGBoost achieved accuracy ranging from 0.665 (Diabetes) to 0.831 (HMDA 2019), broadly consistent with published benchmarks for these datasets. Demographic-parity and equalized-odds differences on each dataset's dominant protected attribute (race, except age for Diabetes) ranged from 0.14 (HMDA 2024) to 0.78 equalized-odds difference (Diabetes, age), with COMPAS showing the largest equalized-odds gap on race (0.611) of any dataset. Race produced consistently larger demographic-parity differences (0.17-0.39) than sex or ethnicity (0.01-0.16) across every dataset, and the Diabetes dataset's largest disparity ran along age rather than race or sex—a finding of independent legal significance, since age discrimination is generally treated as a distinct doctrinal category from race or sex discrimination.

[17] Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[18] Griggs v. Duke Power Co., 401 U.S. 424, 430–32 (1971).
[19] 42 U.S.C. § 2000e-2(k)(1)(A).

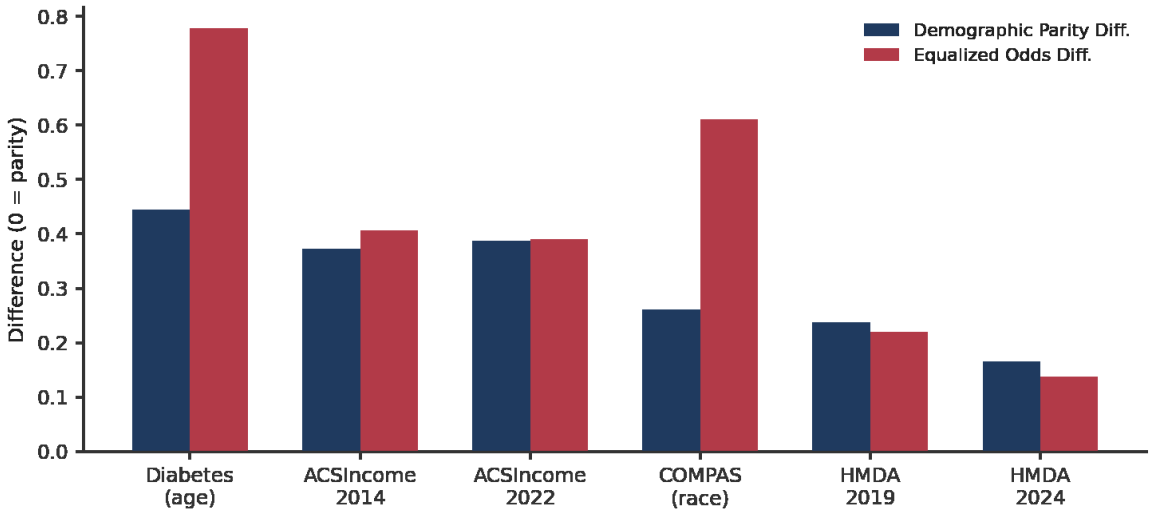

FIGURE 2. *Demographic-parity and equalized-odds differences on each dataset's dominant protected attribute (race, except age for Diabetes), XGBoost model.*

4.2 Legal Significance of Demographic-Parity and Equalized-Odds Differences

The fairness metrics used in this study should not be treated as independent legal tests of discrimination. Demographic parity measures differences in selection or prediction rates between groups, while equalized odds examines whether relevant error rates are distributed similarly across groups. Neither metric is expressly established as a dispositive legal threshold by the AI Act, the GDPR or the DPDP Act. Their legal value is therefore evidentiary rather than determinative.

A substantial equalized-odds difference may demonstrate that an automated system performs differently across protected groups and may justify further examination of its training data, features, proxy variables, decision thresholds and potential harms. It does not, however, establish that the system has committed unlawful discrimination. Statistical disparity may result from differences in base rates, feature distributions, model design or other structural conditions. The legal inquiry must therefore proceed from statistical evidence to the applicable jurisdiction-specific test.

This distinction is consistent with disparate-impact doctrine in U.S. discrimination law, where statistical disparities may form an important part of the evidentiary showing but do not independently resolve the legal question. See *Watson v. Fort Worth Bank & Trust*[20]. The results in Figure 2 should therefore be understood as identifying systems and protected groups requiring closer legal and technical scrutiny, rather than as automatically classifying any model as unlawful.

Supplementary Table 2. XGBoost baseline performance across datasets.

Dataset	Model	Accuracy	AUC	Note
Diabetes	XGBoost	0.665	0.682	Class-imbalance corrected
ACSIncome 2014	XGBoost	0.807	0.899	Balanced classes
ACSIncome 2022	XGBoost	0.804	0.890	Balanced classes
COMPAS	XGBoost	0.672	0.715	Matches published literature range
HMDA 2019	XGBoost	0.831	0.838	Stratified 150k sample

[20] Watson v. Fort Worth Bank & Trust, 487 U.S. 977, 994–95 (1988).

Dataset	Model	Accuracy	AUC	Note
HMDA 2024	XGBoost	0.823	0.861	Stratified 150k sample

The full fairness-metric results underlying Figure 2—accuracy, demographic-parity difference/ratio, and equalized-odds difference/ratio for both the logistic-regression and XGBoost models, across every protected attribute in every dataset—are reported in Appendix A.

5 EXPLAINABILITY AND THE DISCRIMINATORY-PROXY PROBLEM (H1)

For each dataset's XGBoost model, SHAP[21] (SHapley Additive exPlanations) was used to compute per-feature contribution to each prediction, measuring both global feature importance and the divergence of a feature's importance across protected-group categories—the latter being the signature of a feature functioning as a proxy for a protected characteristic even where the protected characteristic itself is excluded from the model.

The results support H1. In ACSIncome, place of birth was the single strongest race-divergent feature in 2014 (age in 2022), a textbook proxy variable since place of birth correlates with race and national origin; hours worked per week diverged most by sex, plausibly reflecting labour-market structure rather than an algorithmic artifact. In COMPAS, prior-arrest count and age dominated both global importance and racial divergence, consistent with longstanding critiques that criminal-history features themselves encode racially disparate policing and arrest patterns, allowing a model to exclude race directly while still reproducing racial disparity through its inputs. In HMDA, debt-to-income ratio was by far the dominant feature globally and the most group-divergent feature across race, sex, and ethnicity in 2019 and by race in 2024 (income diverged most by sex and ethnicity in 2024), indicating a major channel for disparate mortgage outcomes.

SHAP divergence establishes only that a feature's influence differs by group; it does not by itself establish that the feature is an illegitimate discriminatory proxy rather than a feature that is legitimately predictive and happens to correlate with a protected group for independent structural reasons. Resolving that distinction is a legal and causal judgment that lies at the core of disparate-impact doctrine[22] and of Article 22 GDPR's explanation requirement, which scholars have argued cannot be satisfied without feature-attribution methods of exactly this kind.[23]

The legal significance of the SHAP results lies in their ability to identify features that may reproduce disparities even when a protected characteristic is excluded explicitly from the model. Place of birth, geographic variables, criminal-history variables and financial variables may correlate with protected characteristics and thereby contribute to unequal outcomes. Such correlation does not itself establish unlawful discrimination because a feature may be legitimately predictive while also correlating with a protected characteristic. The relevant legal inquiry is whether the feature produces a legally cognizable discriminatory effect, whether its use has a legitimate justification where required by the applicable law, and whether less discriminatory alternatives are available.

The distinction is particularly relevant to the disparate-impact framework recognized in U.S. employment discrimination law. See *Griggs v. Duke Power Co.*24 It also has relevance under the AI Act's data-governance and risk-management framework, which requires attention to the quality and governance of datasets and to

[21] Scott M. Lundberg & Su-In Lee, *A Unified Approach to Interpreting Model Predictions*, 31 ADV. NEURAL INFO. PROCESSING SYS. 4765 (2017).

[22] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 680–93 (2016) (analyzing how facially neutral variables can operate as proxies for protected classes).

[23] Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 843–45 (2018).

[24] *Griggs*, 401 U.S. at 430–32.

risks to fundamental rights. SHAP divergence should therefore be understood as a screening mechanism that identifies features requiring legal and causal investigation, rather than as proof that a feature constitutes an unlawful proxy.

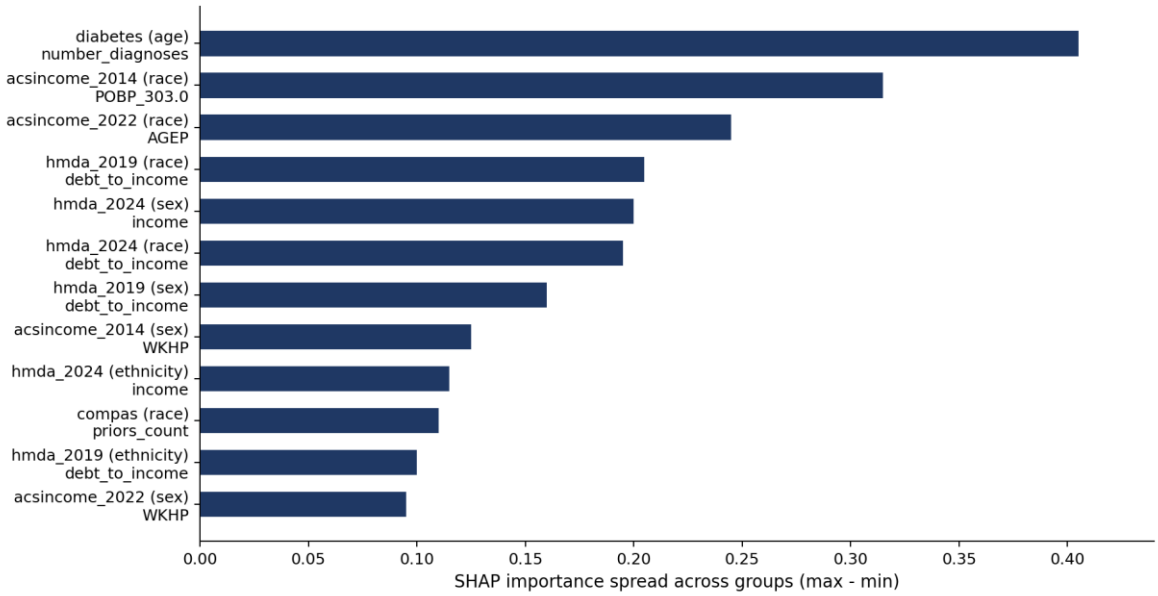

FIGURE 3. *Strongest candidate proxy feature identified per dataset × protected-attribute pairing. Bar length is the spread in SHAP importance for that feature across protected-group categories.*

6 RECENCY AND BIAS (H2)

H2 was tested on the only two datasets with genuine year-over-year structure. In ACSIncome, the race-based demographic-parity difference increased with recency (0.373 to 0.388) while the sex-based disparity decreased (0.082 to 0.062). In HMDA, the race-based difference decreased substantially with recency (0.237 to 0.165) while the sex-based disparity increased slightly (0.057 to 0.108). Recency's effect on bias therefore appears to be domain- and attribute-dependent rather than a uniform mitigant, a more defensible conclusion than a blanket claim that newer data is fairer data, and one with direct policy relevance to any regime—such as the AI Act's data-governance obligations—that treats data currency as a proxy for data quality.

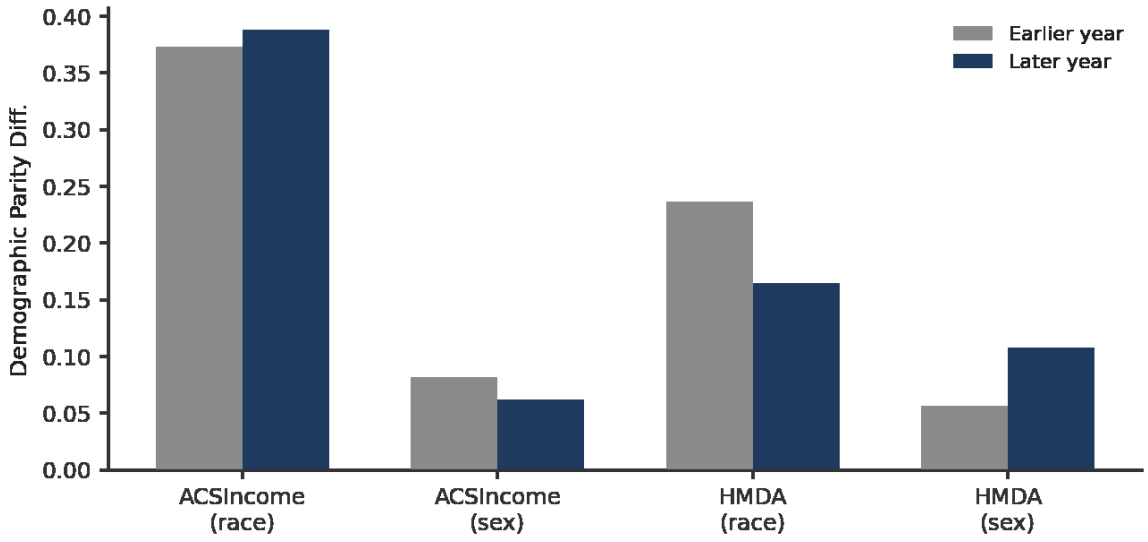

FIGURE 4. *Demographic-parity difference by dataset, protected attribute, and year (earlier year vs. later year).*

The mixed effect of data recency has an important legal implication. Newer data cannot automatically be treated as more representative, accurate or less discriminatory. Article 10 of the AI Act adopts a broader data-governance approach by requiring attention to data quality, relevance, representativeness, preparation, labelling, updating and suitability for the intended purpose.[25] The results in Figure 4 support this approach empirically: racial disparity decreased in HMDA while increasing in ACSIncome despite both datasets being updated to a later period. Temporal recency therefore appears to be an inadequate proxy for fairness. For regulatory purposes, the quality and representativeness of data should be assessed substantively rather than inferred merely from the date on which the data were collected.

7 THE ALGORITHMIC RISK MATRIX AND ITS LEGAL GROUNDING

Each use case was scored on three factors—Likelihood, Severity, and Extent of harm, each on a 1-3 scale—multiplied into a risk score (maximum 27) and mapped to four risk levels (I Low to IV Critical). Likelihood was derived empirically as the worst-case equalized-odds difference found for that dataset's XGBoost model (below 0.20 = Low, 0.20-0.45 = Medium, at or above 0.45 = High). Severity and Extent are necessarily domain-level judgment calls rather than purely data-derived quantities; here they were informed by the AI Act's Annex III high-risk classifications discussed in Part 2.1, offered as a defensible starting point for legal or policy review rather than a final determination.

Table 1. *Algorithmic Risk Matrix by dataset/use case. Score = Likelihood × Severity × Extent (max 27).*

Dataset	Domain	Max EO Diff	Score	Risk Level
COMPAS	Criminal justice	0.611	27	IV — Critical
Diabetes	Healthcare	0.778	18	III — High
ACSIncome 2022	Labour/income	0.390	12	II — Limited
ACSIncome 2014	Labour/income	0.407	12	II — Limited
HMDA 2019	Finance	0.220	12	II — Limited

[25] Regulation (EU) 2024/1689, art. 10, 2024 O.J. (L 1689) 1.

Dataset	Domain	Max EO Diff	Score	Risk Level
HMDA 2024	Finance	0.138	6	I — Low

The resulting ordering—COMPAS (Critical), Diabetes (High), ACSIncome and HMDA 2019 (Limited), HMDA 2024 (Low)—tracks the AI Act's own severity hierarchy reasonably well: criminal-justice risk-assessment and healthcare-access systems sit within Annex III categories that the Act itself treats as most severe, while mortgage lending, though nationally significant in extent, involves reversible economic rather than liberty or health harm. Notably, HMDA moves from Risk Level II in 2019 to Level I in 2024, a concrete link back to the H2 finding that recency reduced HMDA's race-based disparity, illustrating how the empirical results and the risk-classification framework are directly connected and, in principle, could be re-run periodically to support a provider's ongoing Annex III risk documentation obligation.

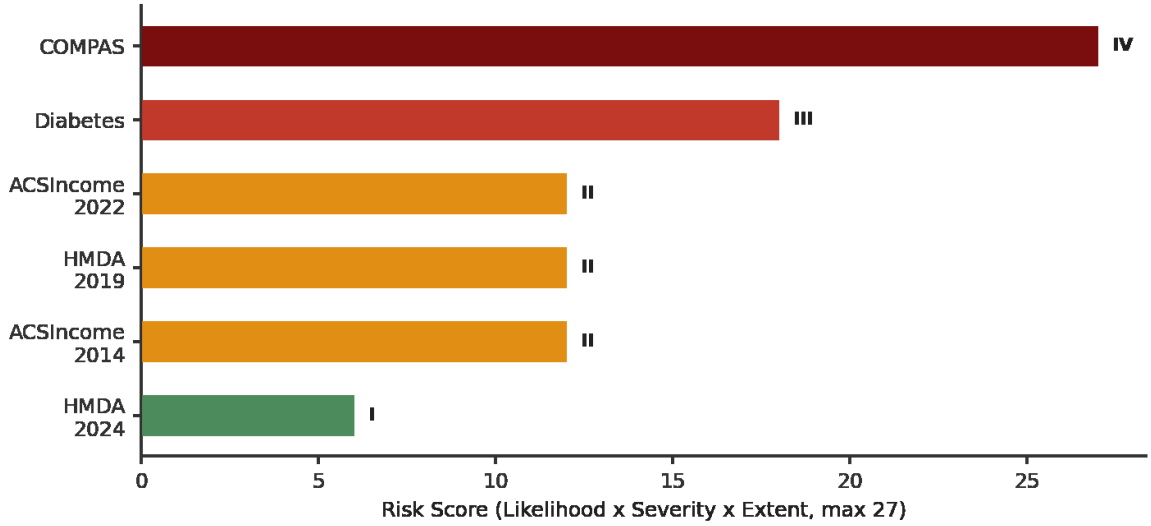

FIGURE 5. *Risk score (Likelihood × Severity × Extent, max 27) by dataset/use case, with mapped risk level, corresponding to Table 1.*

8 MITIGATION EXPERIMENT AND THE LIMITS OF TECHNICAL CORRECTION

The two highest-risk use cases—COMPAS (Level IV) and Diabetes (Level III)—were subjected to a post-processing mitigation experiment using Fairlearn's ThresholdOptimizer, which learns separate decision thresholds per protected group so that false-positive and false-negative rates are equalized across groups.[26] For COMPAS, equalized-odds difference fell 42% (0.611 to 0.354) at a cost of 2.2 accuracy points and a modest recall drop, while demographic parity barely moved (0.261 to 0.256)—an illustration of Kleinberg et al.'s finding that distinct fairness criteria are not generally improved together by the same intervention.

For Diabetes, an initial mitigation attempt produced an apparently spectacular result—accuracy rising from 66% to 89% and the equalized-odds gap collapsing from 0.74 to 0.02—that on inspection proved to be a false positive: the "corrected" model was simply predicting no-readmission for 99.7% of cases against an actual readmission rate of 11%, satisfying equalized odds only because it was equally useless

[26] On the trade-offs inherent in any such fairness-constrained optimization, *see* Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, 8th INNOVATIONS IN THEORETICAL COMPUT. SCI. CONF. 43:1 (2017) (proving that demographic parity and equalized-odds-type criteria generally cannot be simultaneously satisfied except in degenerate cases).

for every age group. Switching the optimizer's objective from accuracy to balanced accuracy corrected this "fairness through triviality" failure mode; the corrected run reduced the equalized-odds gap by 23% (0.741 to 0.571) and demographic parity by 79% (0.488 to 0.102), with recall falling only modestly (0.587 to 0.478). The episode is itself a legally relevant finding: it demonstrates why any compliance claim of bias having been "fixed"—whether under the AI Act's risk-management obligations or a voluntary fairness audit—should be accompanied by recall and utility figures subject to independent verification, since a fairness-metric improvement reported in isolation can conceal a degenerate, non-functional model.

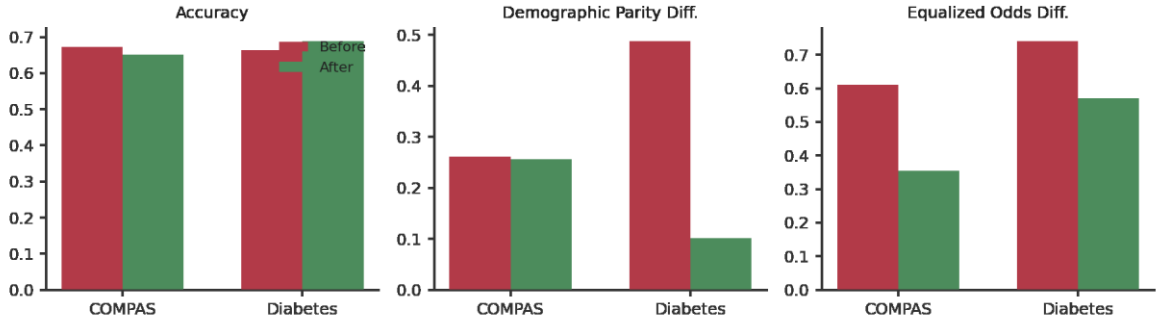

FIGURE 6. *Before/after comparison of accuracy, demographic-parity difference, and equalized-odds difference following ThresholdOptimizer mitigation (balanced-accuracy objective, equalized-odds constraint).*

8.1 Legal Significance of the Mitigation Results

The mitigation experiment demonstrates why legal compliance cannot be reduced to optimisation of a single fairness metric. A post-processing intervention may reduce equalized-odds disparity while simultaneously reducing recall or other measures of predictive utility. This is particularly important in high-impact domains, where the costs of false-positive and false-negative outcomes may themselves affect legally protected interests. The AI Act's risk-management framework similarly requires risks to be identified and mitigated rather than merely producing compliance with one numerical fairness measure.[27] The Diabetes result is especially significant because the initial apparent fairness improvement resulted from a model that predicted the negative outcome for almost all observations. The finding demonstrates that an isolated improvement in a fairness metric cannot establish that an AI system has been meaningfully corrected. A legally relevant audit should therefore report fairness measures together with accuracy, recall, error rates and other context-specific measures of system utility. This supports a multidimensional conception of algorithmic accountability in which mitigation must reduce discriminatory risk without rendering the system substantively non-functional.

Supplementary Table 3. Mitigation results. Recall on the positive class is reported alongside the fairness metrics to guard against the 'fairness through triviality' failure mode described above. The Diabetes before-mitigation figures come from the mitigation run and differ from the baseline figures in Table 1 and Appendix Table A1.

Dataset	Stage	Accuracy	Recall (pos. class)	DP Diff	EO Diff
COMPAS	Before mitigation	0.672	0.625	0.261	0.611
COMPAS	After mitigation	0.651	0.569	0.256	0.354
Diabetes	Before mitigation	0.664	0.587	0.488	0.741
Diabetes	After mitigation	0.689	0.478	0.102	0.571

[27] Regulation (EU) 2024/1689, art. 9, 2024 O.J. (L 1689) 1.

9 Discussion: Bridging the Technical and Legal Findings

Three points of synthesis follow. First, the persistent dominance of race as the largest disparity axis across every domain except Diabetes (where age dominates) supports treating race-based algorithmic disparity as the AI Act's Annex III framework implicitly does: as a first-order systemic risk warranting close regulatory attention regardless of sector, while cautioning that a single dominant-axis metric can obscure real disparities in other protected characteristics that a narrower audit would miss. Second, the SHAP findings in Part 5 give concrete content to the disparate-impact proxy-discrimination inquiry that both EU and Indian discrimination law recognize in principle but rarely operationalize empirically; a regulator applying the AI Act's data-governance requirements could, in principle, require exactly this kind of divergence analysis as evidence of "appropriate" bias examination. Third, and most directly a legal gap, is the DPDP Act's absence of any provision equivalent to GDPR Article 22: an Indian mortgage applicant or job-seeker subject to a model exhibiting the disparities documented here for HMDA- and ACSIncome-type use cases has, under current Indian law, no individual procedural right against the automated decision itself, even though the underlying statistical disparity is of the same order of magnitude as that documented for the EU-regulated equivalents. This asymmetry is not merely theoretical: India's technology and financial-services sectors deploy automated credit and employment screening at a scale comparable to the jurisdictions this paper's data was drawn from, and the absence of an operative automated-decision right leaves a comparable population without a comparable remedy.

The paper's central methodological limitation is one shared by any audit of this kind: SHAP divergence, base-rate gaps, and post-processing mitigation results are correlational and cannot, standing alone, establish that a given disparity is unlawful discrimination rather than a legitimately predictive, structurally correlated feature. Severity and Extent scores in the Risk Matrix remain domain-level judgment calls rather than validated legal categories. These limitations argue for treating the empirical pipeline presented here as a candidate evidentiary tool for regulators and litigants—one that could inform where the AI Act's "significant risk" documentation burden should fall most heavily, or what a GDPR Article 22 explanation should contain—rather than as a self-executing legal determination.

10 Conclusion

This paper finds substantial, measurable disparities across healthcare, labour, criminal justice, and finance, with race the dominant axis of disparity in three of the four domains examined. Explainability tooling meaningfully surfaces dataset-specific proxy variables invisible to aggregate fairness metrics alone, though it cannot resolve whether a given proxy is unlawful in a legal sense. Data recency shows a genuinely mixed effect, cutting against a blanket regulatory preference for newer training data. Identified disparities are only partially correctable through standard fairness-aware post-processing, with a documented failure mode—fairness improvements achieved through a degenerate, functionally useless model—that any legal compliance regime relying on fairness metrics must guard against. None of the three regulatory regimes examined operationalizes its central trigger for obligation using a quantitative fairness threshold; the empirically derived Risk Matrix developed here is offered as one template for what that missing quantitative layer could look like, and as a demonstration that the gap between an EU data subject's Article 22 rights and an Indian data principal's current lack of an equivalent right is not merely doctrinal, but corresponds to comparable underlying statistical disparities documented across this paper's four domains.

DATA AND CODE AVAILABILITY

All four datasets analyzed in this study are publicly available from their original sources. The Diabetic 130-US Hospitals readmission dataset is available from the UCI Machine Learning Repository (archive.ics.uci.edu, dataset ID 296). ACSIncome is available through the Folktables package (github.com/socialfoundations/folktables), which derives it from the U.S. Census Bureau's American Community Survey Public Use Microdata Sample. The COMPAS two-year recidivism dataset is available from ProPublica's compas-analysis repository (github.com/propublica/compas-analysis). The HMDA mortgage-lending data is available from the Consumer Financial Protection Bureau's HMDA Platform (ffiec.cfpb.gov/data-browser). The analysis code implementing the audit pipeline described in Part 3 — data preprocessing, the glass-box and black-box classifiers, the Fairlearn fairness-metric computations, the SHAP explainability analysis, the Algorithmic Risk Matrix scoring, and the ThresholdOptimizer mitigation experiment — is available from the authors upon reasonable request.

APPENDIX A. FULL FAIRNESS-METRIC RESULTS

Full fairness-metric results across all datasets, models (glass-box logistic regression and black-box XGBoost), and protected attributes, underlying Figure 2 in Part 4. DP = Demographic Parity; EO = Equalized Odds. Ratio metrics closer to 1.0 indicate more parity; difference metrics closer to 0.0 indicate more parity.

Appendix Table A1. Full fairness-metric results across all datasets, models, and protected attributes.

Dataset	Model	Attribute	Acc.	DP Diff	DP Ratio	EO Diff	EO Ratio
ACSIncome 2014	LogReg	race	0.812	0.344	0.306	0.396	0.000
ACSIncome 2014	LogReg	sex	0.812	0.081	0.825	0.015	0.928
ACSIncome 2014	XGBoost	race	0.807	0.373	0.300	0.407	0.287
ACSIncome 2014	XGBoost	sex	0.807	0.082	0.827	0.016	0.929
ACSIncome 2022	LogReg	race	0.807	0.382	0.383	0.372	0.000
ACSIncome 2022	LogReg	sex	0.807	0.061	0.884	0.006	0.989
ACSIncome 2022	XGBoost	race	0.804	0.388	0.384	0.390	0.000
ACSIncome 2022	XGBoost	sex	0.804	0.062	0.884	0.005	0.980
COMPAS	LogReg	race	0.668	0.292	0.520	0.527	0.000
COMPAS	LogReg	sex	0.668	0.137	0.727	0.103	0.722
COMPAS	XGBoost	race	0.672	0.261	0.533	0.611	0.000
COMPAS	XGBoost	sex	0.672	0.158	0.665	0.127	0.601
Diabetes	LogReg	age	0.659	0.448	0.000	0.759	0.000
Diabetes	LogReg	gender	0.659	0.013	0.964	0.048	0.915
Diabetes	LogReg	race	0.659	0.229	0.365	0.408	0.274
Diabetes	XGBoost	age	0.665	0.445	0.000	0.778	0.000
Diabetes	XGBoost	gender	0.665	0.021	0.943	0.047	0.923
Diabetes	XGBoost	race	0.665	0.185	0.490	0.285	0.495

Dataset	Model	Attribute	Acc.	DP Diff	DP Ratio	EO Diff	EO Ratio
HMDA 2019	LogReg	ethnicity	0.713	0.112	0.838	0.088	0.811
HMDA 2019	LogReg	race	0.713	0.197	0.721	0.143	0.652
HMDA 2019	LogReg	sex	0.713	0.082	0.881	0.079	0.797
HMDA 2019	XGBoost	ethnicity	0.831	0.071	0.909	0.085	0.774
HMDA 2019	XGBoost	race	0.831	0.237	0.703	0.220	0.393
HMDA 2019	XGBoost	sex	0.831	0.057	0.927	0.078	0.788
HMDA 2024	LogReg	ethnicity	0.627	0.059	0.898	0.106	0.731
HMDA 2024	LogReg	race	0.627	0.297	0.515	0.304	0.494
HMDA 2024	LogReg	sex	0.627	0.122	0.795	0.147	0.654
HMDA 2024	XGBoost	ethnicity	0.823	0.106	0.861	0.059	0.816
HMDA 2024	XGBoost	race	0.823	0.165	0.792	0.138	0.666
HMDA 2024	XGBoost	sex	0.823	0.108	0.861	0.055	0.835

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.8

CHAPTER 8

AI and Legal Accountability in India: Rethinking Responsibility for Automated Decisions

NONY NEMA*

In this algorithmic age, AI systems are impacting a wide range of decisions across organizations. The growing use of AI and automated decision-making systems in India has created new challenges for legal accountability. AI systems are increasingly being used in almost every sector. While these systems can make decisions quickly and efficiently, there are chances that they may produce unfair, incorrect or harmful outcomes. This raises an important question: who should be legally responsible when an automated decision causes harm?

The paper examines whether the existing legal framework in India is sufficient to assign responsibility for harmful or unfair decisions made through AI and automated systems. The paper adopts a doctrinal approach to examine relevant legislation, judicial decisions, regulatory developments in India relating to AI, data protection, liability and automated decision-making. The paper argues that although the existing legal frameworks provide some mechanisms to address the harms associated with AI, they do not yet provide a clear and comprehensive framework for allocating legal responsibility for automated decisions. Responsibility is difficult to determine where multiple actors are involved which include developers, deployers, organizations and human users. This may create an accountability gap when an AI system produces an unexpected outcome.

The paper therefore, argues for a clear framework of responsibility based on the degree of control exercised by different actors in the decision-making process. The paper aims to examine how human oversight, transparency, explainability and effective review mechanisms can improve accountability for AI-based decisions in India.

Keywords: Artificial Intelligence, Automated Decision-Making, Legal Accountability, Algorithmic Bias, Human Oversight

1 INTRODUCTION

Artificial Intelligence is defined as the kind of technology that enables computers to perform functions which usually require human intelligence.[1] Artificial Intelligence is now being used in almost every sector be it healthcare, business, finance, education, etc. It is rapidly changing how decisions are being made in these fields. The growing use of AI may lead to algorithmic bias, errors caused due to automated decision-making and even transparency issues. According to Article 22 of the EU's GDPR, a data subject has the right to reject any decision made by automated systems, including algorithmic profiling, if that decision creates serious legal consequences or deeply impacts their life.[2] While these automated decisions make life faster and more efficient, they also create a major legal problem: who is responsible when an AI makes a mistake? When a technology leads to a financial loss, denies someone a job due to an algorithmic bias

* PhD Scholar at National Law Institute University, Bhopal, Madhya Pradesh, India.

[1] *Artificial Intelligence*, OXFORD REFERENCE, https://www.oxfordreference.com/display/10.1093/oi/authority.20110803095426960 (last visited Oct. 2, 2026).

[2] *Art. 22 GDPR – Automated Individual Decision-Making, Including Profiling*, GENERAL DATA PROTECTION REGULATION (GDPR), https://gdpr-info.eu/art-22-gdpr/ (last visited Sept. 8, 2026).

or gives a wrong medical diagnosis, traditional laws struggle to find the party who is to be held responsible.

Modern AI systems that are built on deep learning process massive amounts of data to find patterns and make decisions. Artificial intelligence is a dynamic system which keeps on learning and this makes it difficult to identify the defect or harm caused. The Consumer Protection Act, 2019 provides a way to address AI-related harm through product liability and deficiency of service provisions. However, since the law was not specifically designed for AI-related risks, its application to AI systems may raise certain legal challenges. There are hardly any court decisions which clearly determine whether algorithmic decision-making systems are covered within the ambit of product liability.[3]

The 'black-box effect' refers to advanced AI systems whose internal decision-making processes are completely hidden and unknown and that creates a gap in the existing legal frameworks because if the law can't trace how a machine reached a conclusion, it cannot fairly determine who to hold accountable for the harm caused. One solution to address this accountability issue is explainability, which involves the process of defining how an algorithm arrived at a particular solution. If an AI system's decision-making process can be defined and understood in a transparent manner then accountability can be determined easily. The EU's AI Act has recognized transparency and explainability as important mechanisms for ensuring accountability in AI-assisted decision-making. Under Article 86 of the act, the affected persons are in certain situations entitled to receive meaningful explanations of the role played by a high-risk AI system in the decision-making process.[4] This approach seeks to ensure that automated systems remain capable of scrutiny and promote fair and accountable decision-making.

The 'OECD AI Principles' are guidelines for developing and using AI that is trustworthy and beneficial to society. These principles were first adopted in 2019 and were the first intergovernmental standard on AI. Among these five principles, there are two principles which focus on 'transparency and explainability' and 'accountability', respectively. Principle 1.3 talks about 'Transparency and Explainability', and emphasizes transparency and responsible disclosure in the use of AI systems, ensuring that individuals are aware when they are interacting with AI and have the ability to question or challenge its outcomes. AI actors should provide meaningful and context-appropriate information about AI systems, consistent with current technological standards. This information should help stakeholders understand the capabilities and limitations of AI, recognize when they are interacting with an AI system, and, where practical, understand the data factors, processes or logic that contributed to an AI-generated prediction, recommendation, content or decision. Such transparency should also enable individuals affected negatively by an AI system to understand and challenge its outputs.[5]

According to Principle 1.5 that is the 'Accountability' principle, individuals and organizations that are involved in the development, deployment or operation of AI systems should be responsible for ensuring that these systems function appropriately and in accordance with the OECD's values-based AI principles. AI actors should assume responsibility for the performance of AI systems and compliance with these principles, taking into consideration their respective roles, the context in which the systems are used, and the level of technological development available. To support this accountability, appropriate traceability

[3] *Who Is Liable When Artificial Intelligence Fails* (Feb. 4, 2026), https://www.lawctopus.com/academike/who-is-liable-when-artificial-intelligence-fails/.

[4] *Article 86: Right to Explanation of Individual Decision-Making | EU Artificial Intelligence Act*, EU AI ACT EXPLORER, https://artificialintelligenceact.eu/article/86/ (last visited Sept. 8, 2026).

[5] *Transparency and Explainability (OECD AI Principle)*, https://oecd.ai/en/dashboards/ai-principles/P7 (last visited Sept. 23, 2026).

mechanisms should be maintained throughout the AI lifecycle, including records concerning the datasets, processes and decisions involved. This enables AI outputs and responses to be examined and assessed when required. Furthermore, AI actors should adopt a continuous and systematic approach to risk management across all stages of the AI lifecycle. Depending on their responsibilities and ability to influence the system, they should take appropriate measures to identify and mitigate risks associated with AI. This may involve cooperation among developers, technology and data providers, users and other relevant stakeholders. Such risks may include algorithmic bias, threats to human rights, safety and security concerns, privacy risks, and issues relating to labor and intellectual property rights.[6]

In India, this lack of clarity regarding accountability creates a serious gap in the legal system. Existing laws such as the IT Act, 2000 and the Digital Personal Data Protection Act, 2023 were implemented to address cybercrimes and data privacy issues but they fail to address the errors caused by an AI system's automated decision-making. As the existing laws do not view AI as a 'person', policymakers find it difficult to figure out who should be held responsible among the software developers, the company deploying the AI, and the end-users.

The research paper aims to examine the current legal landscape with respect to AI accountability in India and highlight the limitations of our existing statutes. The paper aims to analyze how India can close these regulatory gaps and also highlights the need to develop a clear legal framework that balances technological innovation with strict liability and mandatory human oversight.

2 BACKGROUND

At present, India currently lacks a comprehensive statutory framework specifically tailored to regulate artificial intelligence, leaving critical issues such as algorithmic bias, automated decision-making and the problem of explainability unaddressed. The IT Act, 2000 and the DPDP Act, 2023 are the primary legislation in India addressing digital governance. However, they do not directly address the legal complexities introduced by autonomous algorithms.

The IT Act, 2000 was enacted long before the advent of machine learning and primarily focuses on cybercrimes and safe harbor immunities for intermediaries, failing to explain AI decision-making and accountability issues. Similarly, while the DPDP Act, 2023 mandates structural compliances for data processing and user consent, it remains silent on automated profiling and fails to grant individuals the right to seek an explanation for automated decisions.[7]

Algorithmic opacity represents a major risk of these AI systems as they function as 'black boxes'; their decision-making process is frequently unexplainable even to their own developers which creates legal difficulties. If an individual falls victim to an erroneous automated decision – whether through a denied loan, a platform suspension or systemic profiling – a critical question remains: how can decisions made by AI algorithms be contested and what legal channels exist to challenge an invisible algorithm? AI systems are built and used by different actors. Where one person develops the base model, some other person is responsible for supplying training data, someone embeds the model in a service and another person uses the output in order to make decisions. In case of harm, the actors may state that the

[6] *Accountability (OECD AI Principle)*, https://oecd.ai/en/dashboards/ai-principles/P9 (last visited Sept. 23, 2026).
[7] *AI and Data Protection: Challenges in Automated Decision-Making*, IISPPR (Feb. 28, 2025), https://iisppr.org.in/ai-and-data-protection-challenges-in-automated-decision-making/.

responsibility lies somewhere else. Current laws are used to finding a single person at fault, so they fail to assign blame clearly in such cases.[8]

In the case of Pooja Ramesh Singh v. Jammu and Kashmir Bank Ltd. 2026, the NCLT relied on six fabricated case citations generated by an AI tool, which were later discovered and challenged before the Supreme Court. The case highlighted that AI-generated legal content must be independently verified by humans. Responsibility remains with the human decision-maker or lawyer using the AI output while submitting unverified AI-generated material may amount to professional misconduct, and reliance on such material can invalidate a judicial order.[9]

3 ANALYSIS

3.1 What Is Automated Decision-Making?

Automated decision-making involves the use of technology to make decisions autonomously without any human involvement in the process and is supervised by machine learning algorithms. AI systems gather data, analyze and evaluate it and make decisions solely.[10] Automated decision-making systems range from simple rule-based systems that follow pre-set rules, to machine learning models that use patterns from past data to make predictions about new cases, and deep learning systems that use complex layers of mathematical processing that can be difficult even for their developers to understand.[11]

Although the existing legal framework in India does not specifically define the term 'automated decision-making', section 2(b) of the DPDP Act, 2023 defines the term 'automated', as 'any digital process capable of operating automatically in response to instructions given or otherwise for the purpose of processing data'.[12]

AI and machine-learning systems can be difficult to understand because there is a lack of clarity regarding how they arrive at decisions. Organizations often build decision-making codes directly into their systems. Over time, these rules can spread across different systems, making them difficult to maintain.[13]

3.2 Existing Legal Framework in India

3.2.1 Information Technology Act, 2000

The IT Act is the principal legislative framework governing cyber law in India. The IT Act, 2000 was enacted with the main aim of promoting electronic commerce and addressing cybercrimes. The provisions of the act focus on traditional digital issues like hacking, identity theft, data breaches, etc. The act fails to address the complexities caused by AI systems and issues like algorithmic bias, automated

[8] Record Of Law, *Algorithmic Power and Legal Accountability: The Case for AI Regulation in India*, RECORD OF LAW (Aug. 2, 2026), https://recordoflaw.in/algorithmic-power-and-legal-accountability-the-case-for-ai-regulation-in-india/.
[9] Tulip Kanth, *Zero-Tolerance for Using AI-Generated Precedents Without Verification: Supreme Court Sets Aside NCLT, NCLAT Orders Citing Fake Judgments* (July 2, 2026), https://www.verdictum.in/supreme-court/pooja-ramesh-singh-v-jammu-and-kashmir-bank-ltd-2026-insc-668-ai-generated-precedents-1616998.
[10] *What Is Automated Decision-Making?* (Feb. 7, 2025), https://decisions.com/blog/what-is-automated-decision-making.
[11] IJLLR Journal, *Algorithm Bias, Automated Decision-Making, and the Right to Explanation: Comparative Analysis of India's DPDP Act 2023 and the EU AI Act 2024*, IJLLR JOURNAL (July 19, 2026), https://www.ijllr.com/post/algorithm-bias-automated-decision-making-and-the-right-to-explanation-comparative-analysis-of-ind.
[12] Digital Personal Data Protection Act, 2023, § 2(b).
[13] Sparkling Logic, *What Is Automated Decision-Making (ADM)?*, SPARKLING LOGIC (Oct. 11, 2024), https://www.sparklinglogic.com/what-is-automated-decision-making-adm/.

decision-making and lack of transparency; uncertainties regarding who should be held responsible for harm caused by AI remain unresolved.[14]

A major gap in the IT Act is the absence of any provision related to artificial intelligence, machine learning or automated systems. The Act fails to define these technologies nor does it outline their scope which results in ambiguity in statutory interpretation. In the absence of clear definitions, courts and regulatory authorities may face difficulties in applying existing legal provisions to AI-driven activities resulting in inconsistent interpretations and legal uncertainty. When AI systems function autonomously, there is difficulty in assigning responsibility in cases of harm as the IT Act does not clearly identify and state whether the liability in such cases should rest with the developers, manufacturers or users of such AI systems. The act does not impose specific obligations of transparency or explainability in relation to algorithmic decision-making processes. Consequently, individuals affected by automated decisions are not given a clear statutory right to challenge such decisions or obtain reasons for the outcomes, raising concerns regarding procedural fairness, accountability and due process.[15]

Algorithmic decision-making is often opaque, making it difficult for data subjects to understand how and why particular decisions are reached. The inability of such systems to provide meaningful explanations can further impede individuals from effectively exercising rights such as access, correction and objection. Fundamentally this lack of transparency raises concerns regarding fairness, accountability and due process.[16]

When harmful or misleading AI-generated content spreads online, determining who should be held responsible becomes difficult. The existing legal framework mainly provides remedies after harm has occurred such as content takedown, criminal action and civil remedies. However, these measures may not be sufficient to address risks such as deepfakes, automated misinformation and algorithmic discrimination, which can spread rapidly with little or no human involvement. As a result, gaps in accountability may arise, leaving affected individuals with limited and delayed remedies.[17]

While the IT Act, 2000 does not clearly address accountability issues caused by AI decision-making, there are certain provisions in the act which address general digital accountability:

- Section 43A – This section requires a body corporate to implement reasonable security practices to protect sensitive personal data but it does not regulate automated profiling or decision-making.[18]
- Section 79 – This section provides safe harbor protection for intermediaries, requiring them to practice due diligence, which has been expanded via IT Rules to cover deepfakes and synthetically generated content, but it still falls short of governing accountability of AI decision-making algorithms.[19]

[14] Tushar Sahu & Rajeev, *Inadequacy of the IT Act, 2000 in Governing Artificial Intelligence: Need for Legal Reform*, 8 IJFMR - INT'L J. FOR MULTIDISCIPLINARY RSCH. (2026), https://doi.org/10.36948/ijfmr.2026.v08i02.73091.
[15] *Id.*
[16] Gautam Sharma, *Artificial Intelligence and Data Privacy: Ethical and Legal Challenges*, VINTAGE LEGAL (Dec. 26, 2025), https://www.vintagelegalvl.com/post/artificial-intelligence-and-data-privacy-ethical-and-legal-challenges.
[17] Arpita Mishra, Dr Monica Yadav & Dr Pooja Batra Nagpal, *Autonomy, Accountability and Algorithms: India's Cyber Law at a Crossroad*, INT'L J. ADVANCES IN SIGNAL & IMAGE SCI. 2390 (2026), https://doi.org/10.29284/bjwegg22.
[18] IT Act, 2000, § 43A.
[19] *Id.* § 79.

3.2.2 Digital Personal Data Protection Act, 2023

The DPDP Act, 2023 is one of the most significant laws relevant to AI regulation in India currently. But the act does not specifically address automated decision-making of AI systems nor does it require AI systems to provide explanations for their decisions and also fails to provide individuals with a specific right to challenge decisions made by these systems. In contrast, Article 22 of the GDPR provides protection in relation to decisions based solely on automated processing. The absence of similar safeguards in India leaves gaps in algorithmic accountability and may limit the legal remedies available to individuals affected by AI-driven decisions.

Determining liability is a major issue in AI-driven decision-making. When an AI system makes an erroneous decision, like rejecting a job application, denying someone a loan or giving a completely wrong diagnosis of a patient, then who is responsible? Is it the developer, the deploying organization or the government that needs to be blamed? India does not have a dedicated AI-related law currently that assigns liability for AI-related harm. Some legal scholars support a 'human-in-the-loop' approach, under which AI-generated decisions are subject to meaningful human oversight, while others favor specific AI-liability frameworks that would hold AI developers and users legally responsible for errors, harmful decisions, and discriminatory outcomes resulting from the use of AI systems.

There is no separate chapter dedicated to AI in the act. However, the act does mention the processing of digital personal data and the term 'processing' under section 2(x) of the act is defined broadly enough to include automated operations involving digital personal data.[20] The DPDP Act does not contain a standalone provision that expressly prohibits organizations from relying exclusively on automated decision-making. However, such decision-making may still be subject to various provisions of the Act.

Section 8(3) of the DPDP Act requires a data fiduciary to ensure that personal data is complete, accurate and consistent where such data is likely to be used to make a decision affecting a data principal or is intended to be disclosed to another data fiduciary.[21] This provision is particularly relevant to AI systems deployed in contexts involving consequential decisions, including recruitment and employment, credit assessment, insurance underwriting, eligibility determinations, customer profiling and fraud detection. Accordingly, organizations should ensure that personal data used by AI systems is of appropriate quality and is reliable particularly where such data may inform decisions that significantly affect individuals.[22]

The DPDP framework has some relevance to the use of AI and automated decision-making, although it doesn't establish a standalone regulatory regime governing automated decisions nor does it contain any provision relating to the right to demand human review of an automated decision. The act also does not recognize the rights of data principals to obtain an explanation for decisions made through AI-driven systems unlike the GDPR framework which contains more extensive transparency requirements in relation to automated decision-making.

The DPDP Act under section 10(1) enables the Central Government to designate certain data fiduciaries, or classes of data fiduciaries as significant data fiduciaries (SDFs), based on factors such as the volume and sensitivity of personal data processed, the risks to data principals, and the potential impact

[20] Digital Personal Data Protection Act, 2023, § 2(x).
[21] *Id.* § 8(3).
[22] Dhruv Kaushal, *How Does the DPDP Act Regulate AI Training in India?*, KING STUBB & KASIVA (Sept. 9, 2026), https://ksandk.com/data-protection-and-data-privacy/dpdp-act-ai-training-automated-decision-making/.

on matters including the sovereignty and integrity of India, electoral democracy, security of the state and public order. SDFs are subject to enhanced compliance requirements which may become particularly relevant where AI or algorithmic systems are used to process personal data.[23]

Under Rule 13 of the DPDP Rules, SDFs must conduct a Data Protection Impact Assessment (DPIA) and an audit every 12 months and submit reports containing significant observations to the Board. They must also exercise due diligence to ensure that the technical measures they employ including algorithmic software used for specified processing activities are not likely to pose a risk to the rights of data principals.[24] Accordingly, where an AI system forms part of a data-processing operation falling within these requirements, the DPIA, audit and algorithmic due diligence operations may operate as mechanisms for identifying and mitigating risks arising from such systems. However, these provisions do not directly address accountability of AI systems in the decision-making process. The act is relevant primarily through the broader regulation of personal data processing and the additional risk-governance obligations imposed on SDFs.

In 1950, Alan Turing asked the famous question, "Can Machines Think?" in his paper 'Computing Machinery and Intelligence'. This question raises concerns regarding accountability and responsibility. While machines are responsible for executing algorithms, the responsibility for algorithmic outcomes rests with human designers of such systems. The principles of transparency and human oversight in automated decision-making reflect the idea that responsibility for a machine's decisions ultimately rests with the organization that deploys and controls it.[25]

3.2.3 India AI Governance Guidelines, 2025

The guidelines mention seven guiding principles also referred to as 'sutras' which aim to regulate India's AI governance. The fifth principle refers to the principle of 'accountability' which states that AI developers and deployers must operate transparently and remain accountable for the systems they create or use. Accountability should be allocated according to the role performed, the potential risk of harm and the applicable due diligence requirements. Accountability can be supported through a combination of regulatory, technical and market-based mechanisms.[26]

These guidelines mention accountability as the backbone of AI governance. There exists a lack of clarity regarding how liability should be allocated across developers, deployers and end-users. There is a lack of grievance redressal mechanisms available for users. Also, transparency in the design of AI systems, the flow of data and the decision-making processes of organizations remains limited. The probabilistic and adaptive nature of AI may lead to unforeseen outcomes. This creates a need for a governance framework that ensures appropriate oversight and accountability while allowing sufficient space for responsible innovation. The Committee in its recommendations mentions that there is a need to introduce graded obligations and liability frameworks that are proportionate to the functions performed by AI actors, the risks associated with their activities and the extent of due diligence undertaken. Enforcement and

[23] *Id.*

[24] Digital Personal Data Protection Rules, 2025, r. 13.

[25] Adv (Dr) Prashant Mali, *AI and Machine Learning Under DPDPA: Compliance Guide for Organizations*, DPDPA.COM, https://www.dpdpa.com/blogs/ai_machine_learning_dpdpa_compliance_guide.html (last visited Sept. 12, 2026).

[26] *India AI Governance Guidelines*, https://www.pib.gov.in/Pressreleaseshare.aspx?PRID=2228315 (last visited Sept. 13, 2026).

accountability need to be strengthened through mechanisms like transparency reporting, independent audits, and self-certification requirements.[27]

The AI Governance Committee mentions certain practical guidelines for industry and innovators in order to enable responsible implementation of the AI Governance Framework and recommends that all persons involved in the development or deployment of AI systems in India should adhere to the following principles:

- AI developers and deployers should comply with all the applicable laws and regulations in India, including those relating to information technology, data protection, copyright, consumer protection, and the protection of women, children and other vulnerable groups.
- Show compliance with applicable laws and regulations when required by authorities or regulators.
- Adopt voluntary principles, codes and standards to promote privacy, security, fairness, inclusivity, non-discrimination, and transparency, along with appropriate technical and organizational safeguards.
- Set up a grievance system to report AI-related harms and resolve complaints within a reasonable time.[28]
- Publish transparency reports assessing the risks AI systems may pose to individuals and society in India. Sensitive or confidential information should be shared privately with the relevant regulators.
- Make use of technology and legal measures to reduce AI risks, such as privacy tools, machine unlearning, algorithmic audits, and automated bias detection.[29]

3.2.4 Sector-Specific Rules

RBI Model Risk Management Guidance, 2026

The Central Bank released its draft, 'Guidance on Regulatory Principles for Model Risk Management, 2026' which are a set of detailed guidelines for models adopted by banks, NBFCs, and other regulated institutions and also includes AI and machine learning systems. These draft guidelines require organizations to implement model risk management frameworks after being approved by the Board, validate models independently, maintain inventories of algorithms that are used for decision-making and set up human oversight regarding AI-driven decisions. According to RBI, a model includes any system that uses data and analytical techniques (includes AI and ML), in order to produce outputs to be used in decision-making and business operations. This definition also includes algorithms, applications, analytics, spreadsheet-based tools and decision-based rules if they impact business decisions like customer pricing and lending rates.[30]

The draft guidance has also mentioned certain requirements for AI and ML systems which are: assessing risks arising from hallucinations, bias, discrimination in outcomes, adversarial attacks, data drift, etc. Entities are also required to test models under certain conditions and implement required safeguards. RBI has also suggested that regulated entities should establish mandatory human oversight for AI models

---

[27] *Id.*

[28] *Decoding the India AI Governance Guidelines - Saikrishna & Associates*, SAIKRISHNA & ASSOCIATES - A TIER-1 FULL SERVICE FIRM (Nov. 10, 2025), https://www.saikrishnaassociates.com/decoding-the-india-ai-governance-guidelines/.

[29] *Id.*

[30] www.ETBFSI.com, *RBI Proposes AI Governance Framework for Banks, Mandates Human Oversight of Model Driven Decisions*, ETBFSI.COM, https://bfsi.economictimes.indiatimes.com/articles/rbi-unveils-ai-governance-framework-mandating-human-oversight-in-banking/131970975 (last visited Sept. 23, 2026).

especially in cases involving automated decision-making by AI models. A complete human-in-the-loop approach, override capabilities and mechanisms for system suspension and deactivation, including kill switches, need to be implemented in order to ensure effective human oversight and control over AI systems. The Central Bank has also cautioned regarding automation bias and overdependence on model outputs.[31]

SEBI (Intermediaries) (Amendment) Regulations, 2025

'Regulation 16C' was inserted under these regulations, which makes any SEBI-regulated entity solely responsible for any AI/ML tools used by them, whether it has been developed in-house or externally. This regulation was notified in February, 2025 and closely follows SEBI's consultation paper of November, 2024, which proposed that every person regulated by SEBI that uses AI shall be solely responsible for protection and security of investor data, responsibility of any AI-generated output it relies on, and compliance with all applicable laws. This rule applies regardless of the extent of AI use, meaning that even limited or third-party use can result in full responsibility. SEBI may also take enforcement action and impose sanctions for any violation under its general powers.[32]

Previously, AI-related risks could fall into regulatory grey areas, particularly when determining whether responsibility lies with the technology provider or the user. Regulation 16C removes this uncertainty by placing clear responsibility and liability on intermediaries for all AI systems used in their operations. Another concern is regarding 'auditability'. Although intermediaries bear responsibility for AI-related outcomes, regulators still need sufficient visibility into how these systems function. This can be challenging because advanced AI systems, particularly complex machine-learning and generative AI models, often operate as 'black-boxes', making their decision-making processes difficult to understand or inspect. If SEBI is unable to examine the model's logic or data flows, placing full liability on the intermediary may create practical difficulties.[33]

3.2.5 Artificial Intelligence (Ethics and Accountability) Bill, 2025

The Bill was introduced in the Lok Sabha in December, 2025 and is the first dedicated bill with respect to AI. The Bill establishes a legal 'right to explanation' for individuals affected by automated decisions that have significant legal or commercial consequences. It further introduces the concept of 'transparency by design', requiring developers to document aspects such as the model's architecture, sources of training data, and measures adopted to reduce bias before the system is deployed. The Bill also aims to regulate developers, deployers and users and mentions a civil penalty of Rs. 5 Crore for non-compliance with developer or deployer duties. The Bill also provides compensation to individuals who can establish that they have suffered harm as a result of an algorithmic system. Despite these provisions, the Bill fails to

[31] *Press Releases | Official Website of Reserve Bank of India*, https://www.rbi.org.in/Scripts/bs_viewcontent.aspx?Id=5089 (last visited Sept. 23, 2026).
[32] *SEBI | Securities and Exchange Board of India (Intermediaries) (Amendment) Regulations, 2025*, https://www.sebi.gov.in/legal/regulations/feb-2025/securities-and-exchange-board-of-india-intermediaries-amendment-regulations-2025_91809.html (last visited Sept. 23, 2026).
[33] *SEBI's AI Liability Regulation: Accountability and Auditability Concerns*, THE HNLU CCLS BLOG (Oct. 3, 2025), https://hnluccls.in/2025/10/03/sebis-ai-liability-regulation-accountability-and-auditability-concerns/.

determine the allocation of liability for AI-generated content, leaving uncertainty as to whether responsibility should rest with the developer, deployer or user.[34]

4 CONCLUSION & SUGGESTIONS

AI is changing how decisions are made, creating new challenges for legal accountability in India. While there is no specific law regulating artificial intelligence in India, existing laws may address certain AI-related harms, yet they do not provide a clear framework for automated decision-making. India therefore needs a balanced approach that clearly assigns responsibility to AI developers and deployers, ensures transparency and human oversight, and manages risks without slowing responsible innovation.

While the AI Governance Guidelines address the accountability issue of automated decision-making systems to a certain extent, there is a need for an AI-specific law which clearly allocates liability and accountability among the developers, deployers and users of an AI system along with clearly mentioning their responsibilities. A human-in-the-loop approach should be followed to ensure that a designated person retains the authority to review, question or challenge AI-generated recommendations. There is also a need to create traceable workflows, and accountability for automated decisions should be established by assigning clear human ownership and defining explicit decision-making processes, rather than attributing responsibility to the algorithm itself.[35]

India could move towards a structured AI governance framework through proposed legislative measures such as the Digital India Act, which may introduce risk-based regulation, greater transparency, and accountability for AI and emerging technologies. The proposed Artificial Intelligence (Ethics and Accountability) Bill, 2025, also reflects growing interest in stronger safeguards including ethical reviews, bias audits, developer responsibilities, and grievance mechanisms. Although these mechanisms are not yet an established law, they indicate a shift towards clearer legal accountability for AI systems while seeking to support innovation.[36]

* * *

[34] Dhruv, *AI Ethics and Accountability Bill 2025 Explained*, IPLEADERS (May 14, 2026), https://blog.ipleaders.in/ai-ethics-and-accountability-bill-2025-indias-first-dedicated-ai-bill-explained/.

[35] *Build Accountability into AI to Drive Business Value | TechTarget*, AI & EMERGING TECH, https://www.techtarget.com/ai/tip/Build-accountability-into-AI-to-drive-business-value (last visited Sept. 23, 2026).

[36] *AI Laws and Regulations in India as of 2026*, https://www.prashantmali.com/cyber-law-blog-india/ai-laws-and-regulations-in-india-as-of-2026 (last visited Sept. 13, 2026).

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.9

CHAPTER 9

Artificial Intelligence and Accountability of Automated Decisions

GAURANGI MEHROTRA*

Artificial intelligence (AI) has rapidly altered how decisions are made in a number of industries, including public administration, healthcare, finance, employment, education, and criminal justice. Automated decision-making systems are being used more and more to analyze massive amounts of data, spot trends, forecast results, and make choices that directly impact people's rights and interests. When automated decisions lead to prejudice, inaccuracy, or harm, these technologies raise serious questions about accountability and legal responsibility even though they are efficient, consistent, and economical.

Determining who should be held responsible for judgments made or impacted by AI systems is the primary challenge. While AI systems may function with varied degrees of autonomy and complexity, traditional legal frameworks often assign liability to human actors. Affected parties find it challenging to comprehend how a certain decision was made or to successfully contest it due to the lack of transparency, which is also referred to as the "black box" issue. The attribution of culpability is further complicated by algorithmic prejudice, insufficient data, and unpredictable system behavior.

The ethical and legal aspects of accountability in automated decision-making are examined in this research. While taking into account the concepts of transparency, explainability, fairness, human oversight, and practical solutions, it examines the obligations of AI developers, technology suppliers, deployers, organizations, and human decision-makers. According to the study, AI shouldn't be used as a way to evade legal obligations. Rather, a thorough accountability structure should include significant human oversight and set up precise procedures for accountability, review, and remedy. In the end, the report promotes a human-centered strategy that balances technology innovation with institutional accountability, the rule of law, and fundamental rights.

Keywords: Artificial intelligence, automated decision-making, accountability, algorithmic transparency, AI governance

1 OVERVIEW

The study and creation of intelligent beings that observe their surroundings and take actions that increase their chances of success is the focus of artificial intelligence (AI), a subfield of computer science. "The ability to hold two different ideas in mind at the same time and still remain the ability to function" is one definition of artificial intelligence.[1]

AI is being incorporated more and more into financial services, e-commerce, banking, healthcare, education, policing, recruitment, governance, and other facets of social and economic life in India. There is now a lot of potential to improve public administration and efficiency due to the growing use of AI. Simultaneously, it has raised challenging legal issues regarding accountability for decisions made by automated systems.

* Advocate.

[1] Gyanendra Singh, Ajitanshu Mishra & Dheeraj Sagar, *An Overview of Artificial Intelligence*, 2 SBIT J. SCI. & TECH. 1 (2013).

When an automated judgment impacts a person's legal rights, liberty, dignity, means of subsistence, or economic interests, the issue becomes very significant. An AI-based system might deny a loan application, flag someone as a possible fraudster, rate employment candidates, assess a person's eligibility for a service, help with medical diagnostics, or have an impact on political decisions. The impacted individual must have a meaningful way to find out why the decision was taken and who is accountable if it is unfair or discriminatory.

The foundation of the conventional legal system is institutional or human accountability. The law establishes the repercussions when an individual conducts an act, an authority makes a decision, or a business renders a service. This structure is complicated by AI because the final result may be produced by algorithms that have been trained on enormous datasets and may involve multiple parties, such as developers, data providers, technology businesses, governmental bodies, and end users. This leads to what could be called an accountability gap.

Therefore, the crucial question is not just whether AI should be allowed to make judgments, but also whether these decisions can be integrated into India's current legal and constitutional framework without jeopardizing fundamental rights.

An essential place to start is the Indian Constitution. Equal protection under the law and equality before the law are guaranteed under Article 14. While Article 21 safeguards life and individual liberty, Article 19 guarantees certain freedoms. These clauses have been read by Indian constitutional law in a way that prioritizes proportionality, non-arbitrariness, fairness, dignity, and privacy.

This paper's goals are to determine whether India's current legal system can guarantee accountability for judgments made by AI and to pinpoint the fundamental ideas needed to create an all-encompassing Indian AI accountability system.

2 AUTOMATED DECISION-MAKING: MEANING AND NATURE

Systems, software, or procedures that use computation to support or replace government decisions, judgments, and/or policy implementation that affect opportunities, access, freedoms, rights, and/or safety are referred to as "automated decision systems." Predicting, categorizing, optimizing, identifying, and/or recommending are all possible tasks for automated decision systems.[2]

In general, automated decision-making refers to the use of computer-based systems to analyze data and generate conclusions, suggestions, or categorizations pertaining to a person or group.

Automated processing is acknowledged under the Digital Personal Data Protection Act, 2023. According to the Act, "automated" refers to a digital process that can process data or respond to instructions automatically.[3]

There are many levels of automated decision-making.

1. *Completely Automated Making of Decisions* - Information is processed by the system and the outcome is generated without significant human involvement.
2. *AI-Powered Decision-Making* - AI generates a prediction or recommendation, but the ultimate decision is made by a human authority.

[2] Anna Katharina Boos, *Conceptualizing Automated DecisionMaking in Organizational Contexts*, 37 PHIL. & TECH. 2 (2024).
[3] The Digital Personal Data Protection Act, 2023, § 2(b).

3. *AI-Assisted Human Decision Making* - AI only helps the human decision-maker by arranging data or spotting trends.

The distinction has legal implications. Ordinary legal duty can usually stay with the person using the decision-making capacity when AI only helps. However, more accountability measures are required when an organization successfully transfers decision-making to an automated system. Automation bias is a major issue. An algorithmic recommendation may be accepted by human decision-makers without independent review because they believe it to be intrinsically impartial. Therefore, having a human present at the conclusion of an automated process does not always equate to true human oversight.

2.1 Algorithmic Accountability

AI is being used more and more in hiring, worker classification, risk assessment, and other labor choices. However, accountability may get dispersed across developers, vendors, deployers, and organizations when AI makes mistakes. There should always be a clearly identified authority in charge of the result, regardless of whether an algorithm is used or outsourced.

Algorithmic responsibility requires transparency, precise data, human oversight, audits, and clear ownership. Developers should be held responsible for predictable design defects, while deployers and organizations must ensure appropriate use, accurate data, monitoring, evaluation, and effective remedies. Human officials must be able to contest or veto decisions made by AI.

AI should ultimately be viewed as a governance and compliance concern rather than just a technical instrument. Organizations should keep accurate records of their data, decision-making procedures, human evaluations, and remedial measures. The basic idea is that while AI can help with decision-making, it should never take the role of human accountability or the capacity to clarify and correct decisions.

2.2 Algorithmic Accountability and Article 14

Perhaps the most significant constitutional clause governing discriminatory or unfair automated decisions is Article 14. The ruling in *E.P. Royappa*[4] by the Supreme Court held that arbitrary State action is inherently incompatible with equality. This idea was further expanded by the Court in a number of instances.

The connection to AI is simple. Because an algorithm applies the same mathematical procedure to everyone, it may appear neutral. However, equality is not always the result of mathematical neutrality. An AI hiring system that has been educated on past job data, for instance, might replicate past discriminatory trends. The AI might pick up on and replicate a pattern if prior hiring decisions consistently favored a particular group of candidates. Similarly, if the underlying data is erroneous or incomplete, an algorithm that determines a group's eligibility for government benefits may disproportionately exclude that group. These results could give rise to Article 14 issues.

Therefore, rather than only focusing on whether discriminatory instructions were purposefully included into the algorithm, the constitutional test should consider the decision's impact and rationale.

The concept of manifest arbitrariness is also affected by the Supreme Court's ruling in *Shayara Bano v. Union of India*.[5] The doctrine's wider focus on arbitrariness offers a crucial conceptual foundation for analyzing automated governmental decision-making, despite the fact that it was created in the context of

[4] E.P. Royappa v. State of Tamil Nadu, (1974) 4 SCC 3.
[5] Shayara Bano v. Union of India, (2017) 9 SCC 1.

legislation. The simple fact that "the computer made the decision" shouldn't make an AI system constitutionally acceptable. The decision made by the State remains an act of the State even if it uses an algorithm to wield public power.

2.3 Procedural Fairness and Article 21

Because automated judgments may have a direct impact on life, liberty, and dignity, Article 21 is equally important. The Supreme Court ruled in *Maneka Gandhi v. Union of India* that the process envisioned under Article 21 cannot be capricious, unjust, or irrational.[6]

AI is significantly impacted by this idea. Let's say an automated system mistakenly determines that a person is ineligible for a government subsidy. The judgment may create major procedural-fairness issues if the person is not given an explanation or a chance to contest the outcome. Thus, the right to be heard, the audi alteram partem principle, is pertinent.

Typically, an individual impacted by a big automated decision should have:

1. Notification of the ruling
2. Details about automated processing
3. Access to pertinent personal data
4. A clear justification of the decision's main rationale
5. A chance to rectify false information
6. A significant human evaluation
7. A grievance or appeal process.

The severity of the repercussions should determine how strong these protections are.

A recommendation made by AI regarding a person's tastes for entertainment does not need to be protected in the same way as an automated choice that affects work, liberty, welfare benefits, healthcare, or access to basic services.

3 AI AND PRIVACY RIGHTS

Because contemporary AI systems rely largely on personal data, the relationship between AI and privacy is especially crucial in India. The Supreme Court unanimously acknowledged privacy as a fundamental right in *Justice K.S. Puttaswamy (Retd.) v. Union of India*.[7] The ruling linked privacy to liberty, autonomy, and dignity. Informational privacy has been treated as constitutionally significant in the Court's subsequent jurisprudence.

AI raises privacy issues in a number of ways:

1. *Gathering Information – For* training and operation, AI systems need a lot of data. Information privacy may be compromised by excessive collecting.
2. *Data Processing* - It is possible that information provided for one purpose will be handled for another.
3. *Profiling* - AI is able to construct profiles about people by combining various kinds of information.
4. *Inference* - AI has the ability to deduce sensitive traits from seemingly unremarkable data.
5. *Maintaining* - Additional concerns arise when personal data is stored for an extended period of time.

[6] Maneka Gandhi v. Union of India, (1978) 1 SCC 248.
[7] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

Therefore, the fundamental right to privacy necessitates taking into account not only what data is gathered but also how it is analyzed and what judgments are taken based on that analysis. In particular, the proportionality principle is pertinent. A restriction on privacy must be supported by the law, have a justifiable goal, be logically related to that goal, and refrain from unduly interfering with people's rights.

3.1 The 2023 Digital Personal Data Protection Act

An essential part of India's existing regulatory framework for data governance is the Digital Personal Data Protection Act, 2023[8] ("DPDP Act"). The Act aims to control how digital personal data is processed while taking into account both legal processing needs and individual data protection interests.[9] According to the Act, a Data Fiduciary is an individual who, either alone or in concert, chooses how and why to process personal data.[10] This idea is especially pertinent to AI since the organization that decides how and why personal data is processed will often have significant control over an AI system.

The Act offers a number of rights that are pertinent to automated decision-making, such as the right to access personal data information, the right to have personal data corrected and erased, the right to file a grievance, and related rights and obligations for the processing of personal data.[11] The Act also creates the Data Protection Board of India and stipulates penalties and adjudication.[12] The Act has limited but important implications for AI. AI accountability is not the same as data protection.

Even if an AI system processes personal data in compliance with relevant data-protection regulations, it may nonetheless produce a judgment that is discriminatory or unfair. As a result, the DPDP paradigm ought to be viewed as a part of a larger framework for AI accountability. Both are necessary for a thorough system: Data accountability: Was the processing of personal data done in a responsible and legal manner? Decision accountability: Was the choice made using such information impartial, truthful, non-discriminatory, and subject to appropriate review?

3.2 The Information Technology Act of 2000 and the Accountability of Artificial Intelligence

An essential component of India's digital legal system is still the Information Technology Act, 2000[13] (the "IT Act"). The IT Act's provisions regarding electronic records, intermediary responsibilities, cybersecurity, and computer-related offenses may apply to AI-enabled activities depending on the situation, even though it was passed prior to the widespread use of modern generative AI and sophisticated machine-learning systems. Nevertheless, the IT Act is not a comprehensive law pertaining to AI accountability.

A comprehensive framework governing algorithmic bias, explainability, AI effect assessments, human monitoring, automated governmental choices, algorithmic audits, and full culpability for actions made by AI is lacking. The necessity for sector-specific and cross-sector AI governance concepts is demonstrated by this regulatory gap.

[8] The Digital Personal Data Protection Act, 2023.
[9] *Id.* pmbl.
[10] *Id.* § 2(i).
[11] *Id.* §§ 11, 12, 13.
[12] *Id.* §§ 18–34.
[13] The Information Technology Act, 2000.

4 THE "RIGHT TO EXPLANATION" AND NATURAL JUSTICE

The need for natural justice is one of the most significant tenets of Indian administrative law. One of the fundamental tenets of administrative and constitutional law is the idea of natural justice, which stands for the legal system's moral conscience. It embodies the values of justice, equity, and fairness that direct judicial and quasi-judicial processes.[14] Generally speaking, a person impacted by an administrative decision should be given a fair chance to argue their position, especially if the decision negatively impacts their rights or legitimate interests. This principle is strongly related to the duty to give reasons. Three goals are served by a well-reasoned decision:

1. It shows that the authority used its judgment.
2. It enables the impacted person to comprehend the rationale for the choice.
3. It makes judicial review easier.

If the authority only claims that an AI system produced the result, automated decision-making could compromise these goals. Consequently, Indian law ought to acknowledge the notion of meaningful explanation in important circumstances. It is not necessary to reveal the entire source code in order to provide a meaningful explanation. Rather, an impacted person should typically be able to learn:

- That AI was used
- The AI system's purpose
- The main factors taken into account
- The pertinent data relied upon
- Whether a human reviewed the outcome
- The system's limitations, if applicable
- The procedure for contesting the result.

This strategy safeguards legitimate interests in cybersecurity and intellectual property as well as transparency.

5 INDIAN EQUALITY LAW AND ALGORITHMIC BIAS

When systematic mistakes in machine learning algorithms result in unjust or discriminating outcomes, this is known as algorithmic bias. Algorithmic bias is the result of AI systems reproducing and maybe amplifying social prejudices and inequities in their decision-making processes after being educated on historical data that reflects these biases.[15] One of the biggest risks associated with automated decision-making is algorithmic bias.

Bias can come from:

1. *Training Information – Incomplete* or unrepresentative training data could lead to uneven results from the final system.
2. *Discrimination Based on History – Discriminatory* patterns ingrained in past decisions can be replicated by AI.

[14] Dr. Sudhir Kumar Dubey & Dr. Rang Nath Singh, *Doctrine of Natural Justice: Evolution, Principles, and Application in Indian Legal System*, 4 INT'L J. JUD. L. 24–27 (2025).

[15] Ramandeep Kaur, *Algorithmic Bias and Constitutional Safeguards in the Indian Judiciary: A Critical Analysis of AI Integration in Legal Adjudication*, 8 INT'L J.L. MGMT. & HUMAN. 2216 (2025).

3. *Proxy Variables – Indirect* correlations between seemingly neutral variables and legally or constitutionally sensitive traits are possible.
4. *Quality of Data – Erroneous* conclusions might result from erroneous or out-of-date information.
5. *Implementation – Applying* a system designed for one demographic to another may result in unreliable outcomes.

In India, this problem becomes particularly important because of the constitutional commitment to equality and social justice. AI systems used by public authorities should therefore be evaluated not merely for overall accuracy but also for differential impact upon different groups. The State should be required to demonstrate that a high-risk AI system is based on rational criteria and does not produce unjustified discriminatory outcomes.

5.1 Automated Government Decisions and Proportionality

According to the proportionality doctrine, AI-based government actions that impact basic rights must have a legal foundation, a legitimate purpose, a rational connection, and necessity, while weighing the advantages to the public against the risks to individuals. As acknowledged in *K.S. Puttaswamy (Retd.) v. Union of India*,[16] less restrictive options and privacy issues must be taken into account before adopting AI only because it is technologically feasible.

5.2 Human Supervision and Responsibility

AI-based decision-making requires meaningful human oversight. Officials must be able to obtain pertinent data, question, review, and overturn AI outputs, document justifications, and maintain legal accountability. In areas like policing, welfare, healthcare, employment, and taxation where AI judgments have a substantial impact on fundamental rights, human review should become more important.

5.3 AI Developers' and Deployers' Liability

Depending on variables like control, foreseeability, negligence, and causation, AI developers may be held accountable for careless design, insufficient testing, security lapses, predictable hazards, or discriminating systems. Deployers must make sure AI is appropriate for its intended use, employs correct data, is checked for discrimination, and offers human oversight and remedies. Therefore, each actor's job and level of control should determine their level of responsibility.

5.4 Accountability of the Government

Even when private corporations build AI systems, government authorities are nonetheless accountable under the constitution. Legality, equity, nondiscrimination, privacy, proportionality, and efficient remedies must all be guaranteed. Constitutional responsibility cannot be outsourced in the same way as AI technology.

In the private sector, companies like Microsoft (Copilot) and Salesforce (Einstein) are creating AI capabilities that allow businesses to coordinate decision-making across functions at the enterprise level, while enterprise software providers like SAP and Oracle integrate finance, logistics, procurement, and human resources within unified systems. Although this integrated paradigm has not yet been fully

[16] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

embraced by public administration, the direction is obvious. The next phase of change will take place at the level of governance architecture rather than at the level of specific instruments.[17]

6 IMPORTANT JUDICIAL DECISIONS IN INDIA

6.1 E.P. Royappa v. State of Tamil Nadu[18]

The Supreme Court linked non-arbitrariness and equality. Because an algorithmically generated judgment might still be arbitrary, the principle is applicable to automated decision-making. Technology cannot change a state's arbitrary actions into ones that are lawfully permissible.

6.2 Maneka Gandhi v. Union of India[19]

The Court mandated fairness and rationality in processes impacting individual liberty and rejected a limited interpretation of Article 21 (the connection between Articles 14, 19, and 21). The case backs up the idea that meaningful procedural protections should be provided to anyone impacted by important automated government decisions.

6.3 Justice K.S. Puttaswamy (Retd.) v. Union of India[20]

The nine-judge bench acknowledged privacy as a basic right. Because automated systems rely on the gathering, processing, analysis, and inference of personal data, the ruling is especially pertinent to artificial intelligence. The Court acknowledged the intimate relationship between privacy and autonomy and dignity.

6.4 Shayara Bano v. Union of India[21]

A helpful framework for analyzing irrational or arbitrary State behavior is provided by the Supreme Court's consideration of manifest arbitrariness. Similar worries can be raised by an automated decision that yields results that are illogical or irrational, especially when basic rights are at stake.

6.5 Anuradha Bhasin v. Union of India[22]

The Supreme Court focused on legality, reasonableness, and proportionality while examining limitations on basic freedoms in relation to internet restrictions.

6.6 Internet & Mobile Association of India v. Reserve Bank of India[23]

The Supreme Court examined regulatory limitations pertaining to virtual currencies in applying proportionality standards. The case illustrates the need for constitutional principles to strike a balance between technological progress and regulatory authority.

---

[17] Michael A. Santoro, *Where Is Accountability When Governments Deploy AI?*, TECHPOLICY.PRESS (Mar. 30, 2026), https://www.techpolicy.press/where-is-accountability-when-governments-deploy-ai/.
[18] E.P. Royappa v. State of Tamil Nadu, (1974) 4 SCC 3.
[19] Maneka Gandhi v. Union of India, (1978) 1 SCC 248.
[20] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[21] Shayara Bano v. Union of India, (2017) 9 SCC 1.
[22] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.
[23] Internet & Mobile Ass'n of India v. Reserve Bank of India, (2020) 10 SCC 274.

7 DEVELOPMENTS IN INDIAN REGULATION REGARDING AI

Instead of having a single comprehensive AI law, India has historically adopted a technology-friendly approach to AI, depending on innovation, responsible AI, and sector-specific regulation. Constitutional law, data protection, information technology laws, administrative law, judicial review, and government policies are all included in its regulatory framework. The government is putting more emphasis on responsible, secure, and reliable AI governance, especially through MeitY. Instead of depending only on current general rules, India requires a clear accountability framework that guaranties efficient oversight as the usage of AI grows.

Under the IndiaAI Mission, MeitY published the India AI Governance Guidelines in November 2025. Seven fundamental ideas—often referred to as "sutras"—form the basis of the guidelines:

1. Trust as the foundation
2. A human-centric (people-first) perspective
3. Responsible innovation, or innovation over restraint
4. Justice, equity, and nondiscrimination
5. Accountability
6. Transparency and comprehensibility based on design
7. Security, safety, and sustainability[24]

8 AI ACCOUNTABILITY: PROPOSED INDIAN FRAMEWORK

India ought to think about implementing a thorough, risk-based framework built on the following tenets.

1. *Required AI Impact Evaluation* - Before being deployed, high-risk AI systems should go through an impact assessment. The evaluation ought to take into account confidentiality, equality, prejudice, precision, safety, human rights, possible harm to society, and whether remedies are available.
2. *Algorithmic Audits* - For high-risk systems, independent audits ought to be carried out on a regular basis. Both technical performance and legal compliance should be looked at during the audit.
3. *Required Human Evaluation* - When a choice has a significant impact on basic rights, AI shouldn't have the last say.
4. *The right to a meaningful explanation* - Reasonable explanations for important automated decisions should be provided to affected parties.
5. *The Right to Contest* - A person ought to be able to contest a decision made by AI in front of a separate human authority.
6. *Accuracy of Data* - When making important decisions, organizations should confirm the relevance and quality of the data.
7. *Testing for Algorithmic Equality* - It is important to examine high-risk systems for discriminating results.
8. *Audit Trails* - Important AI systems ought to save enough documentation to reconstruct the decision-making process.

[24] INDIA AI GOVERNANCE GUIDELINES.

9. *Responsibility Distribution* - The obligations of developers, deployers, and data providers should be specified in contracts. However, statutory or constitutional requirements should not be eliminated by contractual agreements.
10. *Remedies and Compensation* - Effective remedies should be available to those who suffer legally recognized injury as a result of unlawful or careless AI deployment.

9 DIFFICULTIES IN INDIA WITH AI ACCOUNTABILITY

A number of issues need to be resolved.

- Absence of an All-Inclusive AI-Specific Law - India currently heavily depends on current laws and emerging regulatory frameworks.
- Complexity of Technology - It's possible that judges, attorneys, and regulators lack the specialized technical expertise needed to assess complicated AI systems.
- Trade Secrets - Businesses may object to algorithm publication on the grounds of intellectual property protection.
- Insufficiently Standardized Auditing - For algorithmic testing and certification, uniform criteria are required.
- Dispersed Accountability - Developers, data producers, cloud service providers, and deployers may be involved in AI systems.
- Quick Changes in Technology - As AI technology advances, laws may become out of date.
- Harmonizing Regulation and Innovation – While under-regulation may expose people to grave rights breaches, over-regulation may deter technical advancement.

These difficulties show why India should embrace adaptable values as opposed to strict technology regulations.

9.1 The Necessity of a Human-Centered Strategy

AI governance must guarantee that technology upholds constitutional principles and does not take the role of human accountability. AI can help decision-makers, but it cannot be used to evade responsibility. People who are impacted by AI choices need to be able to recognize the responsible authority, comprehend the rationale behind the decision, and request a meaningful review. Therefore, in order to prevent technological efficiency from superseding fundamental rights, AI governance in India must safeguard equality, liberty, dignity, and the rule of law.

10 CONCLUSION

By enhancing productivity, data analysis, and governance, artificial intelligence is revolutionizing decision-making in India. However, automated decisions may also be opaque, discriminating, erroneous, and challenging, raising significant accountability issues. Important protections are provided by the Indian Constitution, especially Articles 14, 19, and 21, natural justice, judicial review, and the right to privacy acknowledged in Puttaswamy. Although data privacy is further strengthened by the Digital Personal Data Protection Act, 2023, a more comprehensive framework for automated decision-making is required. Transparency, explainability, human oversight, nondiscrimination, audits, and efficient remedies should

all be guaranteed by such a system. In the end, AI should continue to be governed by human accountability and the rule of law, guaranteeing that technology advancement fosters efficiency without jeopardizing equality, privacy, dignity, or fundamental rights.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.10

CHAPTER 10

Beyond the Black Box: A Multi Tiered Framework for Algorithmic Accountability and Liability in Autonomous Decision Making Systems

LAKSHMI NIVEDITHA S[*] AND ANNA MARIAM IYPE[**]

As artificial intelligence systems become increasingly capable of making decisions with limited human intervention, questions of responsibility and accountability have become more difficult. This paper examines two major challenges created by autonomous decision-making systems: the "black box" problem, where the reasoning behind an AI decision may be difficult to understand, and the "many hands" problem, where responsibility is spread across several actors involved in the development and use of the system. The paper further uses actor mapping to identify the different participants involved in the AI lifecycle and examines why identifying these actors alone may not be enough to determine responsibility.

To address this gap, the paper proposes a Multi-Tiered Accountability Framework based on three connected levels: technical accountability, actor-based accountability and responsibility allocation. The framework considers factors such as an actor's contribution, level of control, foreseeability of risk and ability to prevent or reduce harm. It aims to provide a clearer and fairer approach to accountability without treating AI autonomy as a reason to remove human responsibility. The paper ultimately argues that as AI becomes more autonomous, legal accountability must develop alongside technological innovation.

Keywords: Artificial Intelligence; Autonomous Decision-Making Systems; Algorithmic Accountability; Black Box Problem; Many Hands Problem; Actor Mapping; AI Liability; Human Oversight

1 INTRODUCTION

Artificial intelligence has moved beyond being a tool that simply follows instructions given by humans. AI systems are increasingly capable of analysing large amounts of information, identifying patterns, making predictions and producing decisions with limited human intervention. The European Union's AI Act itself recognises that AI systems may operate with different levels of autonomy and may generate outputs such as predictions, recommendations and decisions that influence physical or virtual environments.[1] This increasing autonomy creates an important legal question: When an AI system causes harm, who should be held responsible?

The question becomes more difficult because an AI system does not operate in isolation. Behind a single AI-generated decision may be a developer who designed the system, a company that supplied or trained it, a data provider, an organisation that deployed it, and a human operator who relied upon its output. Therefore, simply asking who "made" the decision may not always provide a satisfactory answer.

[*] Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.
[**] Student at Christ (Deemed to be University), Bengaluru, Karnataka, India.

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, art. 3(1).

The problem becomes even more serious when the reasoning behind an AI decision cannot be clearly understood. Machine-learning systems may process enormous quantities of data and identify patterns that are difficult for even their creators or users to follow. This is commonly referred to as the "black box" problem. The concern is not merely that AI is complicated, but that a person affected by an AI decision may be unable to understand why that decision was reached or who should explain it.[2]

At the same time, responsibility may be spread across several human and institutional actors. This creates what has been described as the "many hands" problem, where several actors contribute to an outcome but no single actor appears to have complete responsibility for it.[3] In the context of AI, the developer may argue that the system was used differently from its intended purpose, while the user may argue that the system was designed by someone else. The result can be a gap between the occurrence of harm and the identification of a responsible party.

This paper therefore proceeds from the position that AI should not simply be restricted because it is becoming more autonomous. Technological development is inevitable, and the law cannot realistically expect technology to remain within the boundaries of older legal categories. However, allowing technology to develop does not mean allowing accountability to disappear with it.

The central argument of this paper is that responsibility for autonomous decision-making systems should be determined through a multi-tiered framework. Rather than automatically placing responsibility upon the developer, manufacturer, user or organisation, the law should examine the different actors involved, their contribution, their degree of control, the risks they were capable of identifying, and the safeguards they were expected to maintain. This approach begins with understanding the nature of autonomous decision-making systems, followed by an examination of the black box and many hands problems, before mapping the actors involved in the AI lifecycle.

2 AUTONOMOUS DECISION-MAKING SYSTEMS

The term "autonomous decision-making system" refers to an AI-based system that is capable of processing information and producing outputs without requiring a human being to make every individual decision. Autonomy, however, should not be understood as meaning that the AI system has become a completely independent human-like actor. Rather, it refers to the degree to which the system can operate without direct human intervention at each stage.

This distinction is important because not every automated system creates the same legal problem. A simple system that follows a fixed instruction is relatively easier to understand. For example, a calculator will produce an answer based on a mathematical formula. A more advanced machine-learning system, however, may identify patterns from data and produce an outcome that was not individually programmed by a human. The more such systems are capable of adapting, learning or making decisions with limited human intervention, the more difficult it becomes to trace the exact chain that led to a particular outcome.

The growing use of large datasets is one of the reasons for this difficulty. Modern AI systems can process extremely large amounts of information and identify patterns across that information. This ability is one of their greatest advantages, but it can also create risks. If the information used to train or operate

[2] Jenna Burrell, *How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOC'Y 1 (2016).
[3] Filippo Santoni de Sio & Giulio Mecacci, *Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them*, 35 PHIL. & TECH. 1057 (2021).

an AI system contains bias, inaccuracies or incomplete information, the resulting system may reproduce or even amplify those problems.

Therefore, the question is not simply whether AI is "intelligent". The more important question from a legal perspective is "What happens when an AI system acts on the basis of its own processing and that action causes harm?"

For instance, consider an AI system used by an organization to assess applications for employment, loans, insurance or other services. If the system produces an unfavorable decision, the affected individual may naturally ask why the decision was made. If a human employee makes the decision, it is generally possible to identify the person, ask for the reasoning behind the decision and examine whether that person acted negligently, unfairly or unlawfully. With an autonomous AI system, however, the process may involve several stages of data collection, training, programming, deployment and operation.

This creates what can broadly be described as a responsibility gap. Scholars have argued that increasing autonomy can make it difficult to determine whether responsibility lies with the programmer, manufacturer, operator or another actor when an autonomous system produces an undesirable outcome.[4] At the same time, it would be too simplistic to assume that every autonomous system automatically creates a responsibility gap. Whether such a gap exists depends upon the particular circumstances, including the nature of the system, the degree of human control and the roles played by different actors.[5]

This is where the present research takes a middle position. AI systems should not be treated as inherently unsafe merely because they are autonomous. At the same time, autonomy should not become an excuse for avoiding human responsibility. The law must instead examine the circumstances surrounding each system and determine which actors were in a position to prevent, control or respond to the harm.

Autonomy, therefore, creates the starting point of the accountability problem, but it does not by itself answer the question of liability. To answer that question, it is necessary to examine what happens inside the system and how its decision-making process can be understood.

2.1 The Black Box Problem

One of the most important challenges created by AI-based decision-making is the black box problem. In simple terms, a black box refers to a system where the relationship between the information given to the system and the final output is difficult to understand or explain.

The problem is particularly significant in machine-learning systems. Unlike traditional computer programs, where a programmer may directly specify the rules that determine the output, machine-learning systems can learn patterns from data. As a result, even when the system is functioning according to its technical design, it may be difficult for a human being to explain exactly why a particular result was produced.

Jenna Burrell identifies different forms of algorithmic opacity, including secrecy, lack of technical understanding and opacity arising from the complexity and scale of machine-learning systems themselves.[6] This distinction is important because the black box problem does not always mean that a company

[4] Peter Königs, *Artificial Intelligence and Responsibility Gaps: What Is the Problem?*, 24 ETHICS & INFO. TECH. 28 (2022).
[5] *Id.*
[6] Burrell, *supra* note 2.

is deliberately hiding information. Sometimes, the problem is that the process is technically difficult to interpret even for people involved in developing or operating the system.

This creates a serious problem when AI is used in areas affecting people's rights and interests. If an AI system rejects a loan application, identifies a person as a security risk, rejects a job application or makes another consequential classification, the affected individual may need to know the basis of the decision. Without meaningful information about how the decision was reached, challenging the decision becomes considerably more difficult.

The problem becomes even more significant when responsibility is being considered. Suppose an AI system produces a harmful decision and the developers themselves cannot fully explain why the system produced that result. The question then becomes: How can responsibility be fairly assigned?

This does not necessarily mean that such systems should never be used. Instead, they should be used with an appropriate level of caution, particularly where the consequences of error are serious. The National Institute of Standards and Technology (NIST), for example, treats accountability and transparency as important characteristics of trustworthy AI and links meaningful transparency to information concerning design decisions, training data, model training, system structure, intended uses and decisions made during deployment.[7]

This demonstrates why transparency cannot simply mean revealing the source code. A person affected by an AI decision may not benefit from being given thousands of lines of code if they still cannot understand how that code affected the outcome. Meaningful accountability requires information that is relevant to the particular stage of the AI lifecycle and the people interacting with the system.

The black box problem also creates an evidentiary difficulty. In a traditional dispute, the parties may attempt to establish what happened by examining documents, decisions and the conduct of the relevant person. With an AI system, the relevant evidence may be spread across training data, model design, system logs, deployment decisions and human interactions with the system.

This is why explainability and transparency are important, but they should not be treated as complete solutions to the problem. Making an algorithm more explainable may help an affected person understand the decision, but it does not automatically tell us who should be legally responsible for the consequences. Responsibility may still be distributed among several actors.

The black box therefore represents one side of the accountability problem: we may not understand how the decision was produced. The many hands problem represents the other side: even if we understand how the system was created and used, several different actors may have contributed to the outcome.

2.2 The Many Hands Problem

The second major difficulty is the many hands problem. Unlike the black box problem, which focuses mainly on the opacity of the decision-making process, the many hands problem focuses on the distribution of responsibility between different actors.

An AI system is rarely created, controlled and used by one person. A single system may involve software developers, data scientists, data providers, manufacturers, technology companies, organisations that purchase or deploy the system, human operators and end users. Regulators may also influence how the system can be developed and used.

---

[7] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (NIST AI 100-1, 2023); *see also* NIST, *AI Risks and Trustworthiness*.

Recent scholarship specifically connects AI responsibility gaps with this problem of responsibility distribution. Where several agents contribute to an outcome, identifying the relevant actors is only the first step; determining how much responsibility each actor should bear remains a separate question.[8]

Consider a hypothetical AI system used to decide whether individuals qualify for a particular service. One company develops the underlying model. Another supplies the data. A third organisation integrates the system into its own platform. A fourth organisation deploys it in practice, while employees use the system's output to make decisions affecting individuals.

If the system produces a discriminatory or otherwise harmful result, it would be unreasonable to automatically conclude that the final user alone is responsible. At the same time, it may also be unfair to automatically blame the original developer without examining how the system was deployed or whether the user ignored warnings and safeguards.

Therefore, responsibility should be divided according to the contribution, control and circumstances of each actor.

This does not mean that every person involved should automatically share equal liability. Equal participation does not necessarily mean equal responsibility. A developer who knowingly creates a defective system, a company that knowingly deploys it in a high-risk environment and an employee who unknowingly relies on its output may have very different levels of responsibility.

The appropriate approach must therefore remain contextual. The same AI system could produce different liability questions depending on how it was used, what information was available to the relevant actors, what safeguards existed and what degree of control each actor possessed.

At the same time, the rules used to make these assessments should not change unpredictably from one situation to another. There needs to be a stable framework through which courts and other decision-makers can assess the roles of different actors. Otherwise, uncertainty itself may become a barrier to responsible AI development.

This is where the many hands problem becomes particularly important to the present research. It demonstrates why simply identifying "a human behind the AI" is not sufficient. There may be many humans behind the AI, each performing a different function.

The challenge is therefore to move from the broad question "Who is responsible?" to the more useful questions:

1. Who created or designed the system?
2. Who supplied or controlled the relevant data?
3. Who decided to deploy the system?
4. Who had control over its use?
5. Who could reasonably have identified the risk?
6. Who had the ability to prevent or reduce the harm?
7. Who failed to take an appropriate safeguard?

Answering these questions provides the basis for a more structured approach to accountability.

[8] Filippo Santoni de Sio & Giulio Mecacci, *Four Responsibility Gaps with Artificial Intelligence: Why They Matter and How to Address Them*, 34 PHIL. & TECH.

2.2.1 Actor Mapping

Before liability can be fairly assigned, the actors involved in an autonomous decision-making system must first be identified. This process can be described as actor mapping.

Actor mapping does not mean deciding in advance who is legally liable. Instead, it creates a structured picture of the different participants involved in the AI lifecycle. This is important because responsibility may arise at different stages and for different reasons.

TABLE 1. *Actors in the AI lifecycle and possible accountability concerns*

ACTOR	ROLE IN AI LIFECYCLE	POSSIBLE ACCOUNTABILITY CONCERN
Developer	Designs and develops the AI system	Design defects, inadequate safeguards or foreseeable risks
Data provider	Supplies training or operational data	Poor-quality, inaccurate or biased data
Manufacturer	Integrates the AI system into a product or service	Integration or technical failures
Deployer	Introduces the system into a real-world environment	Inappropriate deployment or inadequate monitoring
Operator/User	Uses or supervises the system	Misuse, failure to follow safeguards or unreasonable reliance
Organization	Determines how and why the system is used	Governance, oversight and risk-management failures
Regulator	Establishes and enforces applicable standards	Regulatory oversight and compliance

The purpose of this map is not to suggest that each actor will necessarily be liable whenever something goes wrong. Instead, it helps identify where responsibility could potentially arise.

For example, if an AI system produces a harmful result because its training data contained a serious and foreseeable bias, the data-related actor and developer may require examination. If the system itself was adequately designed but an organisation deployed it in a context for which it was never intended, the focus may shift towards the deployer. Similarly, if adequate safeguards existed but an operator deliberately ignored them, the operator's conduct may become relevant.

This approach also reflects the fact that accountability exists throughout the AI lifecycle rather than only at the moment when the final decision is made. NIST's AI Risk Management Framework similarly emphasises the importance of identifying roles and responsibilities across the AI lifecycle and treating AI risk management as an ongoing process rather than a one-time exercise.[9]

Actor mapping therefore acts as a bridge between the conceptual problems identified above and the question of legal liability. The black box problem tells us that we may struggle to understand how the system reached a particular outcome. The many hands problem tells us that several actors may have contributed to that outcome. Actor mapping provides a method of identifying those actors and examining their respective roles.

The central principle emerging from this analysis is that accountability should follow meaningful involvement and control rather than simply follow ownership of the AI system. A person should not escape responsibility merely because the final decision was technically produced by an algorithm. At the same time, an individual should not automatically be blamed merely because they were the last person to interact with the system.

[9] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (NIST AI 100-1, 2023).

Ultimately, the law must distinguish between different forms and degrees of involvement. The greater an actor's ability to anticipate, control, prevent or mitigate a particular risk, the stronger the basis for examining that actor's responsibility. This does not create an automatic rule of liability, but it provides a rational starting point for determining where liability should be placed.

This actor-based approach forms the foundation for the subsequent stages of the proposed multi-tiered framework, which examine how responsibility can be translated into appropriate forms of accountability and liability.

3 ACCOUNTABILITY THEORY

Accountability is a fundamental concept in understanding how power and decision-making should be exercised, particularly where decisions can affect individuals and society. Traditionally, accountability refers to a relationship in which an actor is required to explain and justify their conduct to a forum that has the authority to question, evaluate and potentially impose consequences.[10] This traditional understanding assumes that the person or institution making a decision can be identified and held answerable for that decision.

With the increasing use of artificial intelligence and automated decision-making systems, this traditional understanding of accountability has become more complicated. Algorithmic accountability extends the concept of accountability to systems in which decisions may be influenced or produced by algorithms rather than solely by human decision-makers. It therefore requires attention not only to the final outcome but also to the design, development, deployment and use of the AI system.[11]

Accountability should also be distinguished from related concepts such as transparency and explainability. Transparency generally concerns access to information about how a system operates, while explainability focuses on making the reasoning or output of an AI system understandable. Accountability is broader because it concerns who must answer for a decision, to whom they must answer, according to what standards, and what consequences may follow. Thus, transparency or explainability may support accountability, but they do not by themselves establish accountability.

Similarly, accountability, responsibility and legal liability are not interchangeable concepts. Responsibility concerns who is expected to perform or control an action, while accountability involves being answerable for conduct and its consequences. Legal liability, on the other hand, concerns whether a person or organisation can be held legally responsible under a particular legal rule. Keeping these concepts separate is particularly important in autonomous AI systems because identifying the legally liable party may become difficult when decisions result from complex interactions between developers, deployers, users and the system itself.

This difficulty is reflected in the idea of a responsibility gap. Andreas Matthias argues that learning and autonomous systems can behave in ways that could not have been fully predicted by their designers or operators. As a result, traditional approaches to assigning responsibility may struggle to identify a clear human actor responsible for a particular outcome.[12] In AI decision-making, responsibility may therefore be distributed across several actors rather than resting with a single individual.

---

[10] Mark Bovens, *Analysing and Assessing Accountability: A Conceptual Framework*, 13 EUR. L.J. 447 (2007).

[11] Claudio Novelli, Mariarosaria Taddeo & Luciano Floridi, *Accountability in Artificial Intelligence: What It Is and How It Works*, 39 AI & SOC'Y 1871 (2024).

[12] Andreas Matthias, *The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata*, 6 ETHICS & INFO. TECH. 175 (2004).

The problem becomes more significant as AI systems gain greater autonomy. Where human intervention is limited, it becomes necessary to examine whether meaningful human control existed over the system and its decisions. The concept of meaningful human control therefore provides an important basis for determining how human actors should remain connected to autonomous systems and their outcomes.[13]

These difficulties can ultimately produce what may be described as an accountability gap—a situation in which an AI-generated decision has significant consequences, but no actor can be clearly identified as answerable for the decision. The complexity of algorithmic systems, combined with their capacity to operate with limited human intervention, can make conventional accountability mechanisms inadequate.

A human-centred approach to accountability seeks to address this problem by ensuring that accountability remains attached to human and institutional actors rather than treating the AI system itself as the accountable party. This requires identifying the relevant actors throughout the AI lifecycle and establishing mechanisms through which their decisions and actions can be questioned and evaluated.

A further approach is to understand accountability as distributed across the AI lifecycle. Instead of assigning accountability only at the point where an automated decision produces harm, responsibility and answerability can be considered at different stages, including system design, development, deployment, monitoring and use. This approach recognises that different actors may exercise different forms of control over an AI system and may therefore have different accountability obligations.

Overall, the accountability framework for autonomous decision-making systems must move beyond the idea of identifying a single decision-maker. Traditional accountability theory provides the foundation, while algorithmic accountability adapts that foundation to systems involving multiple actors, limited predictability and varying degrees of human control. The responsibility gap and the need for meaningful human control demonstrate why accountability must be distributed across the AI lifecycle rather than concentrated solely on the final user or decision-maker.

This theoretical understanding of accountability provides the foundation for examining the next issue: legal liability. While accountability determines who should be answerable for AI-related decisions, legal liability asks a more specific question—who can actually be held legally responsible when an autonomous AI system causes harm or violates a legal right?

4 LEGAL LIABILITY IN AUTONOMOUS AI DECISION-MAKING

Legal liability refers to the legal consequences that arise when an individual or organization breaches a legal duty and causes legally recognized harm. In the context of artificial intelligence, the question of liability becomes more complex because the harmful outcome may result not from the direct action of a single person, but from the interaction between developers, providers, deployers, users and an increasingly autonomous system. It is therefore important to distinguish liability from accountability and responsibility. While accountability concerns the obligation to explain and justify conduct, and responsibility may refer to the duties or obligations associated with an actor's role, legal liability determines whether an actor can be held legally responsible and required to provide a remedy for harm. The distinction becomes

[13] Filippo Santoni de Sio & Jeroen van den Hoven, *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, 5 FRONTIERS IN ROBOTICS & AI 15 (2018).

particularly significant where an AI system produces an outcome that no individual actor directly intended or fully anticipated.[14]

4.1 Attribution of Liability

The central problem in AI liability is therefore one of attribution: identifying the person or legal entity to whom a harmful AI decision can properly be attributed. An AI system does not operate in isolation; its behavior may reflect decisions made during its design, training, deployment, monitoring and use. Consequently, liability may potentially involve the developer who designed the system, the provider who placed it on the market, the organization that deployed it, or the operator who relied upon its output. The difficulty lies in determining which actor exercised sufficient control over the relevant risk to justify imposing legal liability. This is particularly challenging where several actors contribute to the operation of the system and their respective roles overlap.[15]

4.2 The Liability Gap in Autonomous AI

Autonomous and adaptive AI systems create what has been described as a liability gap. Traditional liability rules generally operate on the assumption that a harmful act can be connected to an identifiable human actor, a breach of duty or a defective product. However, autonomous systems may produce outcomes that are difficult to predict even by their designers or operators. This does not mean that the AI system itself should automatically become the bearer of legal liability; rather, it exposes a gap between the occurrence of harm and the ability of existing legal rules to identify an appropriate defendant. The more autonomous, complex and unpredictable the system becomes, the more difficult it may be to determine where one actor's legal responsibility ends and another's begins.[16]

4.3 Fault-Based Liability and Negligence

One possible route for addressing AI-related harm is fault-based liability, particularly through negligence. Traditionally, negligence requires the existence of a duty of care, a breach of that duty, a causal connection between the breach and the harm, and legally recognized damage. These principles can potentially apply to AI systems where, for example, a developer fails to conduct adequate testing or a deployer fails to exercise reasonable supervision. However, their application becomes difficult where the specific harmful behaviour of an autonomous system could not reasonably have been anticipated. The concept of foreseeability, which is central to many negligence-based approaches, becomes particularly difficult to apply when an AI system can learn from new data, adapt after deployment or generate unexpected outputs.[17]

4.4 Strict Liability and Product Liability

These difficulties raise the question of whether strict or no-fault forms of liability may be more appropriate for certain high-risk AI applications. Under a strict-liability approach, the claimant does not necessarily have to establish the same level of fault or negligence on the part of the defendant. Such an approach

---

[14] Mark Bovens, *Analysing and Assessing Accountability: A Conceptual Framework*, 13 EUR. L.J. 447 (2007).
[15] Miriam C Buiten, *Product Liability for Defective AI*, 57 EUR. J.L. & ECON. 239 (2024).
[16] Simon Burton et al., *Mind the Gaps: Assuring the Safety of Autonomous Systems from an Engineering, Ethical, and Legal Perspective*, 279 ARTIFICIAL INTELLIGENCE 103201 (2020); *see also* Miriam C Buiten, Alexandre de Streel & Martin Peitz, *The Law and Economics of AI Liability*, 48 COMPUT. L. & SEC. REV. 105794 (2023).
[17] *Id.*

could provide stronger protection where the risks associated with highly autonomous systems are significant but difficult to predict. Product liability is particularly relevant because AI may form part of a physical product or operate as software in its own right. Modern product-liability approaches increasingly recognise that software and AI can create legally relevant risks. For example, the EU's revised Product Liability Directive expressly treats software, including AI systems, as a product for the purposes of its no-fault product-liability regime.[18]

However, product liability does not completely resolve the problem. AI systems may change through updates, learning processes or interaction with their environment, making it difficult to determine whether the relevant defect existed when the product entered the market. The question therefore becomes not simply whether an AI system was defective, but who had control over the relevant risk and at what stage of the system's lifecycle that risk arose. This makes the allocation of liability between producers, providers and users particularly important.[19]

4.5 Allocation of Liability Across the AI Lifecycle

Rather than treating AI liability as a question of identifying one person to blame, a more appropriate approach may be to consider how liability should be allocated across the AI lifecycle. Developers may be responsible for design and development failures; providers may bear obligations concerning the system as supplied; deployers may be responsible for inappropriate implementation or inadequate monitoring; and users or operators may be liable where they misuse the system or disregard required safeguards. This approach recognises that control over an AI system is often distributed rather than concentrated in one actor. It also avoids automatically transferring liability to the person who happens to be closest to the final harmful decision when the underlying risk may have originated much earlier in the system's development.[20]

4.6 Human Oversight and Remedies

Human oversight therefore remains an important consideration when determining liability for autonomous decision-making. Where a system is capable of producing significant consequences, the presence or absence of meaningful human supervision may influence whether a deployer or operator has fulfilled its legal duties. Effective oversight should not merely consist of having a human formally present; it should involve the ability to understand relevant limitations, monitor the system and intervene when necessary. The EU AI Act, for example, requires appropriate human oversight for high-risk AI systems and recognises the need for humans to monitor, interpret and, where appropriate, override or stop the system.[21]

Ultimately, the objective of an AI liability framework should not simply be to identify a defendant after harm has occurred. It should also ensure that individuals affected by autonomous decision-making have a meaningful avenue for redress and compensation, while creating incentives for developers, providers and deployers to manage foreseeable risks before harm occurs. The central challenge is therefore

[18] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on Liability for Defective Products. The Directive expressly includes software, including AI systems, within its product-liability framework.
[19] *Id.*
[20] Bart Custers, Henning Lahmann & Benjamyn I Scott, *From Liability Gaps to Liability Overlaps: Shared Responsibilities and Fiduciary Duties in AI and Other Complex Technologies*, 40 AI & SOC'Y 4035 (2025).
[21] Regulation (EU) 2024/1689 (Artificial Intelligence Act), art. 14.

to construct a liability model that reflects the distributed nature of AI development and use without allowing autonomy or technological complexity to become a means of escaping legal responsibility.

5 RESEARCH GAP

Existing discussions on AI accountability have largely focused on individual concerns such as transparency, explainability, human oversight, bias and liability. These issues are important, but they are often considered separately. The black box problem explains why it may be difficult to understand how an AI system reaches a decision, while the many hands problem explains why it may be difficult to identify who should be held responsible for that decision.

The difficulty becomes greater when these problems occur together. Even if an AI system can be examined technically, this does not automatically tell us which actor should be legally responsible when harm occurs. Similarly, identifying all the people and organisations involved in an AI system does not mean that all of them should be treated as equally responsible.

There is therefore a gap between understanding how an AI system operates and determining who should be accountable for its outcomes. A clear framework is needed to connect these two aspects. Such a framework should consider not only the technical functioning of the system, but also the role, contribution, control and ability to prevent harm of the different actors involved.

This paper addresses this gap by proposing a Multi-Tiered Accountability Framework. The framework aims to ensure that increasing autonomy in AI systems does not result in a corresponding loss of accountability.

6 MULTI-TIERED ACCOUNTABILITY FRAMEWORK

The proposed framework consists of three connected tiers: technical accountability, actor accountability and responsibility allocation. The tiers are designed to work together rather than as separate stages. The first helps understand the system, the second identifies the actors involved, and the third determines how responsibility should be allocated.

6.1 Tier One: Technical Accountability

The first tier focuses on the AI system itself. Before responsibility can be meaningfully discussed, there must be enough information to understand what happened within the system and how it was developed and deployed.

This does not mean that every AI system must be completely explainable in simple human terms. Complex machine-learning systems may not always allow their decisions to be reduced to a simple chain of reasoning. However, there should be sufficient documentation and monitoring to examine important aspects such as the data used, the purpose of the system, its design, known limitations, testing processes and significant changes made after deployment.

This tier directly addresses the black box problem. When an AI system produces a harmful result, complete opacity can make it difficult to determine whether the problem arose from the design of the system, the data used to train it, the way it was deployed or the way it was operated.

Therefore, technical accountability should involve appropriate testing, documentation, monitoring and record keeping. The question should not only be "Can we explain exactly why the AI made this decision?", but also "Do we have enough information to investigate what happened?" This makes technical transparency an important foundation for accountability.

6.2 Tier Two: Actor Accountability

The second tier addresses the many hands problem. Autonomous AI systems are usually not created, controlled or used by one person alone. Developers, data providers, manufacturers, technology companies, organisations, operators and other actors may all have different roles in the system.

The purpose of this tier is therefore to identify these actors and understand their individual roles. Actor mapping should consider who designed the system, who supplied or controlled the data, who decided to deploy it, who operated it and who was responsible for monitoring it.

However, identifying an actor does not automatically make that actor liable. The framework recognises that different actors may have different levels of responsibility.

For example, a developer may be responsible for a defect in the design, while an organisation that deploys the system may be responsible for using it in an unsuitable context. An operator may also bear responsibility where they had the ability to intervene but failed to do so.

The framework therefore asks:

1. Who created or developed the system?
2. Who supplied or controlled the relevant data?
3. Who decided to deploy the system?
4. Who had control over its use?
5. Who could reasonably have identified the relevant risk?
6. Who had the ability to prevent or reduce the harm?

These questions help ensure that responsibility is not placed automatically on the person who happened to be closest to the final decision.

6.3 Tier Three: Responsibility Allocation

The final tier determines how responsibility should actually be distributed between the actors identified in the previous tier.

The framework proposes that responsibility should be based on four main considerations: contribution, control, foreseeability and prevention.

First, contribution considers what role an actor played in creating or contributing to the harmful outcome. Second, control considers the extent to which the actor could influence the relevant part of the AI system. Third, foreseeability considers whether the risk could reasonably have been identified. Finally, prevention considers whether the actor had a realistic opportunity to prevent or reduce the harm.

These factors allow the framework to recognise shared but differentiated responsibility. More than one actor may be responsible for an outcome, but this does not mean that every actor should bear the same degree of responsibility.

For example, if a developer releases a system with a serious and foreseeable defect, the developer may bear greater responsibility. If an organization knowingly deploys the system for a purpose for which

it was not designed, responsibility may instead shift towards the deployer. Similarly, where meaningful human oversight existed but was ignored, the failure of that oversight may also become relevant.

The framework therefore avoids two extremes: treating AI as completely independent and placing responsibility on no one, or treating one human actor as automatically responsible for everything the AI does.

The underlying principle is simple: the more control and ability an actor had to identify, prevent or reduce a particular risk, the stronger the basis for holding that actor responsible for the resulting harm.

7 Conclusion

Autonomous decision-making systems are becoming increasingly capable of performing tasks that were once carried out directly by humans. While this development can provide significant benefits, it also creates new challenges for accountability. The difficulty is not simply that AI can make decisions autonomously, but that those decisions may be difficult to understand and may involve several different actors.

The black box problem creates uncertainty about how an AI system reaches its decisions, while the many hands problem creates uncertainty about who should be held responsible when those decisions cause harm. Actor mapping helps identify the participants in the system, but identifying them alone is not enough. There must also be a clear method for determining the extent of each actor's responsibility.

The Multi-Tiered Accountability Framework proposed in this paper attempts to address this gap by connecting technical accountability, actor accountability and responsibility allocation. It recognises that accountability should exist throughout the AI lifecycle rather than only after harm has occurred.

The framework does not argue that autonomous AI should be stopped or that every unexpected AI decision should result in human liability. Instead, it argues that greater autonomy should not mean weaker accountability. Responsibility should depend on the circumstances of each case and should be connected to an actor's contribution, control, ability to foresee the risk and ability to prevent or reduce the harm.

Technology is evolving, and law has to evolve with it. The challenge for the law is therefore not to prevent AI from becoming more autonomous, but to ensure that autonomy does not create a space where responsibility disappears. A system may be autonomous in its decision-making, but the legal system must still be able to answer the question: When something goes wrong, who was in the best position to prevent it, and why?

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.11

CHAPTER 11

Autonomous Weapon Systems in International Humanitarian Law: Command Responsibility and Accountability

MADHIROJA P S[*] AND VIGNESH ARUN PRAKASH S R[**]

Autonomous Weapon Systems (AWS) have emerged as one of the most urgent challenges for International Humanitarian Law (IHL), testing not only its doctrinal resilience but also its moral responsibility to safeguard human dignity in armed conflict. Defined as systems capable of selecting and engaging targets without direct human intervention, AWS raise profound concerns regarding compliance with the principles of distinction, proportionality and precaution. Simultaneously, they create a responsibility gap that weakens enforcement and undermines the doctrine of command responsibility. This study traces the evolution of discourse on AWS from early military applications to contemporary debates in United Nations forums and International Committee of the Red Cross expert consultations. It critically examines treaties, jurisprudence and institutional responses that have advanced normative principles without establishing binding accountability mechanisms. Recent conflicts reveal the unpredictability of autonomy and its humanitarian consequences, while the ethical insights emphasize preserving human agency and responsibility over machine decision-making. The novelty of this research lies in its integration of doctrinal analysis with ethical inquiry, offering a holistic framework that not only identifies the accountability vacuum but also proposes embedding meaningful human control as a foundational principle of IHL. By situating AWS within both legal and moral paradigms, this study advances a contemporary vision of IHL as the custodian of responsibility and humanity. It further affirms the necessity of codifying accountability to safeguard humanitarian principles and to protect future generations from the dangers of dehumanized warfare.

Keywords: Autonomous Weapon Systems; International Humanitarian Law; Command Responsibility; Accountability; Human Control

1 INTRODUCTION

The emergence of Autonomous Weapon Systems (AWS) in the twenty-first century represents one of the most significant challenges to International Humanitarian Law (IHL). The delegation of lethal decision-making to machines disrupts traditional doctrines of responsibility, tests the adequacy of humanitarian principles and exposes gaps in existing accountability mechanisms.[1] The growing integration of autonomous technologies into military operations has raised serious concerns regarding compliance with the principles of distinction, proportionality and precaution, which have historically depended upon human judgment and control.[2] Unlike conventional weapons, AWS operate with varying degrees of

[*] LL.M. Student at Government Law College, Madurai, Tamil Nadu, India.
[**] LL.M. Student at Government Law College, Madurai, Tamil Nadu, India.

[1] KENNETH ANDERSON & MATTHEW C. WAXMAN, LAW AND ETHICS FOR AUTONOMOUS WEAPON SYSTEMS: WHY A BAN WON'T WORK AND HOW THE LAWS OF WAR CAN (2013).

[2] Michael Schmitt, *Autonomous Weapon Systems and International Humanitarian Law*, 4 HARV. NAT'L SEC. J. 1 (2013); YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT 89–94 (2012).

autonomy, creating uncertainty in the attribution of responsibility when unlawful harm occurs and intensifying debates regarding command responsibility and legal accountability.[3] These concerns are compounded by ethical objections to delegating life-and-death decisions to machines and by the absence of a universally accepted regulatory framework governing their use.[4] Consequently, the study of AWS is not merely a technological or legal inquiry; it is fundamentally concerned with protecting civilians, preserving human dignity and ensuring that evolving methods of warfare remain subject to meaningful humanitarian and accountability standards.

The scholarly discourse on AWS within IHL has focused extensively on questions of responsibility attribution and command accountability. Early scholars identified the emergence of a "responsibility gap," arguing that autonomous decision-making disperses liability among programmers, military commanders and states.[5] Further, it was contended that delegating lethal decisions to machines undermines human moral agency and weakens traditional notions of accountability.[6] In contrast, the existing principles of IHL possess sufficient flexibility to regulate AWS when properly interpreted.[7] The International Committee of the Red Cross has consistently reinforced this position by advocating meaningful human control as an essential safeguard against humanitarian violations.[8] The principles of IHL lay the foundation of lawful military operations.[9] However, several studies suggest that AWS may struggle to distinguish combatants from civilians in complex battlefield environments.[10] Likewise, proportionality assessments and precautionary obligations require contextual human judgment that current autonomous technologies cannot reliably replicate.[11] Comparative analyses of jurisprudence from the International Criminal Tribunal for the Former Yugoslavia (ICTY) and the International Criminal Court (ICC) indicate that existing legal precedents were developed for human actors, leaving significant uncertainty regarding their application to autonomous weapons and machine-driven decision-making.[12]

Recent literature has increasingly shifted towards institutional and regulatory responses. Debates within the United Nations Convention on Certain Conventional Weapons (UNCCW) reveal a continuing divide between states advocating a pre-emptive prohibition on AWS and those favouring gradual regulation through existing legal frameworks.[13] Contemporary scholars called for stronger accountability mechanisms and the incorporation of ethical safeguards into international law.[14] Conversely, some commentators caution that overly restrictive regulation may impede technological innovation.[15] More recent studies emphasize that principles such as distinction, proportionality and precaution presuppose human

---

[3] Noel Sharkey, *Saying 'No!' to Lethal Autonomous Targeting*, 9 J. MIL. ETHICS 369 (2010).
[4] Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007).
[5] Andreas Matthias, *The Responsibility Gap*, 6 ETHICS & INFO. TECH. 175 (2004).
[6] Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007); Peter Asaro, *On Banning Autonomous Weapon Systems: Human Rights, Automation, and the Dehumanization of Lethal Decision-Making*, 94 INT'L REV. RED CROSS 687 (2012).
[7] Michael Schmitt, *Autonomous Weapon Systems and International Humanitarian Law*, 4 HARV. NAT'L SEC. J. 1 (2013).
[8] INT'L COMM. OF THE RED CROSS, VIEWS ON AUTONOMOUS WEAPON SYSTEMS (2016).
[9] YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT (2012); Christopher Greenwood, *The Concept of Proportionality in International Humanitarian Law*, 1 Y.B. INT'L HUMANITARIAN L. 3 (1998).
[10] Noel Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. RED CROSS 787 (2012).
[11] RONALD ARKIN, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS (2009); MARCO SASSÒLI, INTERNATIONAL HUMANITARIAN LAW: RULES, CONTROVERSIES, AND SOLUTIONS TO PROBLEMS ARISING IN WARFARE (2019).
[12] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999).
[13] U.N. Convention on Certain Conventional Weapons, *Report of the Group of Governmental Experts on Lethal Autonomous Weapons Systems*, U.N. Doc. CCW/GGE.1/2019/3.
[14] HUMAN RIGHTS WATCH & INT'L HUMAN RIGHTS CLINIC, HARVARD LAW SCH., LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS (2012); Rebecca Crootof, *War Torts: Accountability for Autonomous Weapons*, 164 U. PA. L. REV. 1347 (2016).
[15] Gary Marchant & Braden Allenby, *Regulating Emerging Military Technologies: Ethics and Law*, J.L. TECH. & POL'Y 1 (2015).

judgment and that autonomous systems risk creating accountability vacuums by diffusing liability among programmers, commanders and states.[16] Against this backdrop, the present study examines the challenges of responsibility, legal regulation and accountability posed by AWS within the framework of IHL.[17]

2 RESPONSIBILITY GAP IN AWS DEPLOYMENT

The deployment of Autonomous Weapon Systems (AWS) without meaningful human oversight has generated profound concern within international legal and ethical discourse. At the heart of this debate lies the "responsibility gap", a doctrinal uncertainty regarding who bears liability when autonomous systems act unpredictably. This section examines the contours of responsibility attribution, situating the discussion within established principles of command responsibility, while acknowledging the human dimension of accountability in warfare.

2.1 Command Responsibility

The doctrine of command responsibility, crystallized in precedents such as *Yamashita* and codified under Article 28 of the Rome Statute, requires military superiors to exercise effective control over subordinates.[18] In the context of AWS, however, the "subordinate" is no longer a human actor but a machine operating with varying degrees of autonomy. This substitution challenges the doctrinal foundation of liability: can a commander be held responsible for the unpredictable actions of an algorithm? Scholars argue that the principle of "effective control" must be reinterpreted to encompass technological oversight, ensuring that accountability remains anchored in human decision-making[19]. The human touch here is evident: commanders must not abdicate moral responsibility simply because machines execute their orders.

2.2 Responsibility Attribution

Responsibility attribution in AWS deployment is complicated by the diffusion of roles among programmers, manufacturers, military operators and state authorities.[20] Unlike traditional weapons, where liability flows through a clear chain of command, AWS introduce multiple nodes of accountability. The unpredictability of machine learning systems exacerbates this diffusion, creating scenarios where violations of IHL may occur without a discernible human agent.[21] Diplomatic discourse within the UN CCW has emphasized the need for "meaningful human control" as a normative safeguard, underscoring that responsibility must remain human-centred even in technologically mediated warfare.[22] This reflects a broader humanitarian concern: victims of unlawful attacks deserve justice and justice requires identifiable human accountability.

---

[16] INT'L COMM. OF THE RED CROSS, AUTONOMOUS WEAPON SYSTEMS AND INTERNATIONAL HUMANITARIAN LAW: SELECTED ISSUES 8–14 (Oct. 2025); OMAR SAAD AHMED & EMAN NAJI ABDULMAJEED, AUTONOMOUS WEAPONS SYSTEMS UNDER INTERNATIONAL HUMANITARIAN LAW: A LEGAL–ETHICAL AND DATA-DRIVEN ANALYSIS OF MODERN WARFARE 45–52 (2025).

[17] Bonnie Docherty, *Mind the Gap: The Lack of Accountability for Killer Robots* (Human Rights Watch Briefing Paper 2015); WILLIAM BOOTHBY, WEAPONS AND THE LAW OF ARMED CONFLICT (2d ed. 2013).

[18] *In re* Yamashita, 327 U.S. 1 (1946).

[19] Rome Statute of the International Criminal Court art. 28, July 17, 1998, 2187 U.N.T.S. 90.

[20] Andreas Matthias, *The Responsibility Gap*, 6 ETHICS & INFO. TECH. 175 (2004).

[21] Noel Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. RED CROSS 787 (2012).

[22] ICRC, VIEWS ON AUTONOMOUS WEAPON SYSTEMS (Geneva 2016).

2.3 Accountability Gap

The accountability gap emerges when existing legal frameworks fail to provide mechanisms for enforcement against AWS-related violations. Current jurisprudence, designed for human actors, struggles to accommodate autonomous systems that lack intent or moral agency.[23] This gap risks undermining the deterrent function of IHL as violations may go unpunished due to doctrinal ambiguity. Institutional actors such as the ICC and ICRC have called for proactive measures ranging from interpretive expansions of existing law to the negotiation of new treaties.[24] The human dimension of this debate cannot be overstated: accountability is not only about punishing wrongdoers but also about affirming the dignity of victims and reinforcing the moral fabric of international law.

Responsibility in warfare is not merely a legal construct but a moral commitment to human dignity. AWS challenge this commitment by introducing unpredictability into lethal decision-making. This section has traced the doctrinal roots of command responsibility, examined the diffusion of liability and highlighted the accountability gap. Together, these analyses underscore the urgent need for legal and institutional innovation to preserve the humanitarian essence of IHL.

3 ADEQUACY OF IHL IN REGULATING AUTONOMOUS TARGETING

International Humanitarian Law (IHL) is built upon principles designed to safeguard human dignity in armed conflict. Yet, the rise of Autonomous Weapon Systems (AWS) has exposed doctrinal vulnerabilities. The principles of distinction, proportionality and precaution are the cornerstones of humanitarian law which are conceived for human judgment not algorithmic decision-making. This section examines whether existing IHL frameworks are adequate to regulate autonomous targeting, balancing doctrinal analysis with the ethical imperative to protect civilians and preserve humanitarian values.

3.1 Distinction Principle

The principle of distinction obliges parties to differentiate between combatants and civilians.[25] AWS complicate this obligation by relying on sensors and algorithms that may misclassify targets in complex environments.[26] Empirical studies highlight the difficulty of encoding contextual cues such as surrender gestures or civilian behaviour into machine logic.[27] The human dimension here is critical: misclassification risks catastrophic humanitarian consequences and erodes trust in IHL. The ICRC has emphasized that distinction requires human judgment, reinforcing the need for "meaningful human control" to ensure compliance.[28]

3.2 Proportionality Principle

Proportionality requires balancing anticipated military advantage against collateral civilian harm.[29] Human commanders exercise moral and contextual judgment in this calculus; AWS by contrast rely on

[23] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment (ICTY 1999).
[24] U.N. CCW, *Report of the Group of Governmental Experts on Lethal Autonomous Weapons Systems*, U.N. Doc. CCW/GGE.1/2019/3.
[25] Additional Protocol I to the Geneva Conventions art. 48, June 8, 1977, 1125 U.N.T.S. 3.
[26] Noel Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. RED CROSS 787 (2012).
[27] Peter Asaro, *On Banning Autonomous Weapon Systems*, 94 INT'L REV. RED CROSS 687 (2012).
[28] ICRC, VIEWS ON AUTONOMOUS WEAPON SYSTEMS (Geneva 2016).
[29] Additional Protocol I to the Geneva Conventions art. 51(5)(b), June 8, 1977, 1125 U.N.T.S. 3.

algorithmic models that may oversimplify complex realities.[30] Arkin's work on "ethical governors" suggests that machines could theoretically enforce proportionality,[31] but critics argue that algorithms cannot replicate human empathy or moral reasoning.[32] Diplomatic debates underscore this concern; proportionality is not a mere mathematical equation but a moral safeguard against excessive harm.[33] Without human oversight, AWS risk undermining the humanitarian essence of proportionality.

3.3 Legal Frameworks

The international legal framework governing AWS is rooted in the Geneva Conventions of 1949 and Additional Protocol I of 1977, which enshrine the principles of distinction, proportionality and precaution.[34] These instruments, alongside customary international law, presuppose human agency in decision-making. AWS, however, introduce autonomous processes that challenge the interpretive scope of these provisions. Scholars such as Schmitt argue that IHL retains resilience if applied rigorously,[35] while others caution that the absence of explicit references to autonomy leaves dangerous gaps.[36]

At the national level, India's legal framework reflects both constitutional and statutory dimensions. The Indian Constitution enshrines the protection of life and personal liberty under Article 21,[37] which indirectly informs humanitarian obligations in armed conflict. India is a party to the Geneva Conventions and has incorporated their provisions into domestic practice through military manuals and operational directives.[38] The Indian Army Doctrine emphasizes compliance with IHL principles, particularly distinction and proportionality in targeting operations.[39] However, India has not yet adopted specific legislation addressing AWS, leaving doctrinal uncertainty in regulating autonomous targeting.[40] This gap underscores the need for harmonization between international obligations and national practice, ensuring that technological advances do not erode humanitarian safeguards.

This section demonstrates that while IHL provides a robust doctrinal foundation, its adequacy in regulating AWS remains contested. The principles of distinction and proportionality demand human judgment that machines cannot replicate. International frameworks, though resilient, require reinterpretation and possibly codification to address autonomy explicitly. National frameworks, such as India's, highlight the importance of aligning constitutional protections and military doctrine with evolving technologies. The human touch is indispensable in safeguarding civilians and preserving humanitarian values, and it must remain at the heart of IHL, even in an era of technological warfare.

---

[30] YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT (2012).
[31] RONALD ARKIN, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS (2009).
[32] Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007).
[33] MARCO SASSÒLI, INTERNATIONAL HUMANITARIAN LAW: RULES, CONTROVERSIES, AND SOLUTIONS TO PROBLEMS ARISING IN WARFARE (2019).
[34] Geneva Convention (IV) Relative to the Protection of Civilian Persons in Time of War, Aug. 12, 1949, 75 U.N.T.S. 287.
[35] Michael Schmitt, *Autonomous Weapon Systems and International Humanitarian Law*, 4 HARV. NAT'L SEC. J. 1 (2013).
[36] Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837 (2015).
[37] INDIA CONST. art. 21 (1950).
[38] GOV'T OF INDIA, MINISTRY OF EXTERNAL AFF., RATIFICATION OF GENEVA CONVENTIONS (1950).
[39] INDIAN ARMY, DOCTRINE OF THE INDIAN ARMY 112–18 (2004).
[40] Anviksha Pachori & Abhishek Bhati, *India's Regulatory and Ethical Stance on Autonomous Weapon Systems*, 6 ASIA-PAC. J. INT'L HUMANITARIAN L. (2025), https://apjihl.org/article/indias-regulatory-and-ethical-stance-on-autonomous-weapon-systems/.

4 MECHANISMS FOR ACCOUNTABILITY IN AWS DEPLOYMENT

Accountability is the cornerstone of International Humanitarian Law (IHL). Without clear mechanisms to attribute responsibility, violations risk going unpunished, eroding both deterrence and the moral fabric of humanitarian law. Autonomous Weapon Systems (AWS) complicate accountability by diffusing liability across programmers, manufacturers, commanders and states. This section explores institutional, doctrinal and ethical mechanisms for accountability with particular emphasis on the Rome Statute, the role of international institutions and normative ethical perspectives. The human dimension is central; accountability is not only about punishment but about affirming justice for victims and reinforcing trust in humanitarian law.

4.1 Rome Statute and International Criminal Court (ICC)

The Rome Statute provides the most comprehensive framework for accountability in armed conflict, codifying crimes such as war crimes, crimes against humanity and genocide.[41] Article 25 establishes individual criminal responsibility, while Article 28 codifies command responsibility.[42] Yet AWS challenge these provisions by introducing actors without intent or moral agency. Scholars argue that liability must be extended to commanders who authorize AWS deployment without adequate safeguards.[43] International criminal jurisprudence, though not yet tested on AWS, offers analogical guidance: responsibility must remain human-centered, ensuring that technological mediation does not dilute accountability.[44]

4.2 Institutional Roles

International institutions play a pivotal role in shaping accountability mechanisms. The United Nations Convention on Certain Conventional Weapons (UNCCW) has convened expert groups to deliberate on AWS, reflecting both consensus and division among states.[45] The ICRC has consistently advocated for "meaningful human control" as a normative safeguard.[46] Regional institutions, such as the African Union, support a legally binding instrument on AWS and stress meaningful human control.[47] National governments, including India's, acknowledge the need for definitional clarity but remain cautious about binding commitments.[48] The diplomatic tone of these debates underscores the importance of balancing humanitarian imperatives with technological innovation and state sovereignty.

4.3 Ethical Views

Ethical inquiry enriches accountability discourse by grounding legal mechanisms in moral reasoning. Walzer's just war theory insists that lethal decisions must remain within human moral agency.[49] Kantian traditions emphasize dignity and restraint, warning against dehumanization in warfare.[50] Contemporary

---

[41] Rome Statute of the International Criminal Court, July 17, 1998, 2187 U.N.T.S. 90.
[42] *Id.* arts. 25, 28.
[43] Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837 (2015).
[44] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment (ICTY 1999).
[45] U.N. CCW, *Report of the Group of Governmental Experts on Lethal Autonomous Weapons Systems*, U.N. Doc. CCW/GGE.1/2019/3.
[46] ICRC, VIEWS ON AUTONOMOUS WEAPON SYSTEMS (Geneva 2016).
[47] African Union, Remarks by Dr. Alhaji Sarjoh Bah, Director of Conflict Management Directorate in the Political Affairs, Peace and Security Department, to the Workshop on Autonomous Weapons System: An ECOWAS Perspective (Freetown, Apr. 17, 2024).
[48] Tejas Bharadwaj & Charukeshi Bhatt, *India's Normative Stance on Lethal Autonomous Weapons Systems*, CARNEGIE INDIA (Feb. 26, 2024).
[49] MICHAEL WALZER, JUST AND UNJUST WARS (4th ed. 2006).
[50] IMMANUEL KANT, THE METAPHYSICS OF MORALS (translated ed., Cambridge Univ. Press) (1797).

scholars such as Crootof and Docherty advocate embedding ethical safeguards into treaty law, ensuring that accountability mechanisms reflect both legal and moral imperatives.[51] The human touch here is vital: accountability is not only about codified rules but about affirming the dignity of victims and preserving the moral conscience of the international community.

AWS challenge accountability by diffusing responsibility and complicating enforcement. Yet the Rome Statute, international institutions and ethical traditions provide pathways to safeguard humanitarian values. Accountability must remain human-centered, ensuring that victims of unlawful attacks receive justice and that humanitarian law retains its credibility. This section underscores the need for proactive institutional reform and ethical integration, reaffirming that accountability is both a legal necessity and a moral obligation.

5 FINDINGS, CONCLUSION AND SUGGESTIONS

This final section synthesizes the research problems explored in the preceding sections, offering findings, practical suggestions and a concluding reflection. The analysis underscores that AWS pose doctrinal, institutional and ethical challenges to IHL. Yet, these challenges also present opportunities for reform, innovation and reaffirmation of humanitarian values.

5.1 Findings

5.1.1 Responsibility Gap in AWS Deployment

The study establishes that AWS create a significant responsibility gap by diffusing liability among programmers, commanders and states. Unlike traditional weapons, where accountability flows through a clear chain of command, AWS introduce multiple nodes of responsibility, complicating attribution. The doctrine of command responsibility, as codified in Article 28 of the Rome Statute, presupposes human subordinates whose actions can be controlled and sanctioned.[52] In AWS deployment, however, the "subordinate" is a machine, often operating unpredictably due to algorithmic learning. This unpredictability undermines the principle of effective control, raising doubts about whether commanders can be held liable for outcomes they neither intended nor foresaw. Scholars such as Matthias highlight the "responsibility gap" as a structural deficiency in current law.[53] The human dimension is crucial: victims of unlawful attacks deserve justice and justice requires identifiable human accountability. Without reform, the deterrent function of IHL risks erosion, leaving civilians vulnerable and commanders shielded by technological ambiguity.

5.1.2 Inadequacy of IHL Principles in Regulating Autonomous Targeting

The principles of distinction and proportionality, foundational to IHL, demand human judgment that machines cannot replicate. Distinction obliges parties to differentiate between combatants and civilians,[54] yet AWS rely on sensors and algorithms that may misclassify targets in complex environments. Empirical studies reveal that contextual cues such as surrender gestures or civilian behaviour are difficult to encode

[51] Bonnie Docherty, *Mind the Gap: The Lack of Accountability for Killer Robots* (Human Rights Watch Briefing Paper 2015); Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837 (2015).
[52] Rome Statute of the International Criminal Court art. 28, July 17, 1998, 2187 U.N.T.S. 90.
[53] Andreas Matthias, *The Responsibility Gap*, 6 ETHICS & INFO. TECH. 175 (2004).
[54] Additional Protocol I to the Geneva Conventions art. 48, June 8, 1977, 1125 U.N.T.S. 3.

into machine logic.[55] Proportionality requires balancing anticipated military advantage against collateral civilian harm,[56] a calculus that involves moral and contextual judgment. AWS, by contrast, rely on algorithmic models that risk oversimplifying complex realities, reducing moral safeguards to mathematical equations.[57] While some scholars, such as Arkin, propose "ethical governors" to enforce proportionality,[58] critics argue that algorithms cannot replicate human empathy or moral reasoning.[59] The inadequacy of IHL in regulating AWS thus lies not in the absence of principles but in their reliance on human judgment. Without meaningful human control, these principles risk becoming hollow, undermining the humanitarian essence of IHL.

5.1.3 Absence of Codified Accountability Mechanisms for AWS

Current international law lacks binding frameworks specifically addressing AWS. The Rome Statute and international criminal jurisprudence provide analogical guidance, but they were designed for human actors, not autonomous systems.[60] Institutional fragmentation exacerbates this gap: while the UNCCW has convened expert groups to deliberate on AWS, consensus remains elusive.[61] Reports by Human Rights Watch and Harvard Law School's International Human Rights Clinic emphasize the urgency of codifying accountability mechanisms, warning that delays risk normalizing AWS deployment without safeguards.[62] National frameworks such as India's military doctrine acknowledge humanitarian obligations but remain silent on AWS, leaving doctrinal uncertainty.[63] The accountability gap undermines enforcement, allowing violations to go unpunished and weakening the credibility of humanitarian institutions. The human touch here is vital: accountability is not only about punishing wrongdoers but about affirming the dignity of victims and reinforcing trust in humanitarian law. Without codification, the international community risks failing both legally and morally.

5.2 Suggestions

i. Codify Meaningful Human Control: States should negotiate binding instruments requiring human oversight in AWS deployment.
ii. Expand Command Responsibility: Interpret Article 28 of the Rome Statute to encompass technological oversight, ensuring commanders remain accountable for AWS actions.
iii. Develop National Legislation: Countries such as India should adopt domestic laws regulating AWS, harmonizing constitutional protections with international obligations.
iv. Institutional Reform: Strengthen the role of the UNCCW and ICC in addressing AWS, ensuring coherence across international institutions.
v. Ethical Integration: Embed ethical safeguards into treaty law, affirming human dignity and moral responsibility in technologically mediated warfare.

---

[55] Noel Sharkey, *The Evitability of Autonomous Robot Warfare*, 94 INT'L REV. RED CROSS 787 (2012).
[56] Additional Protocol I to the Geneva Conventions art. 51(5)(b), June 8, 1977, 1125 U.N.T.S. 3.
[57] YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT (2012).
[58] RONALD ARKIN, GOVERNING LETHAL BEHAVIOR IN AUTONOMOUS ROBOTS (2009).
[59] Robert Sparrow, *Killer Robots*, 24 J. APPLIED PHIL. 62 (2007).
[60] Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment (ICTY 1999).
[61] U.N. CCW, *Report of the Group of Governmental Experts on Lethal Autonomous Weapons Systems*, U.N. Doc. CCW/GGE.1/2019/3.
[62] HUMAN RIGHTS WATCH & INT'L HUMAN RIGHTS CLINIC, HARVARD LAW SCH., LOSING HUMANITY: THE CASE AGAINST KILLER ROBOTS (2012).
[63] INDIAN ARMY, DOCTRINE OF THE INDIAN ARMY 112–18 (2004).

5.3 Conclusion

AWS represent both a technological innovation and a humanitarian challenge. This study has demonstrated that responsibility gaps, doctrinal inadequacies and accountability deficits threaten the integrity of IHL. Yet, through codification, reinterpretation and ethical integration, the international community can reaffirm its commitment to humanitarian values. The novelty of this research lies in its synthesis of doctrinal, institutional and ethical perspectives, offering a holistic framework for regulating AWS. Ultimately, accountability in warfare must remain human-centred, ensuring justice for victims and preserving the moral conscience of international law.

* * *

PART III

Artificial Intelligence, Courts and the Delivery of Justice

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.12

CHAPTER 12

Artificial Intelligence in Judicial and Administrative Decision-Making: Reassessing Rights, Regulation, and Justice

SAUDAMINI GUPTA[*] AND PRIYANSHU[**]

Artificial intelligence has entered courts and administrative agencies more quietly, and a good deal more widely, than the public argument about robot judges would suggest. It is at work in bail and sentencing, in tax assessment and immigration screening, and in the daily administration of welfare. This article asks what becomes of due process once statistical inference begins to supplement, and then to shape, official discretion. There are a few recurring problems. Model reasoning is opaque, which sits badly with the principle of reasoned decisions. Training data carries the imprint of past discrimination, which then reappears wearing the borrowed authority of mathematics. And automation bias hollows out individualized judgment even where nobody intends it to. Drawing on State v. Loomis, the Dutch SyRI litigation and the childcare benefits scandal that followed it, Australia's Robodebt scheme, and the framework now emerging in India such as the Draft Regulations for Use of Artificial Intelligence in Courts, 2026, I argue for interpretability calibrated to what is at stake, independent audit, contestability designed in rather than bolted on, and human oversight built to survive a heavy docket.

Keywords: algorithmic decision-making; due process; natural justice; administrative law; explainability; automation bias; risk assessment; AI regulation

1 INTRODUCTION

The debate on the use of artificial intelligence in the justice system is often found in the future tense. People went on envisioning that one day robot judges would start pronouncing judgments. To delve into this, we must first look at the factors that bring forth such ideas. In contemporary times, the citizens have a case against the slow response of the judicial system. Artificial intelligence is seen as an unbiased and effective alternative to bring in reforms to the issue of judicial delays. The use of artificial intelligence as per popular imagination has a shock value attached to it. The underlying reality, however, is less dramatic and considerably harder to regulate. The basis of all machine learning is the raw data that it builds its model upon.

Use of artificial intelligence isn't detached from human intervention. When artificial intelligence is applied to create filtering mechanisms, it would have practical implications. It would mean that the magistrate would read a risk score before ruling on bail. A revenue officer might examine a file because an artificial intelligence classifier flagged it. A social welfare employee may suspend a benefit because the fraud-detection model classified a household beyond the threshold. Each of these situations would have a machine learning model flagging a case and in each such flagging, the final decision would be attributed

[*] Junior Standing Counsel at Income Tax Department.
[**] Junior Research Fellow at National Law University Odisha, Cuttack, Odisha, India.

to a human. Every resultant human action has several steps wherein artificial intelligence intervened and changed the course of decision making. The interventions impacted the attending, framing and default position of the authority. When the end result attains finality, neither the certifying human authority nor the affected person can inspect the parameters of decision.

In the traditional sense, administrative and judicial laws are grounded in the principles of natural justice. This creates a duty upon the decision maker to apply her own mind to the case brought before her. She is required to make a reasoned decision and spell out the basis of her decision. A reasoned decision allows the affected person to know and contest those reasons. These reasons can be tested against the record by a reviewing court. This basic rule is disrupted by the introduction of a predictive system. Under the new arrangement, the order may still state the reasons but there may be no visible connection between the order and the outcome.

It is essential to have an effective legal system. In India, unfortunately, we have a large backlog of cases especially at the lower courts. There are times where a petitioner may die before their claim is decided. In such a scenario, automation systems can serve as an assistive mechanism. They can be helpful to translate judgments, transcribe proceedings, summarise records and triage filings. However, there is a difference in using assistive technology in court administration and using predictive technology for adjudication. It becomes pertinent to distinguish between a system that expands a decision-maker's capacity to reason about a case and a system that applies the general population level inferences to an individual case. The difference lies in the function performed. Any deployment of artificial intelligence must begin with the demarcation where its impact as a productivity tool ends and the role of adjudication begins. In each such case, the principles of natural justice that are applicable to adjudicatory acts must be applied to automated systems.

2 DEPLOYMENT AND RISK ASSESSMENT

As highlighted earlier, the use of artificial intelligence has begun in some areas of the justice system across jurisdictions. In this light, it becomes necessary to understand the level of risks across deployment.

Case Management and Drafting Support – One of the most extensive uses of automated systems is seen for functions such as to index filings, propose cause lists, transcribe hearings and translate judgments. India's SUVAS and SUPACE[1] tools utilise such deployments to further procedural convenience rather than impacting substantive outcomes. However, there lies a risk of mistranslation. For example, a paragraph may have been relied upon by the defense and can be dropped by the summariser. Such a problem may be classified as a quality-control problem.

Triage and Resource Allocation – One of the deployments of automated systems is for the purpose of tax assessment. In this area, while automated systems may not decide the case, they can still decide the people whose files would be examined, returns be audited and claims be investigated. This exercise may impact millions of people while being subjected to negligible checks and balances. When a particular trade or region is flagged disproportionately it marks the onset of an enforcement pattern that would not be revealed in any individual assessment order.

[1] Supreme Court Vidhik Anuvaad Software (SUVAS) and Supreme Court Portal for Assistance in Court's Efficiency (SUPACE); *see* SUPREME COURT OF INDIA, ANNUAL REPORT (2021).

Risk and Needs Assessment – Automated systems are also used to predict recidivism or absconding possibilities. Such predictive inputs end up impacting bail, sentencing, parole and prison classification. At times due to bias in training data, the system attaches a probabilistic statement about a reference class. When a member of such a reference class seeks a remedy, a general probabilistic statement would be applied ignoring the individual differences of the particular case. This has a counterproductive result and ends up increasing the volume of litigation.

Eligibility Determination – There is a growing trend of using automated systems in welfare, subsidy, pension and tax administration for matching, scoring and calculation. Similarly, automated systems are used in visa triage and immigration risk tiering. The key concern lies in the profiles of subjects to such scrutiny. The beneficiaries of these programs are often those who have access to limited resources. In the event of any adverse decision, such people do not have the means to go for an appeal.

The underlying issue is that the harm caused cannot be mitigated by introducing sophisticated systems. For instance, the Australian government came up with the Robodebt scheme for automated debt assessment and recovery. The scheme did not use any machine learning. Yet, it resulted in one of the largest administrative disasters in Australian history. The problem lies with the automation of judgment on a large scale and the reversal of onus of proof. In the real world, the entire process is riddled with a huge power asymmetry. It is viable to see sentencing regulations being litigated because defendants have counsel and appeal rights. On the other hand, welfare based classifications are less likely to be litigated.

3 STRUCTURAL PROBLEMS

3.1 Dichotomy of Natural Justice and Black Box

The idea of natural justice requires the presence of reasoned decisions. This requirement serves a larger purpose by reducing the presence of bias and prejudices and inducing accountability. When decision makers are required to give reasons behind the outcomes, they are subjected to scrutiny and any irrelevant considerations are exposed. They let the affected party know the case against her. The requirement of reasoned orders makes supervisory review possible. The reasons stated in an order are assessed by the higher courts at the stage of appeal. Indian Courts have held that recording reasons is a facet of natural justice in quasi-judicial and administrative decision-making. An order stands and falls on the reasons stated in it.[2]

The introduction of predictive models disrupts the mechanism of checks and balances. It happens in two ways that run parallel. The initial issue is that of technical opacity. The mechanism of deep-network/gradient-boosted ensembles is based on weights. This violates the reasoned decision requirement of administrative law. This works on the idea of post hoc explanation which follows the idea that if the variables are changed, the output would change too. At the outset, the explanation resembles reason but in reality, it is an approximation of one model generated by the other. Rudin argues that such approximations can be unfaithfully precise where the stakes are highest. It is better to use an inherently interpretable model to reach consequential decisions rather than explaining an opaque one after the fact.[3]

---

[2] Kranti Assocs. Pvt. Ltd. v. Masood Ahmed Khan, (2010) 9 SCC 496; Mohinder Singh Gill v. Chief Election Comm'r, (1978) 1 SCC 405.

[3] Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACH. INTELL. 206 (2019).

The other issue is that of legal opacity. The first step in deploying an automated model involves a contract between the vendor and the justice department. The vendors want to preserve their trade secrets and resist disclosing training data, feature sets and validation results. As seen in Loomis' case, the defendant could not examine how COMPAS produced his score because the methodology was proprietary.[4] In the *Houston Federation of Teachers*[5] case, a federal court recognised the impact of secrecy on procedural due process. The dismissed teachers had alleged that procedural due process was not followed because they were dismissed on the basis of proprietary value-added scores and they could not verify or contest their score. The federal court upheld the claim of violation of procedural due process. This reasoning deserves wider adoption since a person would be unable to check the validity of a score even if the most mathematically accurate model is deployed.

The duty does *not* require a complete causal account of the decision-maker's mental processes since human reasons are themselves post-hoc rationalisations to an unknown degree.[6] It requires grounds specific enough to be tested for legal permissibility and rational connection to the outcome. For an algorithmic system this means that the input variables, the logic of their combination, the reference population, and the validation record are disclosable.

3.2 Training Data Bias and Cyclic Discrimination

A supervised machine model needs training data to get the output. With time the model ends up reproducing the patterns gathered through training data. When the input data itself suffers from bias and prejudices, it creates a discriminatory training database. The model reproduces this cycle of discrimination and confers it the authority of statistics. This is usually an indirect mechanism. One may try to resolve it through formally excluding protected characteristics. Such exclusions, however, serve little benefit because characteristics may be constructed from information such as postcode, occupation and prior interaction with police. The basis of the recidivism model becomes rearrest rather than reoffending. In an unevenly distributed policing system, the attention cannot be uniform. The model learns the distribution of police attention as much as the distribution of criminality.

This problem was practically experienced in the COMPAS controversy. There was a ProPublica analysis on trends of risk classification conducted in 2016. It looked at the cases where defendants did not end up becoming repeat offenders and highlighted that in this category, the black defendants were more likely to be classified as high risk. On the other hand, while examining the cases where defendants did end up becoming repeat offenders, it found that white defendants were more likely to be classified as low risk.[7] The vendor defended the model by stating that it was calibrated. This meant that irrespective of which race a person belonged to a given score carried the same empirical reoffending rate. This was used to establish the relevant fairness criterion. The subsequent literature shows that both were right about their own metrics. However, the metrics are irreconcilable. Where base rates differ, calibration within groups and equality of false positive and false negative rates cannot both be satisfied except in degenerate

[4] State v. Loomis, 881 N.W.2d 749, 761 (Wis. 2016), *cert. denied*, 137 S. Ct. 2290 (2017).

[5] Hous. Fed'n of Teachers, Local 2415 v. Hous. Indep. Sch. Dist., 251 F. Supp. 3d 1168 (S.D. Tex. 2017).

[6] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018); Sandra Wachter, Brent Mittelstadt & Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the GDPR*, 7 INT'L DATA PRIVACY L. 76 (2017).

[7] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

cases.[8] The selection of fairness criterion is a normative act. It ends up determining how the burden of predictive error is distributed between social groups. When a vendor's engineering meeting decides this normative question, it ends up circumventing the question of constitutional equality.

This issue can be looked at from the lens of Indian equality doctrine. Article 14 of the Indian Constitution prohibits arbitrariness. In case a classification is to be made, it is permissible only after establishing an intelligible differentia. Such an intelligible differentia must have a rational nexus to the object sought to be achieved. An automated system that is built upon discriminatory data cannot become an intelligible differentia. When cases of violation of equality doctrine are brought in, in absence of disaggregated performance data, a petitioner cannot establish the discriminatory nature of the output made by an opaque process.

An extension of this argument that is often overlooked is the nuanced nature of adjudication. When judges are adjudicating cases, they do not design outcomes based solely on precedents. They perform a normative task that is constantly evolving. They overrule precedents, read down colonial era statutes and reconcile the conception of rights to the changing social realities. An automated model that uses historical data to perform a predictive task is structurally incapable of doing this. It is optimised to reproduce the past distribution of outcomes. Its margin of errors lies precisely in cases where the law ought to correct itself. Allowing automated systems to perform adjudicatory functions does not simply encode the prejudices of the past. It also removes the mechanism through which the justice system can evolve.

3.3 Automation Bias and the Erosion of Individualized Discretion

The subtlest risk in automation lies in the focus on human oversight in frameworks across jurisdictions such as the European Union's, Council of Europe and India's draft judicial regulations. This may end up defeating the safeguards currently being drafted. The empirical literature on human oversight offers very little comfort.

Research on different arenas of automation bias be it cockpit studies, clinical decision support or criminal justice risk management has led to some common observations. Operators tend to accept incorrect recommendations which they would have rejected had they not involved automated systems. They also fail to detect the issues that were not flagged by the system.[9] Both these issues become manifold under time pressure and heavy caseloads. Supervisors are poor at identifying where a score is wrong.[10] This results in a wrong score distorting judgments while the supervisor remains unaware. The oversight requirements are often satisfied on paper without offering any functioning protection in practice. The system is legitimised as being human control and the automation is not put for scrutiny.[11]

The use of automation also creates an accountability asymmetry independent of psychology. If a decision maker departs from a recommendation and is wrong, he is exposed. On the other hand, a decision maker who follows it and is wrong can shift the onus on the system. When an institution audits only

[8] Jon Kleinberg, Sendhil Mullainathan & Manish Raghavan, *Inherent Trade-Offs in the Fair Determination of Risk Scores*, arXiv:1609.05807 [cs, stat] (2016), https://arxiv.org/abs/1609.05807; Alexandra Chouldechova, *Fair Prediction with Disparate Impact*, 5 BIG DATA 153 (2017).

[9] Linda J. Skitka, Kathleen L. Mosier & Mark Burdick, *Does Automation Bias Decision-Making?*, 51 INT'L J. HUM.-COMPUT. STUD. 991 (1999); *see also* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008).

[10] Ben Green & Yiling Chen, *Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments*, *in* PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 90 (2019).

[11] Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, 45 COMPUT. L. & SEC. REV. 105681 (2022).

deviances and not concurrences, it elevates an advisory tool to the position of a de facto rule without any formal expression.

This problem is addressed in administrative law though it is rarely applied to software. The rule against acting under dictation holds that an authority vested with discretion must exercise it itself and may not surrender it to another's direction.[12] The principle that "he who decides must hear" condemns dividing the adjudicative function between the entity that receives the evidence and the entity that reaches the conclusion.[13] An authority that in substance ratifies a vendor's output performs both the functions. The accountability cannot be diluted merely because the dictating authority is a software. On the contrary, the need for accountability is aggravated because a superior officer can at least be asked to state reasons.

4 COMPARATIVE LENS

4.1 United States: The Loomis Case and the Limits of Cautionary Instructions

Eric Loomis was sentenced in Wisconsin in a case of drive-by shooting. The sentencing court had referred to a COMPAS risk assessment in the presentence report. On appeal to the Wisconsin Supreme Court, he argued that due process was violated since he was unable to examine the proprietary methodology. He alleged that the instrument had used gender as an input and produced group based predictions[14].

The court held that a COMPAS score could be considered but could not be used to determine whether an offender should be incarcerated or fix the severity of a sentence. It introduced procedural safeguards wherein any presentence report containing a COMPAS assessment should be accompanied by written cautions such as that the methodology is proprietary and undisclosed, that the instrument was normed on national rather than Wisconsin populations, that studies had questioned whether it disadvantages minority offenders, and that it was designed for correctional rather than sentencing decisions.

The judgment serves as an unintentional demonstration of why disclaimers do not work. The court acknowledged that the tool may be racially skewed and that it was normed on a different population. The sentencing court was still asked to consider it, while forbidden from relying on it decisively. The problem is that there is no cognitive procedure by which a judge can consider the unverified data without being influenced by it. The margin of error cannot be reviewed and practically a defendant can almost never establish that the score was biased.

The Wisconsin Court's reasoning is complicated. Dressel and Farid[15] found COMPAS's predictive accuracy comparable to that of untrained laypeople given a handful of variables and suggested that the gains purchased at such constitutional cost may be modest. Further, the Houston teachers case shows that where a court squarely confronts trade secrecy against the right to contest, the analysis is not difficult.

4.2 The Netherlands: SyRI and the Aftermath

In The Netherlands, the SyRI (*Systeem Risico Indicatie*) was deployed. The data linked to SyRI spanned across employment, tax, housing, and benefits registries. The objective was to generate risk assessment reports and identify individuals who were likely to have committed fraud. It was deployed selectively, in

[12] Comm'r of Police, Bombay v. Gordhandas Bhanji, AIR 1952 SC 16; Purtabpore Co. v. Cane Comm'r of Bihar, (1969) 1 SCC 308.
[13] Gullapalli Nageswara Rao v. Andhra Pradesh State Rd. Transp. Corp., AIR 1959 SC 308.
[14] *Loomis*, 881 N.W.2d at 753–72.
[15] Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, 4 SCI. ADVANCES eaao5580 (2018).

practice concentrating in low-income neighborhoods. The model or the indicators were not published and the individuals were not informed that they had been assessed unless a report was generated.

In February 2020 the District Court of The Hague held that the enabling legislation violated Article 8 of the European Convention on Human Rights.[16] The court accepted that combating fraud was a legitimate claim but held that the legislation failed the fair balance test. The absence of transparency about the model and indicators made it impossible to verify how conclusions were reached or to defend against them, and the system carried an unaddressed risk of discriminatory effect and stigmatization. The UN Special Rapporteur described the case as an early skirmish over the digital welfare state.

The SyRI judgment tests the legislation on the grounds of proportionality under an existing rights framework. The same structure is available under Article 21 in India after *Puttaswamy*.[17] It treats verifiability as a component of legality. In other words, a system that cannot be checked cannot be justified. It also recognises that the harm falls on everyone processed, irrespective of whether they are found liable or not.

Within a year of the SyRI judgment, the Dutch childcare benefits case came up. The Dutch tax authorities had used risk classification to identify possible fraud in childcare allowance claims, treating indicators including dual nationality as markers of elevated risk. As a result, tens of thousands of families were wrongly accused and ordered to repay large sums. The data protection authority found the processing unlawful and discriminatory, and the government resigned in January 2021.[18] The impact of the scandal was huge because of its scale, the presumption of guilt and the impossibility of contesting a classification whose basis was undisclosed. The Netherlands had a strong administrative judiciary, an active regulator, and a fresh precedent. Yet, striking down such a monitoring system could not prevent the harm.

4.3 Australia: Robodebt and the Reversal of Onus

The Australian government operated an automated compliance programme called Robodebt between 2015 and 2019. The program compared annual taxation income data against fortnightly welfare declarations. In cases where the annual figure was averaged across fortnights, irregular earnings, such as casual, seasonal, and gig work, generated apparent discrepancies and hence apparent debts. This led to an automatic issue of notices and the burden of disproving the debt fell on recipients. In many cases there were no corresponding payslips.

The Federal Court held that a debt calculated by income averaging was not properly made. The issue ended with a class action settled for approximately AUD 1.8 billion in debts being written off and compensation paid. A Royal Commission reported in 2023 that the scheme had been unlawful from the outset.[19]

In this case rather than a machine learning model, there lay instead the substitution of a statistical assumption for individual assessment, a reversal of the burden of proof, and the removal of the officers previously required to investigate a discrepancy before asserting a debt.

---

[16] NJCM v. The Netherlands (SyRI), Rb. Den Haag, Feb. 5, 2020, ECLI:NL:RBDHA:2020:1878 (Neth.).

[17] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

[18] AUTORITEIT PERSOONSGEGEVENS, DECISION ON THE PROCESSING OF NATIONALITY BY THE TAX ADMINISTRATION (2021); PARLIAMENTARY INTERROGATION COMM., ONGEKEND ONRECHT (Neth. 2020).

[19] Amato v. Commonwealth, No. VID611/2019 (Fed. Ct. Austl. Nov. 27, 2019) (consent judgment); Prygodicz v. Commonwealth (No. 2) [2021] FCA 634 (Austl.); ROYAL COMMISSION INTO THE ROBODEBT SCHEME, REPORT (Austl. July 2023).

4.4 India: Rapid Adoption, Emerging Governance

India's position is distinctive and, in a sense, more hopeful. Much of the framework is still being written and its jurisprudence on privacy, arbitrariness, and procedural fairness is unusually developed. The manner of adoption within the judiciary has so far been assistive rather than adjudicative. The adoption has been confined, as of now, to translation of judgments into regional languages, case-file analysis, digitization of the district judiciary under e-Courts Phase III, and live transcription piloted at the Supreme Court.

The governance response has been swift. The Ministry of Electronics and Information Technology released the India AI Governance Guidelines in November 2025. The guideline adopted a principles-based, light-touch model anchored in existing law rather than a standalone statute.[20] In June 2026 the Supreme Court's AI Committee published Draft Regulations for Use of Artificial Intelligence in Courts, 2026.[21] The draft creates a classification and states that no AI system may perform the function of adjudication or sentencing. It mandates that deployed systems would only be used in assistive capacity. It requires advocates and litigants to disclose AI use in preparing pleadings, submissions, or evidence. This is in response to the emerging international problem of fabricated citation. It proposes to permanently establish an Apex Body at the Supreme Court. Such a body would set standards, approve tools, coordinate with High Court committees, handle grievances, and publish annual reports. Its scope extends beyond the constitutional courts to tribunals and statutory bodies exercising adjudicatory powers, where high-volume deployment is most likely.

While this serves as a stronger starting position in comparison to most other jurisdictions it still has gaps. The assistive versus adjudicative line is drawn at the wrong place. A system that produces a recommended outcome, a risk tier, or a ranked disposition is assistive in form and determinative in effect. The preferred demarcation should lie where it separates systems that expand the decision-maker's engagement with the individual record from those offering a conclusion derived from a reference class. The latter ought to be prohibited rather than focusing only on formal delegation. This would serve the purposes better.

The judicial framework does not reach the administrative state, where the volume of automated determination is far greater. Revenue, subsidy, and identity-linked benefit systems affect a greater number of people than court-annexed tools. Further, their outputs are typically challengeable only through writ jurisdiction most affected persons cannot access. The Digital Personal Data Protection Act, 2023 contains no equivalent to Article 22 of the General Data Protection Regulation. Thus, there exists no right against decisions based solely on automated processing. In other words, no right to human intervention and no right to contest.[22] With broad powers to exempt State instrumentalities, this leaves the individual materially worse protected than in the European Union. The single most consequential reform available is to fill that gap either by rule, amendment, or judicial development under Articles 14 and 21.

On the practical side, procurement is unregulated as a rights question. The crucial decisions such as whether the model is interpretable, whether the vendor must disclose validation data, whether disaggregated performance is published, who owns the audit trail are made when a contract is signed. The terms fixed in a contract begin impacting rights long before a petitioner enters the courtroom. The certification

[20] MINISTRY OF ELECTRONICS & INFO. TECH., GOV'T OF INDIA, INDIA AI GOVERNANCE GUIDELINES (Nov. 2025).
[21] SUPREME COURT OF INDIA, ARTIFICIAL INTELLIGENCE COMMITTEE, DRAFT REGULATIONS FOR USE OF ARTIFICIAL INTELLIGENCE (AI) IN COURTS, 2026 (June 3, 2026).
[22] Digital Personal Data Protection Act, 2023; *cf.* Regulation 2016/679, art. 22, 2016 O.J. (L 119) 1 (EU).

pathway the draft regulations envisage is the right vehicle for imposing those terms, provided the criteria are published and certification decisions are reviewable.

4.5 The European Union and the Council of Europe

The European Union's AI Act classifies as high-risk both AI intended to assist a judicial authority in researching and interpreting facts and law and in applying the law, and systems used by public authorities to assess eligibility for essential public benefits or to detect fraud in that context.[23] The classification triggers obligations on risk management, data governance, documentation, logging, accuracy, and human oversight. As such public-body deployers must conduct a fundamental rights impact assessment. Post adjudication the affected persons must have a right to an explanation of the system's role in decisions producing legal effects.

The implementation record is itself cautionary. Obligations for standalone high-risk systems were to apply from 2 August 2026. Following the Commission's Digital Omnibus proposal of November 2025, an amending Regulation entered into force on 27 July 2026. The amended regulations deferred the Annex III high-risk obligations to 2 December 2027 and those for artificial intelligence embedded in regulated products to 2 August 2028, with exemptions for systems already on the market.[24] It was justified on the ground that harmonised standards, conformity assessment infrastructure, and national competent authorities were not ready. The binding force of an artificial intelligence statute depends on institutional capacity that takes years to build. The compliance deadlines are political variables rather than fixed points.

The Council of Europe's Framework Convention takes a different route. It requires parties to ensure that activities across the AI lifecycle are consistent with human rights obligations and to provide accessible remedies and procedural safeguards.[25] It can be easily transplanted across jurisdictions. Its rights-first framing serves as the more useful model for a jurisdiction that is unlikely to enact a comprehensive AI statute soon.

The Court of Justice held in *SCHUFA* that automated generation of a probability value can itself constitute an automated individual decision where a third party draws strongly on that value in deciding the individual's rights. This removes the formal distinction between recommendation and decision on the same lines as Part 3 suggests is necessary.[26] In *Bridges* the Court of Appeal held facial recognition deployment unlawful partly because the police force had not taken reasonable steps to satisfy itself that the software did not have a racial or gender bias. It treated the failure to investigate discriminatory performance as itself a breach of the public sector equality duty.[27] Such a decision solves the difficulty of proving algorithmic discrimination by shifting the onus from the claimant and transforming it into an obligation on the public body.

---

[23] Regulation 2024/1689, arts. 6, 14, 27, 86 & annex III, 2024 O.J. (L 1689) 1 (EU).

[24] Regulation 2026/1744 (Digital Omnibus on AI), 2026 O.J. (L 1744) (EU) (in force July 27, 2026).

[25] Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, C.E.T.S. No. 225 (opened for signature Sept. 5, 2024); EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (CEPEJ), EUROPEAN ETHICAL CHARTER ON THE USE OF ARTIFICIAL INTELLIGENCE IN JUDICIAL SYSTEMS AND THEIR ENVIRONMENT (Dec. 2018).

[26] Case C-634/21, SCHUFA Holding AG, ECLI:EU:C:2023:957 (Dec. 7, 2023).

[27] R (Bridges) v. Chief Constable of S. Wales Police [2020] EWCA Civ 1058 (Eng.).

5 LOCATING THE HARMS IN EXISTING DOCTRINE

A recurring assumption that algorithmic decision-making requires wholly new law is often false. Instead, it requires the application of existing doctrine to a new object supplemented with evidentiary rules that make that application possible.

Reasons. An order reciting conclusions while the operative ground was an undisclosed score is, on the *Mohinder Singh Gill* principle, an order supported by reasons other than those stated, and should be challengeable on that basis alone.

Hearing. The right to be heard is meaningless if the case to be met is undisclosed. Where an adverse inference is drawn from a model output, *audi alteram partem* should require that the person be told a system was used, what it produced, on what inputs, and how it was validated.[28]

Non-delegation. Where a statute confers discretion on a named authority, that authority must exercise it. The systematic adoption of model outputs, particularly where deviation carries institutional cost, is discretion exercised under dictation.

Arbitrariness and equality. A learned proxy producing systematically worse outcomes for a protected group satisfies neither limb of the Article 14 test. A state that has not tested for such an effect has not discharged its obligation to act non-arbitrarily.

Proportionality and privacy. After *Puttaswamy*, state action interfering with informational privacy must satisfy legality, legitimate aim, necessity, and proportionality, with procedural safeguards. Data linkage and profiling across administrative databases such as the SyRI pattern falls squarely within this test.

A petitioner cannot plead disparate impact without disaggregated error rates, nor irrelevant consideration without the feature list, nor dictation without deviation statistics. A disclosure obligation is triggered by its reliance, so that documentation, feature sets, validation results, and disaggregated performance data are producible in proceedings, with trade secrecy a ground for protective orders rather than withholding; mandatory retention of decision logs recording the model output, the human decision, and any deviation; and an adverse inference where a body cannot produce this material.

6 A WORKABLE REGULATORY RESPONSE

Increasing transparency through interpretability. Where a system informs a decision affecting liberty, family integrity, immigration status, or subsistence, the model should be inherently interpretable. In practice a scoring rule, decision list, or sparse regression whose operation can be stated in a paragraph and reproduced by hand. Post-hoc explanation should not satisfy the requirement. For triage, post-hoc explanation plus published aggregate performance may suffice whereas for translation or retrieval, verification against source is the appropriate control. The accuracy cost of interpretable models in these domains is, on the evidence, small while the cost of opacity is a right rendered untestable.

Having an independent, published audit. There must be an audit conducted by an entity independent of both the vendor and deploying body. Such an audit should be repeated after any material retraining. Findings should be admissible, and the absence of audit should count against the public body.

Application of human oversight. Given the automation bias evidence, a requirement that a human "review" the output is close to worthless. Effective oversight requires presenting the case record before the

[28] State of Orissa v. Dr. (Miss) Binapani Dei, AIR 1967 SC 1269.

model output, so the human forms an independent view first. There must be recorded reasons for concurrence as well as deviation which would remove the asymmetric cost of disagreement. The concurrence rates should be audited since a rate approaching unity indicates rubber-stamping rather than agreement. The workload standards should make genuine review possible. There should be blind testing of reviewers against known-erroneous outputs. These belong in the certification criteria for any approved tool.

Inclusion of opportunity to contest. The affected person should be notified that an automated system was used, told the output and principal inputs, given a route to correct input data, given a right to fresh determination by a human who has not seen the output, and given reasons addressing her individual circumstances. That last point matters because an explanation stating that people with similar characteristics reoffend at a certain rate is not a reason for deciding individual cases.

Prohibition in certain areas. Some deployments should be off the table irrespective of accuracy. These may include (i) fully automated determination of guilt, sentence, or deprivation of liberty; (ii) automated suspension of subsistence benefits without prior human determination and an opportunity to be heard; and (iii) predictive systems resting on characteristics the constitution treats as suspect, or on proxies not separable from them. The Indian draft regulations' prohibition on AI adjudication should be extended in substance to administrative determinations affecting subsistence.

Procurement as the primary instrument. The state's leverage is greatest at the point of purchase. Standard contractual terms should require interpretability where the tier demands it. It should include disclosure of training data to auditors and to courts under protective order, state ownership of logs, disaggregated performance reporting as a condition of continued payment, and express waiver of trade secrecy objections in proceedings challenging a decision informed by the system.

Increasing institutional capacity and allowing legal aid. None of this functions without institutions able to exercise it. The Apex Body contemplated by India's draft judicial regulations is a promising model provided it is resourced with technical staff rather than constituted solely of judges and officials. It must publish its certification criteria and decisions that are subject to review. The European experience of deferring obligations for want of standards and competent authorities is a warning to build capacity before the deadline. Equally, those most exposed to automated decision-making are least able to litigate. Without funded representation, relaxed standing for public-interest challenges, and a route to aggregate remedies, the rights above will be exercised principally by those who have access to expert counsel. The SyRI litigation happened in the Netherlands because civil society brought it, and the Dutch benefits scandal continued because its victims could not.

7 OBJECTIONS

Uniformity reduces bias. The strongest objection is that discretion is already inconsistent and prejudiced, and that a model is at least uniform and testable. This objection is persuasive but uniformity remains a virtue only where the underlying rule is just. A litigant facing a biased judge may draw a different judge on appeal, while one facing a biased model faces the same model everywhere. Human error is also diverse and partially self-cancelling, whereas model error is correlated across every decision the model touches. A model error would end up converting individual mistakes into systemic ones. The unaudited proprietary model should be replaced with an audited interpretable model and structural guidance.

Scope of interpretation diminishes accuracy. This *claim* may be true in some cases but not always in the tabular, low-dimensional settings typical of justice and welfare administration, where simple models frequently match complex ones. Where a real trade-off exists, it is between marginal predictive gain and a person's ability to contest a decision about her own liberty.

The issue of backlog would be unchanged by regulation. The backlog argument is serious. However, it argues for aggressive investment in exactly the assistive tools this paper does not propose to restrict. What is restricted is the substitution of statistical inference for individual determination, which does not address backlog at all. It rather processes cases faster by not deciding them.

Disclosure would lead to circumvention. Publishing exact feature weights may allow evasion where a system detects deliberate wrongdoing. That justifies tiered disclosure that mandates full access for auditors and courts under protective order, aggregate publication otherwise. This does not lead to secrecy from the affected individual, who learns nothing exploitable from being told why she in particular was flagged.

8 CONCLUSION

Artificial intelligence in its assistive forms is doing valuable work in systems where delay is itself an injustice. The question is which decisions may be informed by statistical inference about a reference class, on what terms of visibility, and with what capacity for a person to answer back.

The present research puts forth some observations. The harm caused cannot be mitigated by technical sophistication. As seen earlier, Robodebt was arithmetic and SyRI a matching system. In practice both were catastrophic because each substituted an assumption for an assessment at scale. It places the burden of correction on the person least able to bear it. It is very common to see formal safeguards failing. Cautionary instructions did not constrain Wisconsin's sentencing courts. Similarly a judgment striking down SyRI did not prevent the Dutch benefits scandal. The comprehensive European statute has had its central obligations deferred by sixteen months for want of institutional readiness. Safeguards depending on a busy official noticing that a machine is wrong will not hold. Existing doctrine relating to the duty to give reasons, the right to be heard, the rule against dictation, the prohibition on arbitrariness, the proportionality standard for privacy intrusions is substantially adequate, and fails not for want of principle but for want of facts. Petitioners lose because they cannot examine the basis of the output.

The reform agenda requires interpretable models where liberty and subsistence are at stake. The price of placing reliance ought to be disclosed. Trade secrecy should be treated as a reason for protective orders rather than silence. There must be independent audits, and the results must be disaggregated and published. Human oversight must be engineered so that it survives contact with a heavy docket, and its performance be measured. The results of adjudication must be notified and explained independently along with providing a route to a second human determination. All these terms should be included in the contract. The agencies must also focus on capacity building initiatives to be able to enforce the obligations.

India is unusually well placed to get this right. Adoption is still mostly assistive, the jurisprudence on arbitrariness and proportionality is strong, and the governance framework is being drafted now rather than retrofitted later. The 2026 draft regulations correctly identify adjudication as non-delegable and

propose a sensible central approval pathway. The need is to extend the same to give the framework evidentiary teeth, without which every right in it is decorative.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.13

CHAPTER 13

Constitutional Dimensions of AI-Driven Judicial Decision-Making

SARIKA BALODA[*] AND KIRTI GOYAL[**]

Integration of rapidly growing artificial intelligence into judicial decision-making poses progressive opportunities and challenges for traditional system of administration of justice. It is witnessing a huge shift in legal systems worldwide. This paper examines how AI-powered judicial decision-making affects the constitutional principles of due process, impartiality and transparency. It also examines capabilities of AI in performing judicial works. The paper throw light on AI based legal start-ups and the utility of AI. It further deals with risks of bias and erosion of judicial discretion. It evaluates how AI tools poses challenges to constitutional provisions. Further comparative analysis, case studies, statistical data and normative inquiry for ethical deployment of AI in judicial decision-making is discussed. The central theme is to deal with advantages, disadvantages, challenges and solutions for proper implementation of AI in judicial decision-making. Lastly, concluded with some remarks for harmonisation of technology with fundamental principles of the constitution.

Keywords: Judicial Decision-Making, Equal Protection, Legal Ethics, Potential Bias, Accountability.

1 INTRODUCTION

AI is defined as follows: "[Artificial intelligence] (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to pre-defined parameters) to achieve the given goal; AI systems can also be designed to learn to adapt their behavior by analysing how the environment is affected by their previous actions".[1]

The struggle for justice has been a foundational element of human civilization. The legal system has been evolved over the centuries. The development of justice finds its base in the idea of equity, fairness, natural justice and the preservation of public order. The advent of Artificial Intelligence (AI) has affected various aspects of day-to-day life of human civilization including its applicability over judicial decision making. The judiciary has been long regarded as an institution of human discretion but is now facing technological transformation. Artificial intelligence, once used only in science fiction, is being actively used in courtrooms around the world nowadays. AI is now mediating decisions on automated mode that is remarkably affecting individuals' lives.

> Machines are taking over the work of people everywhere. In the not too distant past, spell-checking and search engines were regarded by many as 'intelligent' information technology. Today, facial

[*] Research Scholar at Symbiosis International (Deemed University), Pune, Maharashtra, India.
[**] Research Scholar at Amity University Rajasthan, Jaipur, Rajasthan, India.
[1] Pawel Marcin Nowotko, *AI in Judicial Application of Law and the Right to a Court*, 192 PROCEDIA COMPUT. SCI. 2220–28 (2021), https://www.sciencedirect.com/science/article/pii/S1877050921017324 (last visited May 7, 2025).

> recognition routinely checks travellers at our airports. Google Maps gives unsolicited advice about destination: "the restaurant may be closed". Tablet and mobile phone answer spoken questions with friendly spoken replies. The newspapers talk about 'robot' justice. There are claims that algorithms can accurately predict court decisions, and that we won't need human judges anymore. We enjoy talking about things that don't exist yet, and fantasize about how they will make our lives easier.

[2]AI can be described as "allowing a machine to behave in such a way that it would be called intelligent if a human being behaved in such a way".[3] This is the definition that John McCarthy, considered to have invented the term "Artificial Intelligence", gave to AI in 1956. This is important to establish, defining human intelligence as the measure of what AI does. Intelligence is the ability to reason abstractly, logically and consistently, discover, lay and see through correlations, solve problems, discover rules in seemingly disordered material with existing knowledge, solve new tasks, adapt flexibly to new situations, and learn independently, without the need for direct and complete instruction.[4]

This use of AI is expected to reduce human error and enhance access to justice. AI is capable of analysing vast datasets, identifying patterns and making predictions and is hence used in the judicial process now. To recommend a sentence it can analyse evidence and conduct legal research regarding the suitability of the sentence for a particular category of offence and an offender.

Artificial Intelligence (AI), powered by advanced algorithms and machine learning, has become increasingly vital in modern society. It has been integrated into criminal justice systems worldwide, significantly influencing the administration of justice. The adoption of AI is expected to grow substantially in the future, reshaping how justice is delivered. While AI is valued for its perceived reliability and objectivity, it also presents notable ethical and legal challenges. Despite its benefits, concerns have emerged that machine learning may adversely affect the fairness of judicial processes. The incorporation of cutting-edge technology and robotics in AI holds potential to enhance economic and social aspects of human life.[5]

However, its deployment raises certain questions regarding the foundation of constitutional law, like: is it possible for a machine to follow the constitutionally guaranteed justice? Can it reassure the principle of equal protection under the law? Can it be a threat to impartial judiciary? In aggregate it can be said that excessive reliance on AI raises complex constitutional questions that can be a threat to the democratic governance and the rule of law.

In March 2025, Justice B.R. Gavai, who later served as Chief Justice of India from May to November 2025, expressed serious reservations about usage of AI in legal organizations. At a Nairobi seminar on "Leveraging Technology within the Judiciary," Justice Gavai said that although technology has many advantages, its use needs to be carefully considered in light of its drawbacks.[6]

While AI promises enhanced proficiency, stability and access to justice, it also constitutes significant constitutional obstacles. China has been a pioneer in incorporating AI into its court system, hence the constitutional framework of China with special reference to the role of AI-driven judicial decision-making

---

[2] A.D. Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN. (2020), https://iacajournal.org/articles/10.36745/ijca.343, DOI:10.36745/ijca.343 (last visited May 7, 2025).

[3] *Id.*

[4] *Id.*

[5] Stanley Greenstein, *Preserving the Rule of Law in the Era of Artificial Intelligence (AI)*, 30 ARTIF. INTELL. & L. 291–330 (2022), https://doi.org/10.1007/s10506-021-09294-4 (last visited May 9, 2025).

[6] *The Delicate Balance of AI in Judiciary*, https://www.drishtijudiciary.com/editorial/the-delicate-balance-of-ai-in-judiciary (last visited May 7, 2025).

needs to be analysed to frame an outline of Indian law. Apart from it comparative insights from global jurisdictions shall also be considered. AI is the system that produces intelligent results. Nowadays, using AI in the administration of justice has become common. AI has a significant role in the routine activities of tribunals and courts.[7] While using AI it must be ensured that the pursuit of technological advancements does not come at price of cherished legal principles and individual rights.

2 AI IN THE JUDICIARY: APPLICATION AND CAPABILITIES

AI can be useful in many different ways to meet different requirements. Sales talk on AI for courts is abundant. It has been argued that 'it would make it fairer, and moreover, unlike human judges, AI does not get tired and does not depend on its glucose levels to function'.[8] The working speed of the judiciary is enhanced due to use of AI and expedites dispensation of justice. AI assists judges by predicting vital information regarding the case. AI is not only used in the judiciary, but police also use it for investigation purposes. The police can acquire the requisite information by use of AI without any difficulty. The police acquired vital information using AI, i.e., information about the accused, charge sheet filing date, number of witnesses examined during evidence, the reason for adjournment, the quantum of punishment, and compensation awarded.[9] Therefore, analyzing all information through AI helps judges make better decisions and strategies. Across the world, different countries used AI in their judicial proceedings, especially the UK, Austria, France, Finland, and the Netherlands. AI is used in the administration of justice.[10] AI use in the legal domain can give extraordinary assistance to the judges. However, it cannot replace a judge's expertise. Algorithms help to analyze a large set of data to find solutions and make predictions as the AI has predictive nature. This has led to an increase in its use in criminology, forensics, and legal areas. AI technology helps in the translation of documents and legal research. AI assistance supports judges' decision-making and helps in judicial decisions. It has been suggested that robot judges could replace human judges and that hearings could be conducted automatically.[11] AI tools are used in the criminal justice system. AI introduced algorithms informing the decisions about parole, probation, sentencing (acquittal or conviction), and bail.[12]

AI technologies in judicial systems include processing language along with predictive analysis of cases to enhance decision-making and efficiency. These tools access vast datasets to identify patterns, but their reliance on set historical data can lead to bias, often called a "black box". Although AI can simplify tasks and minimize human mistakes, its complex nature requires careful examination to ensure compliance with law and justice. Estonia is the country that has made an experiment with AI-based judges to handle small claim cases. Common applications of AI include:

---

[7] Maryam Mukhtar & Ayesha Siddiqah, *Critically Evaluate the Use of AI in the Judicial Proceedings*, SSRN (Oct. 22, 2024), https://ssrn.com/abstract=4995240 or http://dx.doi.org/10.2139/ssrn.4995240.

[8] A.D. Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN. (2020) (citing D. KAHNEMAN, THINKING, FAST AND SLOW 43 (2011)).

[9] Kongze Zhu & Lei Zheng, *Based on Artificial Intelligence in the Judicial Field Operation Status and Countermeasure Analysis*, MATH. PROBS. ENG'G (2021), https://doi.org/10.1155/2021/9017181 (last visited May 9, 2025).

[10] Stanley Greenstein, *Preserving the Rule of Law in the Era of Artificial Intelligence (AI)*, 30 ARTIF. INTELL. & L. 291–330 (2022), https://doi.org/10.1007/s10506-021-09294-4 (last visited May 12, 2025).

[11] Tania Sourdin, *Judge v Robot? Artificial Intelligence and Judicial Decision Making*, 41 U.N.S.W. L.J. 1114–33, ISSN 0313-0096, https://www.unswlawjournal.unsw.edu.au/wp-content/uploads/2018/12/sourdin.pdf (last visited May 10, 2025).

[12] Christopher Rigano, *Using Artificial Intelligence to Address Criminal Justice Needs*, 280 NAT'L INST. JUST. J. 1–10 (2019) (NCJ No. 252031), https://www.ojp.gov/pdffiles1/nij/252038.pdf (last visited May 7, 2025).

2.1 Case Management and Court Administration

As per the information provided by the Supreme Court of India, Artificial Intelligence (AI), and Machine Learning (ML) based tools are being deployed in case management. They are being used in transcribing of oral arguments in Constitution Bench matters. The AI assisted transcribed arguments can be accessed from the website of the Supreme Court. The competent authority has directed to consider the transcribing of oral arguments on regular hearing days i.e. Thursdays.[13] It helps to reduce the burden of administration and improves the efficiency of judiciary.

2.2 Organising Information

Recognising patterns in text documents and files can be useful, for example when sorting large amounts of cases, or in complex cases that contain a lot of information. An example from the United States of America is 'eDiscovery', an automated investigation of electronic information for discovery, before the start of a court procedure. eDiscovery uses machine learning AI, which learns through training what the best algorithm is that is capable of extracting the relevant parts from a large amount of information. Parties agree which search terms and coding they use. The judge assesses and confirms the agreement. This is a method of document investigation recognised by the courts in the United States and the United Kingdom. The method is faster and more accurate than manual file research.[14]

2.3 Legal Research as well as Document Analysis

AI can be used for legal research when we have to go through thousands of documents, which can be very time-consuming. AI can do the same task within a few minutes and obtain accurate and relevant information.[15] AI can process enormous amounts of judgments and statutes more efficiently and accurately. It can understand and interpret legal language, allowing lawyers and judges to perform more detailed research in less time.

2.4 Predictive Analytics and Sentencing

Predictive analytics has become a cornerstone of data-driven decision-making, enabling businesses and organizations to forecast future trends, behaviours, and outcomes. With the integration of Artificial Intelligence (AI), predictive analytics has reached new heights of accuracy and efficiency. AI-powered predictive models can analyse vast amounts of data, identify patterns, and generate actionable insights, transforming industries like finance, healthcare, and retail.[16] Predictive analytics uses historical data to derive outcomes of proceedings of the court. These tools are based on probabilities in a case or estimate the time required to resolve a dispute.

---

[13] Press Release, PIB Delhi, Ministry of Law and Justice, *Use of AI in Supreme Court Case Management*, https://www.pib.gov.in/PressReleasePage.aspx?PRID=2113224 (last visited May 20, 2025).

[14] A.D. Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN. (2020), E ISSN-2156-7964, https://iacajournal.org/articles/10.36745/ijca.343, DOI:10.36745/ijca.343 (last visited May 7, 2025).

[15] Bhavana Kakad, *AI in Legal Research and Document Review*, IPLEADERS BY LAWSIKHO, https://blog.ipleaders.in/ai-in-legal-research-and-document-review/ (last visited May 22, 2025).

[16] Isabella Agdestein, *The Role of AI in Predictive Analytics*, FOCALX (2025), https://focalx.ai/ai/ai-predictive-analytics/ (last visited May 15, 2025).

2.5 Virtual Hearings and Online Dispute Resolution

Online dispute resolution (ODR) consists of online alternative dispute resolution (OADR) and online courts. OADR is dispute resolution outside the courts, which originally emerged in the mid-1990s as an adjunct to various forms of alternative dispute resolution (ADR) and as a response to disputes arising from the expansion of ecommerce. As a result, it focussed on using technology to resolve customer complaints and sought to support negotiation, mediation and arbitration. Today it may go further and give rise to innovative ways to resolve disputes beyond the traditional categories of ADR. OADR may be privately run or state-sponsored, such as when it forms part of a consumer redress scheme. It may be synchronous (the participants are all present at the same time) or asynchronous (the participants engage with the process at different times), or a combination at different steps in the process. In contrast, online courts form part of the justice system and are therefore subject to institutional norms and legal requirements derived from the nature of the judicial function.[17]

2.6 Translation and Accessibility

AI-powered translation tools can work efficiently to break down language barriers and work accurately in multilingual legal systems. AI can provide real-time translation of court proceedings, documents or testimonies. This feature helps to ensure that all parties understand the proceedings regardless of the language understood by them. In addition, AI tools can generate audio or visual content for individuals with disabilities, making the judicial system more inclusive and equitable.

3,89,41,148 cases are still ongoing at the Taluka or District levels, besides 58,43,113 cases are still waiting at higher courts, according to the most recent NJDG (National Judicial Data Grid). The effectiveness of the judiciary suffers as a result of such pendency, which eventually limits people's availability of justice.[18]

2.7 AI Related Start-Ups[19]

- Spot Draft: This is an AI based start-up by Shashank Bijapur, who is a Harvard Law School alumnus and Madhav Bhagat an ex-Google employee. This AI driven business has the ability to scrutinise legal documents and it reduces the paperwork by giving the customers an option to create business contracts. This revolutionary platform offers its clients an option to draft and sign contracts; it also has features of automatic reminders and payments.
- CaseMine: This is a legal research platform. This start-up aims at using AI technology in order to draw links between different case laws and thus makes it simpler for the legal researcher to have an in-depth and comprehensive research.
- CaseIQ: This machine learning software acts like a legal assistant in researching case laws; it also analyses the legal language and works like an assistant by pointing out any potential points of law

[17] Michael Legg, *The Future of Dispute Resolution: Online ADR and Online Courts*, 27 AUSTRALASIAN DISP. RESOL. J. 227 (2016), https://www.austlii.edu.au/au/journals/UNSWLRS/2016/71.pdf (last visited May 18, 2025).
[18] *Artificial Intelligence in Judiciary*, https://www.drishtiias.com/daily-news-analysis/artificial-intelligence-in-judiciary (last visited May 8, 2025).
[19] Harshul Gupta, *Scope of Artificial Intelligence as a Judge in Judicial Sector*, 3 INDIAN J.L., POLITY & ADMIN., ISSN: 2582-7677, https://www.ijlpa.com/journal (last visited Oct. 2, 2026).

which might be missing, suggests alternative arguments, highlights relevant judgements and case laws for a comprehensive legal research.

- NearLaw: This Mumbai based start-up offers AI based solutions to legal practitioners and law firms. NearLaw is said to be using NLP technologies to assist in legal matters by understanding case rankings.
- Practice League: This is a Pune based legal tech law firm which has used AI capabilities to transform the working model for more than 8,000 lawyers. Reports highlight that this firm is working with tech giants like Google and Amazon in order to weave AI abilities into its working solutions.

2.8 Benefits of AI in the Judiciary

Nowadays, use of data in an electronic form (E-Form) is gaining importance in the community. The electronic form of data is easily and quickly searchable and processable. Across the world, the development of technology is rapidly increasing. Today, AI is used in many fields, including the judicial system. AI is a system designed by humans; this system achieves complex goals. AI is used in the judicial system in different ways; firstly, AI as a tool in the hands of the judge; secondly, AI as an ADR method; and thirdly, AI as the autonomous decision maker.[20]

AI is used in the justice system for "transparency, justice, fairness, non-discrimination, safety and security, privacy and data protection, and human responsibility for the decision". These all are AI's core ethical principles and are universally accepted. When AI is used in judicial procedures, it redefines work and improves work conditions for judicial officers. AI is applicable in numerous areas with positive results. On the other side, AI also carries some risks. In AI technology, the self-generated algorithms contain more complex risks.[21]

The use of AI tools in Criminal Justice Proceedings provides several opportunities to ensure the legal system's predictability, standardization, and transparency. Law enforcement officials and legal bodies involved in law and criminal proceedings must be aware of opportunities AI might provide for effectiveness and development in real-time. The use of AI in judicial field has positive and negative aspects. When the AI algorithms are used in judicial proceedings, they also take some risk factors, e.g., as a judge has to decide upon release of a pregnant woman offender who is at risk of reoffending. In this case, the AI algorithm should rationally provide the decision against such a woman about her release. Still, on the other side, the human judge has the opportunity to make the decision of this case based on new hierarchy of the victims that would be created once offender becomes a mother; hence, this example shows that the AI cannot resolve all issues which are raised in the criminal proceedings. The principles of AI used in criminal judicial proceedings should be ensured or comply with rule of law, the presumption of innocence, and other general principles established in Article 6 of ECHR. Thus, in criminal proceedings, the interested and aggrieved party has the right to claim or challenge the scientific validity of the use of AI Algorithms.[22]

---

[20] Pawel Marcin Nowotko, *AI in Judicial Application of Law and the Right to a Court*, 192 PROCEDIA COMPUT. SCI. 2220–28 (2021), DOI:10.1016/j.procs.2021.08.235 (last visited May 17, 2025).

[21] N A Wickramarathna & EA TA Edirisuriya, *Artificial Intelligence in the Criminal Justice System: A Literature Review and a Survey*, https://www.researchgate.net/publication/358635259 (last visited May 21, 2025).

[22] Maryam Mukhtar & Ayesha Siddiqah, *Critically Evaluate the Use of AI in the Judicial Proceedings*, SSRN (Oct. 22, 2024), https://ssrn.com/abstract=4995240 or http://dx.doi.org/10.2139/ssrn.4995240.

AI can easily access and analyse information and enables courts to resolve cases more quickly and efficiently. AI can reduce inconsistencies in judicial decisions by relying on data and precedent and can improve access to justice for individuals. AI supports judges with comprehensive data, legal research and precedents to enable courts to make reasoned decisions. AI offers innovative solutions to enhance efficiency, transparency and accessibility.

On the other side, when AI is used in the judicial system, it also has some inherent limits. When algorithmic decision-making is used, it is not easy to achieve precision and objectivity. When algorithmic decision-making is used in difficult nature and complex cases, then it is difficult for AI to achieve substantive justice. AI may be able to promote only formal justice.[23]

In the Court proceedings, the AI can advise on the matter. The advisory AI could be useful for the people, the general public, the parties of the case, and legal professionals. AI technology in the Courts not only looks for relevant and legal information. But, this technology also provides answers to questions. Later on, it is the choice of the user whether the user will act on the advice provided by the AI tools or technology. Hence, this advisory function of the AI tools helps people to resolve their disputes by themselves (outside of the court). For example, the Civil Resolution Tribunal (CRT) was established in British Columbia, Canada. This CRT Tribunal deals with disputes related to subsidized housing and strata. This tribunal was provided free public legal information, available 24/7, calculation aids, and solutions explorer. In 2019, it gradually extended its jurisdiction after it proved successful. The CRT builds its expert system after every three months. Thus, updating this system (CRT) is still done by human experts.[24] Another significant benefit of AI is that it can fix the language barrier in between diversities of 22 official languages of India by translating the text in desired language.

2.9 Constitutional Principles and AI

Principle of Fairness, Impartiality and Transparency is the fourth fundamental principle of the European Ethical Charter on using AI in the judicial system. This principle of the Charter is broader as compared with other ones. This principle deals with use of AI in judicial system with transparency, fairness, and impartiality. There is still a need to pay further consideration and attention to the explainable AI and take initiatives regarding fairness, accountability, and transparency in judicial AI. For example, COMPAS Software (algorithm) was also criticized for its bias. According to the Charter, fairness and transparency are the broader concepts, and when AI is used in the judicial system, many rights of people are concerned in the case. Thus, the algorithmic software and another tool of AI used in the judicial system must act fairly, transparently and impartially.[25] There must be some check and balance or any form of accountability on the tools of AI used in the judicial system.[26] Principle under user control is the 5th fundamental principle of the Charter on using AI in judicial proceedings. This principle supports a lawyer's work in applying AI

[23] Zichun Xu, *Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities*, 36 APPLIED ARTIF. INTELL. (2021), https://doi.org/10.1080/08839514.2021.2013652 (last visited May 15, 2025).
[24] Maria Dymitruk, *Ethical Artificial Intelligence in Judiciary*, 22nd International Legal Informatics Symposium Conference (2019), https://www.researchgate.net/publication/333995919 (last visited May 20, 2025).
[25] A. Završnik, *Criminal Justice, Artificial Intelligence Systems, and Human Rights*, 20 ERA FORUM 567–83 (2020), https://doi.org/10.1007/s12027-020-00602-0 (last visited May 20, 2025).
[26] *Id.*

in the judicial system. According to this principle, the algorithms and software may not be used as prescriptions.[27]

Use of AI in courts sparks constitutional debates about fairness, privacy and justice. Risk of bias in AI might weaken independence of judiciary and hinder handling of private information of individuals. Equality is at risk if AI reflects bias in making its decisions. To uphold constitutional values AI must support and not replace judges. The constitutional dimensions of AI in the judiciary are as follows:

2.9.1 Due Process and the Right to Fair Trial

The right of availing fair trial is enshrined within the provisions of Indian constitution. Fundamental theories of natural justice include entitlement to receive a reasoned decision and right to be heard i.e. *audi alteram partem*. When AI systems participate in judicial decisions there is a risk that these rights may be compromised. If the reasoning behind an AI's recommendation or prediction is inaccessible, litigants may be unable to understand or contest the basis of the judicial outcome. Article 21 of the Constitution of India guarantees both entitlement to fair trial and due process. AI techniques challenge the due process when their decision-making processes are opaque. There is a probability of denying the ability to understand or challenge the basis of decisions affecting the liberty. Trust may be damaged by this lack of transparency in the legal system and violate the fundamental principle that justice must not only be done but must also be seen to be done. Furthermore, the reliance on statistical correlations identified by AI might not adequately suit individual circumstances and may lead to decisions that are procedurally unfair.

2.9.2 Equality Before the Law

The principle of identical protection mandates that all individuals must be treated equally under the law. One of the core promises of AI is the elimination of human bias. However, if trained on historic information that reflects societal bias and prejudices related to race, gender, socio-economic status, etc., AI systems may establish rather than eliminate discrimination. This poses a direct threat to the constitutional guarantee of identical protection under the law. Articles 14–18 of our constitution deal with equality before law. AI used in sentencing might disproportionately recommend harsher penalties for individuals. Reducing algorithmic bias is essential to prevent AI from worsening discrimination in our judicial system. Our constitution disallows difference on basis of place of birth, race, religion, sex, caste, etc. An investigation conducted in 2016 found that COMPAS was biased against African American defendants, flagging them as higher-risk at committal of crime nearly twice the rate of white defendants.[28]

Judicial values such as impartiality and equality intersect with AI tools. AI systems have been used in biased ways in relation to marginalized people historically. AI tools' decision-making led to bias and discrimination. Bias in digital systems may enhance in multiple ways. For instance, systems may overfit training data that is not representative of the broader population. Technologies such as intelligent speech processing could be used to replace court reporters and keep a live transcript of the court proceedings. The lack of interpretation creates a risk that financially capable parties would better understand technological systems. Principles of equality and impartiality require not only that like cases are treated alike but

[27] Cath Corinne, *Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges*, 376 PHIL. TRANS. R. SOC. A. (2018), http://dx.doi.org/10.1098/rsta.2018.0080 (last visited May 20, 2025).

[28] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

also that different cases are treated differently. Digital decision-making through machine mechanisms clusters the cases together that need to be treated differently.[29]

2.9.3 Separation of Powers

AI in the judiciary raises questions about the balance of power among all other branches of the government. It has been embodied in Article 50 of the constitution to prevent the concentration of power and safeguard individual liberties. The increasing dependence on AI in jurisdictive decision-taking could potentially blur the lines between the judicial, legislative and executive branches, thereby violating the principle of separation of powers. The separation of powers doctrine ensures that each branch of government operates within its constitutional limits. If algorithmic tools are developed or controlled by the executive agencies or private corporations, there is a high risk that judicial discretion may be curtailed, undermining the independence of the courts.

2.9.4 Judicial Independence or Impartiality

The Indian Constitution establishes an independent and impartial judiciary as a cornerstone of fair adjudication.[30] AI systems, if overly relied upon, may erode the judicial discretion and produce biased decisions. If judges rely upon AI-generated recommendations without critical review they risk becoming mere spectators rather than justice deliverers.

2.9.5 The Right to Privacy and Data Security

Use of AI in judicial processes involves gathering and analysing large amount of sensitive personal information. It raises concerns regarding right to privacy and data security. Strict legal frameworks and technical safeguards are compulsory to prohibit unconstitutional access of and misuse of sensitive data used and processed by AI.

2.10 Comparative Approaches to AI and Judicial Oversight Worldwide

Across the world, different law enforcement agencies used the "crime forecasting software". This software includes the algorithms, i.e., Precobs, HunchLab, PredPol, MapRevelation, etc. These algorithms of AI are also used in the work of criminal courts. Several algorithms of AI are also used to prevent or stop the commission of criminal acts. These algorithms have been identifying where the criminal act can happen or people who can commit a criminal act.[31]

Different countries are using AI in their legal systems in different ways. Some countries are using AI to make judgments more effective. Other countries are experimenting with using AI in more complex ways. Some other countries are using AI to help judges in making decisions in certain types of petty cases. Few countries have created specific laws and guidelines about how AI can be used in their legal system. These laws might set standards for things like fairness, transparency, and accountability. In some other countries judges themselves are taking help of AI. There are some common principles which are

---

[29] F. Bell, L.B. Moses, Michael Legg, Jacob Silove & Monika Zalnieriute, AI Decision Making and the Courts, A Guide for Judges, Tribunal Members and Court Administrators (Australasian Institute of Judicial Administration 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4162985 (last visited May 20, 2025).

[30] Supreme Court Advocates-on-Record Ass'n v. Union of India, (2016) 5 SCC 1.

[31] Maryam Mukhtar & Ayesha Siddiqah, *Critically Evaluate the Use of AI in the Judicial Proceedings*, SSRN (Oct. 22, 2024), https://ssrn.com/abstract=4995240 or http://dx.doi.org/10.2139/ssrn.4995240.

important in all cases; they are human control, due process, explainability and ongoing evaluation. These approaches illustrate the need for specific and principled regulatory frameworks that uphold constitutional norms. Different nations balance innovation concerns and safety through distinct legal frameworks.

Article 10 of UN UDHR, Article 14 of the UN ICCPR, and Article 47 of Charter of Fundamental Rights of EU provide the right to access justice and guarantee fundamental right of "fair trial within a reasonable time and establishing independent and impartial tribunal."[32] Article 14 of ECHR and Article 21 of Charter of Fundamental Rights of EU stated that any discrimination based on religion, language, belief, genetic feature, political opinion, property, age, sexual orientation, colour, race, ethnic, social origin, sex, birth and membership of a national minority should be prohibited. Thus, AI tools in the judicial system are bound to not discriminate against anyone.[33] Article 6 of ECHR also provides guidelines regarding ethics. This Article of ECHR also provides the rules and standards regarding the proper procedure. The main requirement is that AI is useful in the courts and provides a transparent procedure, equality rights provided to parties during proceedings, and well-founded judgment. Applying AI in the courts could reduce the complexity of the judicial system.[34]

In the European Union (EU) a risk-based regulatory framework governs AI. EU's "Artificial Intelligence Act 2024" is the first ever legal AI framework worldwide. It categorises AI systems in 4 levels as per risk, i.e. minimal, limited, high and unacceptable. High-risk system faces dire need of alterations in AI for transparency and accountability.

The United States adopts a decentralized approach and no comprehensive AI law exists there. U.S. courts play a significant role in tackling issues of facial recognition and misidentification. States like California and New York do not have AI in judicial system but introduced AI-specific laws, on privacy and transparency.

China's approach is state-driven and it prioritises national security and social stability. The Cyberspace Administration of China (CAC) enforces strict AI regulations. It requires approval of the government for data affecting public opinion. China's AI governance emphasizes real-time monitoring of AI systems which reflects a collective approach. In India, AI regulation is still somewhere evolving. The government's National Strategy for AI focuses on its adoption in healthcare and agriculture fields but judicial oversight remains fragmented. The Punjab & Haryana High Court referenced ChatGPT in the *Jaswinder Singh v. State of Punjab*[35] case, for general view on bail jurisprudence.

In 2019, the state of Kazakhstan also implemented "AI to predict the court decisions". This system of AI includes the 1.2 million judicial acts and 120 thousand statements of claims. This system of AI helps the judges regarding decision-making. When the judges feel difficulty in decision-making, they will turn to this system for help. The judges put the request in the system. After that, the system will provide 10 top court cases related to the request (facts, application of law, and related circumstances). Then, the judges consider this system's recommendations and apply them to the case's decision-making. In 2016, the Ministry of Justice of the UK introduced the bill known as "Transforming our Justice System". The primary aim and objective of this bill are to digitize the Court proceedings, and it will "play a significant

[32] Maria Dymitruk, *Ethical Artificial Intelligence in Judiciary*, 22nd International Legal Informatics Symposium Conference (2019), https://www.researchgate.net/publication/333995919 (last visited May 20, 2025).
[33] *Id.*
[34] *Id.*
[35] Jaswinder Singh v. State of Punjab, PHHC 068 250 (2024).

role in ensuring that legal systems in England and Wales provide a rapid and certain Judgments; this saves the time and cost of people and also reduces the impact of legal proceedings".[36]

Courts around the world face a tough task in making sense of complex technology while keeping laws fair. Different countries handle this differently. The European Union plays it safe, focusing on preventing problems. The United States leans toward letting the market decide, with less strict rules. China uses strict government control, while India is still building its system. Bringing these approaches together is hard because each country has its own priorities. Still global talks are working to find common and responsible AI.

3 Case Studies and Examples

While deciding a case, the Manipur High Court noted that it relied on Google and ChatGPT 3.5 for further research, indicating a growing trend of AI use in Indian courts, albeit with a general hesitancy similar to worldwide judicial sentiments over AI integration.[37] AI tools cannot replace human judgment and must be used subject to judicial review; this was held in *State v. Loomis*.[38] The European Court of Human Rights has addressed AI in cases involving automated surveillance and data processing. While fully autonomous AI judges are not yet a reality, various countries are experimenting with AI tools that assist in different stages of judicial processes. The following case studies highlight both the potential benefits and the constitutional challenges:

Germany's AI in Case Management and Judgment Drafting: German courts are utilizing AI systems like OLGA (Oberlandesgerichtsassistent) at the Stuttgart Higher Regional Court and Frauke at the Frankfurt District Court to manage case backlogs and expedite judgment in specific types of cases. OLGA assists in case categorization and information retrieval, while Frauke extracts case-specific data and uses pre-written text to make judgments. These tools enhance efficiency.

China's "Smart Courts": China has invested significantly in "smart court" systems that utilize AI for tasks like case filing, legal research and even suggesting sentences based on precedents. These systems aim to improve efficiency and consistency but still concerns exist regarding the transparency and the extent to which they might influence judicial discretion.

Estonia's Robot Judge for Small Claims: Estonia has experimented with an AI-powered "robot judge" to resolve small claims disputes amounting to €7,000, i.e. approximately 6.58 lakh Indian rupees. It aims to reduce the burden of judges and provide speedy resolution for petty cases.

COMPAS in the United States: The COMPAS Algorithm was developed by the private company "Northpointe". The court houses of the USA adopted widely use of AI tool "COMPAS". This tool works as a management and decision-supporting tool. The COMPAS tool was used in Loomis vs. Wisconsin and Kansas vs. Walls. ProPublica analyzed the outcome of the COMPAS Algorithm. Advanced technologies like COMPAS are employed to evaluate an individual's risk in criminal justice proceedings. This tool relies on statistical data to inform decisions. However, some researchers have criticized

[36] Maryam Mukhtar & Ayesha Siddiqah, *Critically Evaluate the Use of AI in the Judicial Proceedings*, SSRN (2024), https://ssrn.com/abstract=4995240 or http://dx.doi.org/10.2139/ssrn.4995240 (last visited May 15, 2025).
[37] *Artificial Intelligence (AI) in Indian Judiciary*, https://www.drishtijudiciary.com/editorial/artificial-intelligence-ai-in-indian-judiciary (last visited May 9, 2025).
[38] State v. Loomis, 881 N.W.2d 749 (2016).

COMPAS, arguing that its algorithm perpetuates unfairness and bias, including issues of discrimination and racism. These flaws can negatively impact a wide range of individuals' rights.[39]

India's SUPACE and SUVAS: The Supreme Court of India has adopted AI-driven tools such as SUPACE (Supreme Court Portal for Assistance in Court's Efficiency) to support legal research, case analysis, and drafting, and SUVAS (Supreme Court Vidhik Anuvaad Software) for translating legal documents. These tools aim to enhance the efficiency of legal research and improve access to justice by addressing language barriers. The e-Courts initiative in India further leverages AI to streamline case management. While these efforts have boosted operational efficiency, commentators have asked whether they include adequate safeguards for data privacy and against algorithmic bias. Former Chief Justice of India, D.Y. Chandrachud, cautioned that AI could deepen disparities in access to justice, potentially creating a system where only the affluent benefit from high-quality legal assistance. Conversely, he acknowledged AI's potential to make justice more accessible to the broader population.

3.1 Case Study 1: Colombia – Guardianship Action Involving AI Assistance

In 2023, a Colombian judge used ChatGPT to prepare a ruling in a guardianship case filed against a health service provider on behalf of a minor suffering from autism. While reviewing the case, the use of AI by the judges raised concerns about the due process and judicial independence in the Colombian Constitutional Court. This incident highlights the potential risks of relying on AI in legal contexts.[40]

3.2 Case Study 2: The Netherlands – Property Dispute and AI-Generated Facts

In June 2024, a Dutch lower court judge used ChatGPT to determine factual details in a property dispute case.[41] The judge's reliance on ChatGPT was criticised and raised questions about the suitability of AI in legal contexts and risks of relying on unverified information. This case highlighted the necessity for strict guidelines and oversight when integrating AI into judicial processes.

3.3 Difference of Opinion Among Indian Courts

Punjab & Haryana High Court's view: Jaswinder Singh, accused of assault resulting in death, was refused bail by the Punjab & Haryana High Court (Anoop Chitkara J.). In the order, the judge also put to ChatGPT a general question on bail jurisprudence where an assault involves cruelty, and recorded its response. The court explained that ChatGPT's contribution was for a more general legal framework rather than a case-specific viewpoint. The High Court's use of AI for legal research to support judicial thinking is demonstrated in this case.[42]

Delhi High Court's view: In a trademark lawsuit, Delhi High Court Justice Pratibha M. Singh decided in favour of Christian Louboutin. Louboutin's legal team used comments generated by ChatGPT to show how the company is known for its "spike shoe style" with a "red sole," which Louboutin contended

[39] Wilson Otitonaiye, *Impact of Artificial Intelligence Algorithm for Passing Judgement in Judicial System* (2020), https://www.researchgate.net/publication/347444168, DOI:10.13140/rg.2.2.22300.10882 (last visited May 18, 2025).
[40] Corte Constitucional de Colombia, Sentencia T-323 de 2024 (Aug. 2, 2024), file T-9.301.656, https://www.corteconstitucional.gov.co/relatoria/2024/T-323-24.htm (last visited Oct. 2, 2026).
[41] Rechtbank Gelderland, judgment of June 7, 2024, ECLI:NL:RBGEL:2024:3636, case no. 10664071 / CV EXPL 23-2321, https://deeplink.rechtspraak.nl/uitspraak?id=ECLI:NL:RBGEL:2024:3636 (last visited Oct. 2, 2026).
[42] *Artificial Intelligence (AI) in Indian Judiciary*, https://www.drishtijudiciary.com/editorial/artificial-intelligence-ai-in-indian-judiciary (last visited May 9, 2025).

Shutiq had imitated. However, Justice Singh disapproved of ChatGPT's use to decide facts or legal matters in court, citing worries over possible errors, fictitious case law, and creative data produced by AI chatbots. This case highlights the value of human judgment over AI-generated content in court rulings and demonstrates the Delhi High Court's cautious approach to integrating AI in legal processes.[43] Still in India there are no exact guidelines as to the use of AI in judiciary though some tools are being used here but it is not used in making decision but only used in taking help.

3.4 Constitutional Framework for AI in Judiciary

To tackle the constitutional disputes posed by AI, it requires a multi-dimensional approach. For which the following framework is proposed:

3.4.1 Developing Legal and Regulatory Frameworks

Existing legal frameworks are not adequate to tackle the exceptional challenges of AI; hence, there is a need for new legislation and judicial interpretations to govern the implementation of AI in judicial context. This includes establishing standards for accountability as well as transparency in AI system. To ensure the compliance with constitutional principles regulatory bodies must necessarily oversee the advancement as well as implementation of AI in judicial applications.

3.4.2 Transparency Frameworks

The constitutional principle is that legal decision must be well reasoned. Transparency and fairness are the important elements of democratic state. People in democratic state must know how decisions are made and on what grounds they are made. Conversely, many AI systems suffer from a lack of interpretability. Steps should be taken to create AI systems that are more transparent, understandable, and user-friendly.

3.4.3 Moderate Algorithmic Bias

Regular updating is essential to minimize bias and uphold right to equal protection. Special measures are necessary to identify bias in the data used by AI systems. This includes employing techniques for bias detection and correction. Legal and ethical guidelines should emphasize the responsibility of developers and users of AI to address bias.

3.4.4 Upholding Human Oversight and Judicial Discretion

AI offers useful insights and suggestions, but it cannot substitute for human judgment and discernment. Judges must maintain ultimate authority and discretion in rendering judicial decisions by retaining ultimate responsibility for making decisions with AI serving only as an assistive tool.

3.4.5 Public Accountability and Legal Redress

The AI systems used in courts should include input from judges, lawyers, civil society and the public to ensure legitimacy and inclusivity. Stakeholders, including defendants, attorneys and the public should have access to information fed in it. Litigants must have the right to challenge AI-assisted decisions, including the logic behind the decisions.

---

[43] *Id.*

3.4.6 Ethical Governance and Interdisciplinary Collaboration

Addressing the complex constitutional and ethical challenges of AI in the judiciary requires harmonisation among legal experts, computer experts, ethicists and policymakers. Tackling the legal and moral issues in use of AI in courts demands teamwork. Open discussions are essential to create balanced solutions that hold technological progress while upholding the values of justice.

3.4.7 Data Privacy and Security

Strong legal and technical safeguards are essential to protect the privacy and security of data used by AI in the judicial context. This includes implementing strict data protection protocols. Effective mechanisms and privacy regulations are required to be framed to redress the case of data breaches or misuse by AI.

4 NEED OF BALANCING CONSTITUTIONAL RIGHTS AND EFFICIENCY OF AI

Artificial intelligence (AI) is transforming industries and enhancing judicial decision-making. However, its rapid adoption raises critical concerns about safeguarding constitutional rights. Balancing AI's efficiency with individual rights and liberties requires careful consideration. Without strict control over AI it may collect and analyse personal data excessively. To address this, governments and organizations must enforce transparent policies and require warrants for sensitive AI-driven surveillance in legal field. Human control over AI must remain as an essential component to prevent overreach of AI in judicial-discretion. AI used in criminal justice such as risk assessment tools for sentencing can perpetuate biases if trained on fake, false or defective data. Strict focus on equal protection principles is necessary to ensure that AI supports justice. A deep scrutiny of work process of AI in judiciary can establish standards for accountability, ensuring it to respect constitutional boundaries. Citizens and advocates too, must stay informed for their rights in an AI-driven world. Ultimately, AI's efficiency should increase and not decrease constitutional protections. Following these norms society can harness AI's potential while upholding the principles that safeguard individual freedoms.

4.1 Statistical Data on Use of AI in Indian Judiciary

More than 36,000 rulings were translated into Hindi and 17,000 into regional languages by the Supreme Court's SUVAS project, significantly improving access for residents who do not know English.[44] Regarding judicial applications, the Digital Personal Data Protection Act (2023) contains conspicuous gaps. It doesn't outline detailed protocols for gathering biometric data in courtrooms, grant third-party vendors access to private case information, or specify how long AI training data should be stored.[45]

Xu reports that an algorithm based on Supreme Court data from 1791 to 2015 predicted the Court's decisions and the Justices' votes with an accuracy surpassing the 66 percent predictive accuracy of jurists.[46]

---

[44] Press Release, PIB Delhi, Ministry of Law and Justice, *Artificial Intelligence in Judiciary* (Aug. 9, 2024), https://www.pib.gov.in/PressReleasePage.aspx?PRID=2043476.

[45] *Artificial Intelligence in the Indian Judiciary: Balancing Efficiency with Constitutional Rights*, https://lawfullegal.in/artificial-intelligence-in-the-indian-judiciary-balancing-efficiency-with-constitutional-rights/ (last visited May 11, 2025).

[46] Zichun Xu, *Human Judges in the Era of Artificial Intelligence: Challenges and Opportunities*, 36 APPLIED ARTIF. INTELL. (2021), https://doi.org/10.1080/08839514.2021.2013652 (last visited May 14, 2025).

A key role of AI in court systems is "prediction." In judicial proceedings, AI can forecast case outcomes, a practice termed "predictive justice" in the U.S. and British legal systems. The main goal of AI implementation is to minimize risks. As cases grow more complex with additional issues and information, risk factors increase. AI helps mitigate these risks in such scenarios. For instance, several predictive tools are commercially utilized in the U.S. A team of American academics developed a machine learning tool that reportedly predicts U.S. Supreme Court (SCOTUS) case outcomes with 70.2% accuracy and the voting behavior of individual justices with 71.9% accuracy.[47]

4.2 Instances of the Recent Expansion of Artificial Intelligence within India's Judicial Framework

1. Supreme Court Vidhik Anuvaad Software (SUVAS): The Supreme Court of India introduced an AI-powered application designed to leverage machine-assisted translation technology. This tool focuses on converting English legal documents and court orders into nine regional languages, marking a pioneering step in integrating AI into India's judicial system.
2. Supreme Court Portal for Assistance in Court's Efficiency (SUPACE): Recently launched by the Supreme Court of India, this AI-driven tool gathers pertinent legal facts and laws, providing judges with streamlined access to critical information to enhance judicial efficiency. It will yield outcomes customized to the specific requirement of the case and the way the judge thinks.[48]
3. e-Courts Initiative: A significant advancement in judicial reforms, this platform has greatly enhanced efficiency by saving time, effort, and costs, providing free online access to critical data such as judgments and court orders for users.[49]

4.3 Benefits and Challenges of Artificial Intelligence in the Legal Sector

The primary benefit of artificial intelligence (AI) in the legal domain is its ability to significantly reduce time spent on tasks. AI can swiftly and accurately address complex legal issues, achieving results in moments that would take humans far longer. Unlike humans, who are prone to errors, AI systems handle intricate legal tasks with precision and process vast amounts of data in seconds. This efficiency reduces costs by minimizing the number of lawyers needed for a case. AI systems are equipped with predictive algorithms and historical case law data, enabling them to identify and assess potential risks in similar legal disputes. This helps lawyers mitigate risks in court and provide more precise advice to clients regarding legal strategies. Additionally, AI operates tirelessly without requiring breaks, unlike humans who need rest to maintain organized and systematic work. As a result, AI enhances productivity and delivers higher-quality outcomes. In a client-centric legal industry, AI's ability to automate repetitive tasks frees up lawyers' time, allowing them to focus on strengthening client relationships and meeting their evolving demands effectively.[50] But like every system, this one also has its own set of disadvantages.

A key drawback of artificial intelligence (AI) in the legal field is its inability to exhibit originality, creativity, or innovative thinking. Unlike humans, machines cannot fully replicate human judgment and

---

[47] A.D. Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN. (2020), https://iacajournal.org/articles/10.36745/ijca.343, DOI:10.36745/ijca.343 (last visited May 12, 2025).
[48] *AI Portal SUPACE*, https://www.drishtiias.com/daily-news-analysis/ai-portal-supace (last visited May 20, 2025).
[49] Gupta, *supra* note 19.
[50] *Id.*

will repetitively execute tasks unless provided with new directives. This makes it challenging for AI to adapt to the ever-changing legal landscape without human guidance. If AI were to replace judges in delivering verdicts, it would rely solely on pre-existing case data, which may not be suitable for every situation, as each case has unique nuances. The integration of AI into the legal sector remains a nascent concept, particularly in developing nations. Many traditional legal professionals in these regions are hesitant to adopt AI, often due to unfamiliarity with the technology. Additionally, in developing countries, technological infrastructure and equipment often lag behind those in developed nations, rendering them less reliable. A significant concern for lawyers is the potential for AI to cause widespread unemployment. As AI becomes more prevalent, its efficiency in handling repetitive, low-skill tasks may lead industries to invest in machines rather than human labor, potentially displacing workers such as paralegals who perform routine tasks. Many of their tasks can now be performed by machines, which poses a threat to human employment. Another significant drawback is the high cost associated with installing, maintaining, and repairing these complex machines—expenses that only a limited number of organizations can bear. Additionally, the ever-evolving technological environment necessitates frequent software upgrades, which demand substantial financial investment. The legal status of artificial intelligence remains ambiguous, as current laws do not clearly define whether AI and machines fall under the scope of common legal frameworks. This ambiguity leads to confusion in legal disputes involving the rights or responsibilities of robots and machines. Furthermore, one of the foremost concerns related to machines is data security.[51] Though employing AI has many advantages it has some disadvantages too but for taking benefit of AI tools they must be used with great caution and concern.

4.4 The Evolving Role of AI in Judicial Processes

The evolving role of AI in the judiciary lies in developing systems that are transparent, explainable, morally and ethically aligned with judicial values. Rather than full automation, the emphasis will likely be on 'human-in-the-loop' models, where AI assists judges but final decisions rest with humans. Judicial training, public oversight and interdisciplinary collaboration between technology and law are essential. Legal frameworks must evolve to oversee the advancement and implementation of AI systems within courtrooms. Additionally, AI has the potential to enhance access to legal services through innovative solutions, such as the use of AI-based legal advisors to help citizens to understand their rights and legal procedures.

There are many challenges to address in the context of AI. One challenge suggests that AI technology is changing and it can be hard for laws to maintain balance. Another challenge is that judges and lawyers are unlikely to always possess the necessary technical knowledge to comprehend how AI systems work. Despite these challenges, it's clear that the legal system is likely to see a growing influence of artificial intelligence. It is therefore necessary to develop effective approaches to implement judicial discretion in AI and by doing so, we can promote the responsible and ethical application of AI which will be consistent with the principles of justice.

4.5 Moral and Legal Issues in Adopting AI within the Judicial System

Over 25 documents outline ethical guidelines for the use of artificial intelligence (AI). The European Commission for the Efficiency of Justice (CEPEJ), an expert body of the Council of Europe, has

[51] *Id.*

addressed the role of AI in judicial systems, recognizing its growing significance. The CEPEJ established five core principles to guide the ethical application of AI in the judiciary, which are as follows: "First is the principle of respect for fundamental rights; second is principle of non-discrimination, third is the principle of quality and security; fourth is principle of transparency, impartiality, and fairness, and last is principle under user control".[52] AI systems can serve dual purposes in courtrooms. Firstly, as "AI assistants," they aid judges by forecasting outcomes and drafting judicial decisions. Secondly, as "robot judges," they could potentially take over from human judges, independently resolving cases in fully automated judicial processes.[53]

The role of artificial intelligence in judicial frameworks promises efficiency, consistency and data-driven decision-making. Use of AI introduces significant ethical and legal challenges in judicial system in India and to cope with it, the judiciary must collaborate with bodies like NITI Aayog to develop AI frameworks specially developed for India to ensure compliance with the Constitutional principles. The challenges for implementing AI in judiciary are:

Bias and Fairness: AI systems utilize past data to forecast results and assist in decision-making, sentencing and bail determination. If this data reflects any form of bias it can increase inequalities and injustice. Algorithm like COMPAS has faced criticism for superfluously flagging minority defendants as higher-risk of being criminals. This violates principles of fairness and equality. It may risk breaking of anti-discrimination laws guaranteed under the Indian constitution. Reducing bias needs frequent checks and balances but perfect fairness is hard to achieve.

Clarity and Interpretability: The opaque nature of AI systems makes decision processes opaque and poses ethical challenges in a legal system that demands transparency. This opacity can erode public trust and violate legal standards. Ethically stakeholders must balance the efficiency of AI with the right of public to understand the rulings. Developing interpretable models and mandating disclosure of the role of AI in making decisions can be a crucial step to achieve transparency.

Accountability and Judicial Autonomy: It becomes hard to determine judicial accountability when AI influences judicial decisions. There is no clarity that if an error occurs who will be liable, developers, courts or judges? Letting machines make decisions can weaken the human discretion in judgments that are key to a fair justice system. Over-reliance on AI could infringe the constitutional principles. Courts must ensure that judges retain ultimate authority in decision-making.

Privacy and Data Security: AI systems require vast data including sensitive and personal information of persons. Mishandling of such data can breach the trust in this system. It may violate data protection and privacy laws, which makes strict consent and security protocols mandatory. Strong encryption must be used to hide personal details. Following global rules can be the key to manage these risks.

Fair Trial: The use of automated decision-making must not compromise with an individual's right to a fair trial, legal representation or of appeal. AI can be used to support but not to replace the human discretion in serious legal matters.

---

[52] Elisa Bertolini, *Is Technology Really Inclusive? Some Suggestions from States Run Algorithmic Programmes*, 20 GLOB. JURIST J. (2020), DOI https://dx.doi.org/10.1515/gj-2019-0065, https://iris.unibocconi.it/handle/11565/4023626 (last visited May 20, 2025).

[53] Ulenaers Jasper, *The Impact of Artificial Intelligence on the Right to a Fair Trial: Towards a Robot Judge?*, 11 ASIAN J.L. & ECON. (2020), De Gruyter, DOI: 10.1515/ajle-2020-0008, https://ideas.repec.org/a/bpj/ajlecn/v11y2020i2p00n1.html (last visited May 18, 2025).

4.6 Influence of AI Technologies on Essential Judicial Principles

Artificial intelligence can both enhance and challenge core judicial principles. These principles, which include open justice, impartiality, equality under the law, accountability, accessibility to justice, and efficiency, are wide-ranging and subject to varied interpretations. They often intersect and influence one another.[54]

4.7 Barriers to Open Justice Posed by AI Tools

The deployment of AI tools in judicial systems introduces significant challenges to the principle of open justice, which emphasizes transparence and public trust in legal processes. These barriers are outlined as follows:

1. Opacity in AI Functionality: AI systems often conceal details about their operations due to factors such as operational confidentiality, protection of proprietary data, and concerns over personal privacy. For example, the creators of the COMPAS risk assessment and sentencing tool have not fully revealed its methodologies or underlying datasets. This lack of transparency was a central issue in the State v. Loomis case (2016), where the court highlighted that the inability to understand the tool's processes posed a major concern. When AI systems remain non-transparent, individuals subject to their decisions are left without clear explanations for court rulings, which undermines open justice and erodes trust in algorithmic outcomes.
2. Inaccessibility to General Audiences: Another significant hurdle is the public's and legal professionals' limited understanding of how AI systems function. Complex source code is often incomprehensible to those without technical training, creating a gap in accessibility. In judicial contexts, providing clear reasoning for decisions is essential for effective communication between courts and parties. The COMPAS algorithm exemplifies this challenge, as its technical nature creates a disconnect between technology and legal language, making it difficult for lawyers to interpret. Without clear insight into how AI-driven decisions are made, parties are left exposed, and the delivery of justice is hindered.
3. Systemic Risks to Accountability: A third obstacle to open justice stems from the broader implications of AI integration, such as the potential for unchecked biases and lack of accountability. Without transparent and auditable AI processes, judicial decisions risk becoming difficult to scrutinize or challenge, further distancing the legal system from public oversight and understanding. Certain AI systems are intricate, and process-based explanations often fail to clarify their outputs effectively.[55]

Issues related to data protection, privacy, human rights, and ethical considerations will persist and become more pressing as AI technology advances, posing new problems and necessitating strict self-regulation on the part of technology creators. The risk of covert ceding of decision-making authority exists when an AI assistant and a human judge coexist. It is crucial to talk about how AI's recommendations could be implemented in the judiciary as soon as feasible to avoid this. Whether to permit AI to assist or replace human judges to collaborate or to institutionally separate them is the key question in this

[54] Maryam Mukhtar & Ayesha Siddiqah, *Critically Evaluate the Use of AI in the Judicial Proceedings*, SSRN (Oct. 22, 2024), https://ssrn.com/abstract=4995240 or http://dx.doi.org/10.2139/ssrn.4995240.
[55] *Id.*

discussion. One of the main problems in this regard is making it possible for human and AI communications to function. The idea that Judge AI will always be unique from other judges may also not be helpful. As was already mentioned, many technologists predict that humans are unlikely to be completely replaced by AI.[56] Excessive reliance over AI tools in courtrooms can be a major threat to open justice. Advancement in AI system and feeding of multidimensional data in the software can help courts to achieve more accurate results and core judicial values.

4.8 Suggestions for Using Artificial Intelligence in Legal Field

AI where on one hand can be a boon for society, on other hand can be a curse if not used with care and caution, especially in legal field. Every case has different facts and issues and it is difficult for AI to provide accurate decisions and hence certain data must be fed so that it can evaluate each and every scenario in a very detailed manner before coming to any decision. Other than this to use AI efficiently a huge amount of personal data is fed into its memory so in case of any kind of cyber-attack on the system there is a huge risk of data leak of both accused and victims and it can affect society adversely, so certain strict encryption methods and security protocols must be developed to protect data. Other guidelines can be as follows:

Ensure that AI services and tools are designed and implemented with a focus on upholding fundamental rights, aligning with core principles such as privacy, equality, and the right to a fair trial. This requires integrating safeguards for these rights at every stage of AI development and deployment.[57]

Promote equitable treatment by preventing discrimination against individuals and groups in algorithmic applications. Bias can originate from the data used by algorithms or be unintentionally embedded within the algorithms themselves. Addressing both data and algorithmic biases is essential to achieve fair and impartial results.[58]

Transparent Data Processing for Legal Protection: To ensure accountability and compliance with legal precedents, data processing methods should be transparent, auditable, and accessible. Algorithm users must disclose their methodologies, data sources, and assumptions in a timely and comprehensive manner. This transparency enables third-party assessment of decisions, data, and reasoning, thereby providing robust legal protection and facilitating judicial review.[59]

The integration of AI tools in the judicial system can compromise accountability due to their opaque nature, which can hinder the effectiveness of traditional checks and balances. The right to appeal, review, and the requirement for judges to provide reasoned decisions may be undermined when AI-driven processes lack transparency, making it challenging to ensure accountability and fairness in judicial decision-making.[60]

---

[56] Mohit Sharma, *India's Courts and Artificial Intelligence: A Future Outlook*, vol. 15, at 99–120 (2023), DOI:10.18690/lexonomica.15.1.99-120.2023, https://www.researchgate.net/publication/377062808_India's_Courts_and_Artificial_Intelligence_A_Future_Outlook (last visited May 18, 2025).

[57] A.D. Reiling, *Courts and Artificial Intelligence*, 11 INT'L J. FOR CT. ADMIN. (2020), E ISSN-2156-7964, https://iacajournal.org/articles/10.36745/ijca.343, DOI:10.36745/ijca.343 (last visited May 12, 2025).

[58] *Id.*

[59] *Id.*

[60] F. BELL, L.B. MOSES, MICHAEL LEGG, JACOB SILOVE & MONIKA ZALNIERIUTE, AI DECISION MAKING AND THE COURTS, A GUIDE FOR JUDGES, TRIBUNAL MEMBERS AND COURT ADMINISTRATORS (Australasian Institute of Judicial Administration 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4162985 (last visited May 18, 2025).

To mitigate the risks associated with AI, systems should be designed to promote fairness and equity, protecting marginalized communities from discriminatory outcomes. The use of transparent and explainable AI tools is crucial in ensuring accountability. Addressing human rights concerns related to AI requires a multifaceted approach, including the establishment of a robust legal framework that guides public authorities in the use of AI. Governments should conduct thorough impact assessments and review existing legislative frameworks, policies, and practices governing algorithmic systems to ensure they align with human rights standards.[61]

5 CONCLUSION

The application of artificial intelligence into the Indian judicial system presents both substantial opportunities and constitutional challenges. AI-assisted technologies can contribute to the effective management of judicial workloads by facilitating case-file analysis, legal research, evidence management and identification of material information. Such applications may assist courts in addressing delays and case pendency and may subsequently contribute in facilitating access to justice. The ongoing digital transformation of the Indian judiciary illustrates the developing use of technological tools in strengthening the administration of justice.

However, the adoption of AI in judicial decision-making cannot be examined only on the basis of efficiency, accuracy, or the technological capability. In India the judicial system is constitutionally placed within such a framework that protects equality, liberty, dignity, privacy, and procedural fairness. Therefore it becomes more crucial that the use of algorithmic systems must remain consistent with the constitutional protections. Algorithmic bias and discriminatory outcomes may raise concerns regarding these protections and the opaque decision-making might create hardships in ensuring procedural fairness and the judicial accountability. Correspondingly, the collection and processing of personal and sensitive information through an AI system might contravene the constitutional privacy.

These concerns become more serious when AI is employed in the criminal justice system, which involves the exercise of extensive coercive powers of the state that include arrest, detention, search, seizure, surveillance and restriction on individual's personal liberty. Following the lessons from COMPAS there is a need to examine the possibility of discriminatory outcomes and the challenges associated with the opaque algorithmic assessments.

Considering the Indian constitutional framework, it is important that the AI consequently remains an assistive tool rather than an autonomous substitute for judicial reasoning. The highest and final responsibility for adjudication must remain with the judges, who shall critically evaluate AI-generated outputs rather than treating them as conclusive. Human supervision is particularly important where AI-assisted recommendations might influence the outcomes of the judicial proceedings. Transparency, comprehensibility, accountability, data protection, and mechanisms for reviewing misguided or discriminatory outcomes are therefore essential to the constitutionally legitimate use of AI in the judicial decision making.

Therefore it becomes imperative that the future of AI-assisted adjudication in India should be guided accordingly by the principle that the technological advancement must operate within and not outside the constitutional structure. Further, AI can strengthen judicial efficiency, consistency and accessibility only

[61] *Id.*

when its deployment upholds constitutional supremacy, judicial independence and other constitutional principles. The object therefore is not the replacement of human adjudication with automated decision-making but the responsible incorporation of AI into judicial decision making in a manner that promises the constitutional promise of justice.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.14

CHAPTER 14

Two Black Boxes: Reassessing Judicial Deference in India's Emerging AI-Driven Security Apparatus

MAHEK MOTWANI*

The rise of artificial intelligence over the last decade has profoundly transformed the character of national security. Threats once confined to land, sea, air, and space are now identified and acted upon across data networks, satellite systems, and financial infrastructures, increasingly by AI systems rather than human analysts. Constitutional democracies, including India, have long relied on judicial deference doctrines to reconcile national security secrecy with the rule of law. Mechanisms such as state secrets privilege and in-camera review assume that classified decisions, though hidden from public view, rest on human reasoning a court could understand if given access. This paper argues that AI-driven security decisions, from predictive threat scoring to automated watchlist generation, break this assumption by introducing a second, distinct layer of opacity. Classification conceals information that remains knowable to a cleared observer; AI conceals reasoning that may be inscrutable even to its own operators, a problem sharpened where systems are foreign-built or privately developed and thus doubly opaque to the state itself. Courts equipped only for the first opacity are unequipped for the second, leaving algorithmic determinations effectively unreviewable regardless of clearance.

Using India as its primary case study, this study contends that this conflation strains rights to privacy, expression, and equal protection under Articles 21, 19, and 14 of the Indian Constitution. Through a comparative analysis of the Anglo-American doctrine of state secrets privilege and the EU AI Act, it argues that traditional tests of legality, necessity, and proportionality cannot be applied when algorithmic logic resists reconstruction. The paper proposes a bifurcated standard of review separating classified-content review from algorithmic-methodology review, allowing courts to evaluate system interpretability independently of classified inputs. It concludes by addressing executive and technical objections to this reform.

Keywords: Judicial Deference, Algorithmic Opacity, National Security Law, Indian Constitutional Jurisprudence, Proportionality Test

1 INTRODUCTION

Artificial intelligence has increasingly been incorporated in decisions once reserved for human judgment within the machinery of national security. It has entered the space of threat scoring, watch-listing, and surveillance in ways that touch upon and in some cases directly concern constitutional law. This shift has occurred with little corresponding change in the legal architecture meant to hold such decisions accountable. Courts across constitutional democracies, including India, continue to rely on judicial deference doctrines built for a different era: one in which national security decisions, however secretive, were understood to rest on human reasoning that a court could, in principle, examine if granted appropriate access.

This paper contends that this situation represents a diminution of rights under the Indian Constitution. Mechanisms such as state secrets privilege and in-camera review were designed to manage the

* Student at SNDT Women's University, Juhu, Mumbai, Maharashtra, India.

concealment of information, a problem of access, not comprehension. Artificial intelligence introduces a second, structurally distinct problem: the *concealment of reasoning*. Where classified information remains knowable to a sufficiently cleared observer, the internal logic of many AI systems, particularly those built on deep learning architectures, may be unreconstructable even to the engineers and agencies that deploy them. Courts equipped only to negotiate access to secrets are institutionally unprepared to confront a system that offers no comprehensible secret to access in the first place. This problem deepens further where such systems are foreign-built or privately developed, denying even the state itself full insight into their design and function.

The consequence is a quiet but significant erosion of constitutional accountability. Decisions that materially affect due process under Article 21, that raise concerns of arbitrary state action under Article 14, and that touch upon the boundaries of executive power, may now be functionally unreviewable, not because courts have chosen to defer, but because deference has been rendered meaningless by the absence of anything left to defer to.

This paper proceeds in six parts. Part 2 describes the traditional architecture of judicial deference in India in matters of national security. Part 3 develops the paper's central theoretical contribution: a distinction between two forms of opacity, classification and algorithmic inscrutability, that current doctrine conflates. Part 4 grounds this distinction in concrete case studies drawn from India's evolving security apparatus. Part 5 examines the constitutional stakes this conflation raises. Part 6 proposes a reform: a bifurcated standard of judicial review that separates classified-content review from algorithmic-methodology review. Part 7 addresses potential objections.

2 THE TRADITIONAL ARCHITECTURE OF NATIONAL SECURITY DEFERENCE

Judicial deference in national security matters has never meant the absence of a legal framework; it has meant the presence of a particular kind of framework, one built to manage secrecy rather than to dissolve it. In India, this framework developed principally around the interception and surveillance powers vested in the executive under Section 5(2) of the Indian Telegraph Act, 1885,[1] which permits interception of communications in the interest of public safety or during a public emergency. In People's Union for Civil Liberties v. Union of India (1997),[2] the Supreme Court, while declining to strike down the provision, read into it a set of procedural safeguards: interception orders were to be reviewed by a committee of senior bureaucrats rather than by courts, and communicated only in summary form. The judgment did not open the executive's reasoning to judicial scrutiny; it substituted an internal, executive-led review for an external, judicial one, on the understanding that the underlying decision, however secret, remained a reasoned act attributable to an identifiable official who could, in principle, account for it.

The same structural logic appears, in a different form, in the Anglo-American doctrine of state secrets privilege. In United States v. Reynolds (1953),[3] the U.S. Supreme Court permitted the executive to withhold evidence from a civil proceeding on the ground that disclosure would endanger national security, requiring only that a responsible official formally claim the privilege and that a court satisfy itself, without necessarily examining the evidence in detail, that the claim was not implausible on its face. Subsequent

[1] Indian Telegraph Act, 1885, § 5(2).
[2] People's Union for Civil Liberties v. Union of India, (1997) 1 SCC 301.
[3] United States v. Reynolds, 345 U.S. 1, 10–11 (1953).

practice, including in-camera review and the use of security-cleared special advocates in the United Kingdom's closed material proceedings, elaborated this basic model without displacing it: a court forgoes independent assessment of classified material, trusting that a human decision-maker examined it and could defend the conclusion reached, if pressed.

What unites these mechanisms, across jurisdictions, is an assumption so basic that it is rarely stated: secrecy conceals information, not reasoning. A classified file might describe a source, a surveillance intercept, or a threat assessment, but the inferential steps connecting that information to a decision, to intercept a call, to detain a suspect, to deny an entry, were until recently drawn by a human analyst using methods a lawyer or judge could reconstruct in principle, even if barred from doing so in practice. Judicial deference, on this view, is not abdication. It is a calculated wager that behind the veil of classification lies an accountable, comprehensible act of judgment. That wager is the doctrinal inheritance now being tested by artificial intelligence.

3 Two Black Boxes: Distinguishing Classification from Algorithmic Opacity

AI in national-security systems introduces a problem that is distinct from the problem of secrecy that has classically been associated with classification. Much of classification, fundamentally, creates a problem of access: the State has information but decides that it cannot be examined by certain parties. With appropriate legal procedure, the information can be accessed and is in principle intelligible and worthy of assessment. In-camera proceedings, confidential submissions and other methods of restricted disclosure are contingent upon this idea.

However, certain AI and machine-learning systems generate outputs by the complex interplay among their models, training data and large numbers of variables. In these systems, understanding the information provided to the system is not sufficient to understand how particular outputs were generated. The issue is not limited to the court's ability to access the information but whether the court can meaningfully assess the process by which the information was fashioned into a recommendation, risk score, threat analysis, or other security assessment. This distinction is important because access and explanation are not necessarily the same thing. A court may, for example, be given access to the relevant data, the model or even technical documentation without being able to determine whether a particular output was sufficiently reliable to justify state action. Whether the output of a system that is used to identify persons as potential security risks is meaningful is a question of both method and accuracy, not whether the system is available to the court.

The quality of the underlying data is therefore an important part of the problem. An algorithm can process information efficiently without that information necessarily being accurate, complete, current or free from systematic bias. If unreliable or inappropriate data contributes to a security assessment, reviewing only the final output may not be sufficient. The court may need to examine the relationship between the data used, the methodology applied and the resulting assessment.

This forms what this paper calls the second black box. In the first black box, information is hidden from access because it is classified. In the second, the form and content of that which is available for analysis by an AI system and its subsequent output which can affect state actions is hidden. These problems can overlap, but are not the same. While a court may be able to regulate access to classified material, it may still struggle to assess the reasoning and reliability of an AI-assisted analysis and subsequent action.

This problem is compounded where the AI system is designed or supplied by a private or foreign entity and the government has limited access to a system's design, training or running. In this scenario, where neither the affected party nor reviewing court can meaningfully assess the confidence or reasoning of a security assessment assisted by AI, what can be used for the purposes of constitutional scrutiny?

4 NATGRID AND THE EMERGING SECURITY ARCHITECTURE

NATGRID provides a useful illustration of the technological changes taking place within India's national-security architecture. It is designed to connect multiple government databases, including immigration, banking, telecommunications and travel records, with authorised security and law-enforcement agencies. According to the Ministry of Home Affairs, NATGRID enables access to information from these different sources to support the identification of suspicious individuals, investigation of terror networks and intelligence-led investigations. As of March 2026, it connects 11 Central User Agencies, police forces across all States and Union Territories, and 11 data-providing organisations.[4]

The system is also developing more advanced analytical capabilities. Government material identifies GANDIVA as an advanced NATGRID analytics tool that enables multi-source data collection and intelligence analysis for counter-terrorism and criminal investigations.[5] The wider Multi-Agency Centre and Subsidiary Multi-Agency Centre network has also been upgraded with AI- and machine-learning-enabled software for real-time analysis of intelligence inputs. These developments show that the use of computational and AI-assisted analysis in India's security architecture is no longer purely hypothetical.

At the same time, the public information available about these systems does not establish that an AI system independently makes final national-security decisions. That distinction is important. The argument of this paper does not depend on treating AI as an autonomous decision-maker. It is enough that AI-assisted systems increasingly participate in the collection, integration and analysis of information on which security assessments and subsequent state action may depend.

This creates the problem identified in the preceding section. When multiple sources of information are integrated and analysed through advanced computational systems, a court reviewing a resulting state action may have to consider more than whether the underlying information is classified. It may also have to ask whether the data used was sufficiently reliable, whether the analytical process was appropriate for the purpose for which it was used, and whether the resulting assessment can be meaningfully examined. The more such systems become embedded in security decision-making, the more important these questions become for constitutional review.

NATGRID therefore illustrates an emerging legal problem rather than an established example of unlawful or autonomous AI decision-making. The question is not whether the system should be presumed defective, but whether existing doctrines of national-security deference are capable of responding if an AI-assisted assessment is later challenged on constitutional grounds.

[4] PRESS INFO. BUREAU, GOV'T OF INDIA, NATIONAL INTELLIGENCE GRID (NATGRID) (2026).
[5] *Id.*

5 Constitutional Stakes

The conflation of these two forms of opacity is not merely a procedural inconvenience; it implicates the core commitments of Indian constitutional law. Three are especially salient.

First, the right to privacy and personal liberty under Article 21.[6] Puttaswamy held privacy to be an intrinsic facet of Article 21 and required that any state incursion upon it satisfy a proportionality standard, encompassing legality, a legitimate state aim, necessity, and a balancing of the measure's impact against its purpose. A proportionality inquiry of this kind presupposes that a court can identify what the measure actually does and why, in order to weigh it against the individual's interest. Where the operative reasoning is algorithmically opaque, courts are asked to conduct a balancing exercise without one side of the scale being visible, testing proportionality against a process rather than a decision.

Second, the guarantee against arbitrary state action under Article 14. Since E.P. Royappa v. State of Tamil Nadu (1974),[7] arbitrariness has been treated as antithetical to equality in the Constitution, and Maneka Gandhi v. Union of India (1978)[8] extended this reasoning to require that procedures affecting personal liberty be fair, just, and reasonable, not merely prescribed by law. An arbitrariness inquiry asks whether a decision follows from an intelligible, non-capricious basis. A determination generated by a system whose own designers cannot fully reconstruct its reasoning sits uneasily within this framework: it may be statistically well-calibrated and constitutionally arbitrary at once, since calibration answers whether a system is accurate on average, not whether any given individual determination is intelligible on its own terms.

Third, the separation of powers. Judicial deference to the executive in security matters has always rested on the premise that courts are ceding review of a human judgment, made by an accountable official who remains, in principle, answerable through other channels, political, administrative, or eventually judicial. Where the operative determination is substantially shaped by a system no official can fully explain, deference risks becoming something else: not a court declining to review executive judgment, but the effective absence of any institution capable of reviewing the judgment at all, since it is not clear the executive itself possesses a reviewable rationale to defer to.

6 Proposed Reform: A Bifurcated Standard of Judicial Review

If the difficulty is a conflation of two distinct forms of opacity, the remedy this paper proposes is their deliberate separation into two independently reviewable questions.

The first, classified-content review, would proceed largely as it does today: courts, through in-camera examination, cleared special advocates, or comparable mechanisms, assess whether the state's justification for withholding underlying information, sources, raw intelligence, sensitive methods, is genuinely necessitated by security concerns. Nothing in this proposal disturbs that inquiry.

The second, algorithmic-methodology review, would ask a different and presently unasked question: independent of whether the underlying data is classified, can the system's design, training methodology, and general decisional logic be assessed for interpretability, proportionality, and freedom from arbitrary or discriminatory pattern by a qualified, appropriately cleared technical auditor, functioning

[6] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[7] E.P. Royappa v. State of Tamil Nadu, (1974) 4 SCC 3.
[8] Maneka Gandhi v. Union of India, (1978) 1 SCC 248, ¶ 48.

independently of the deploying agency? Such review would not require disclosing the specific classified inputs that produced a given output; it would require disclosing, in a controlled and if necessary security-cleared setting, how the system in general terms weighs categories of factors, what its known error rates and failure modes are across demographic groups, and whether its confidence thresholds meet a defensible standard before being relied upon for consequential decisions. This is analogous to auditing a decision-making process rather than adjudicating a particular case, a review of the recipe rather than the specific meal.

This separation resolves the doctrinal confusion identified in Part 3 without requiring courts to develop new expertise in machine learning from first principles; it requires them, instead, to route the second-order question to institutions equipped to answer it, whether specialised technical tribunals, court-appointed independent auditors, or a dedicated oversight body with standing technical capacity, much as courts already route valuation disputes to accountants or medical questions to physicians rather than deciding them unaided.

The most immediate objection is that methodology, too, can be sensitive: disclosing how a security system weighs risk factors could itself furnish adversaries with a map for evasion. This concern is real but not, on inspection, a reason to collapse the two forms of review back into one. A methodology audit conducted by a cleared technical body, operating under the same confidentiality obligations that already govern classified-content review, need not become public to be judicially useful; what matters is that someone institutionally positioned to assess algorithmic soundness does so, and that a court can rely on that assessment, much as it already relies on in-camera findings it does not itself read into the public record.

There is emerging, if still developing, comparative support for a review of this kind. The European Union's Artificial Intelligence Act requires that high-risk AI systems be designed so that their operation is sufficiently transparent to enable those overseeing them to interpret the system's output and use it appropriately, and separately requires human oversight mechanisms capable of allowing overseers to correctly interpret a system's output and its limitations.[9] These obligations remain, in the assessment of much of the surrounding technical literature, more readily achievable for global explanations of a system's general behaviour than for local, case-specific reasoning, precisely the distinction this paper's bifurcated proposal is built to accommodate: global, methodology-level review as a judicially administrable substitute for case-specific algorithmic explanation that many systems cannot presently provide.

7 ANTICIPATED OBJECTIONS AND LIMITATIONS

The institutional competence objection holds that courts, having long deferred to the executive's assessment of security necessity, are still less equipped to assess the soundness of a machine learning system's design. This paper's proposal does not ask courts to acquire this competence directly; it asks them to institutionalise access to those who already possess it, through appointed technical auditors, in the same way courts have long relied on court-appointed experts in patent, medical negligence, or valuation disputes without becoming engineers or physicians themselves.

[9] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act), arts. 13(1), 14(4), O.J. (L 2024/1689) (July 12, 2024).

The technical limitation objection is more serious: some architectures may resist meaningful interpretability regardless of audit rigour; no external methodology review can explain a genuinely inscrutable model beyond describing its statistical behaviour. Where this is true, the honest constitutional conclusion may be that such systems are unsuitable for consequential, rights-affecting security determinations until interpretable alternatives exist, a conclusion this paper does not shy from, since a system that cannot be reviewed cannot be reconciled with a constitutional order premised on accountable state action.

Finally, feasibility varies across jurisdictions with different institutional capacities. India's comparatively nascent technical-audit infrastructure means implementation would likely begin narrower than the full model sketched here, perhaps confined initially to the highest-stakes categories of algorithmic security determination, before wider application. This is a reason for phased implementation, not for abandoning the underlying distinction.

8 Conclusion

Judicial deference in national security law has always rested on a wager that secrecy conceals reasoning, not the absence of reasoning altogether. Artificial intelligence, deployed at increasing scale within India's security apparatus with only nascent statutory or judicial oversight, threatens to make that wager untenable, not because states have become more secretive, but because some of what they now rely upon may not be explicable to anyone, cleared or not. Treating algorithmic opacity as simply a deeper form of classification risks either abandoning meaningful review altogether or demanding an explanation no system can give. Separating classified-content review from algorithmic-methodology review offers courts a doctrinally coherent path between these failures, preserving legitimate secrecy while insisting that the reasoning beneath a security determination, however produced, remain subject to some independent, competent form of scrutiny. As India's security institutions continue to adopt these tools faster than doctrine has adapted to them, this distinction may prove necessary not only to constitutional principle, but to the basic proposition that the exercise of state power, however sophisticated its instruments, remains an exercise for which someone can be asked to account.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.15

CHAPTER 15

Online Dispute Resolution and Technology in the Delivery of Justice

DR. KARAN SINGH GAUR[*] AND DR. GARIMA YADAV[**]

The administration and delivery of justice have been significantly transformed by the rapid development of digital technology, leading to the emergence of Online Dispute Resolution (ODR) as an innovative mechanism for resolving disputes efficiently and effectively. This article examines the extent to which ODR and emerging technologies can enhance access to justice while ensuring fairness, transparency, and procedural safeguards. The study adopts a doctrinal and analytical methodology, drawing upon primary legal sources, judicial decisions, statutory frameworks, policy documents, and secondary literature relating to ODR, e-governance, artificial intelligence, digital courts, and technology-enabled dispute resolution. It also examines recent developments in India, including online mediation, electronic filing, virtual courts, and technology-driven legal services.

The article argues that ODR has considerable potential to enhance access to justice by reducing geographical barriers, procedural delays, litigation costs, and the burden on conventional courts. Technology can facilitate faster communication, automated case management, online mediation, and simplified dispute-resolution processes, particularly in consumer, commercial, family, and small-value disputes. However, technological innovation alone cannot guarantee equitable access to justice. Challenges relating to the digital divide, data privacy, cybersecurity, algorithmic bias, technological literacy, confidentiality, and the enforceability of online outcomes may create additional barriers for vulnerable and digitally excluded groups. Therefore, the effective integration of technology into justice delivery requires a balanced framework that combines technological innovation with judicial oversight, procedural fairness, accessibility, and robust data-protection safeguards. The article concludes that ODR should complement, rather than replace, traditional justice mechanisms, thereby contributing to the development of a hybrid, accessible, and inclusive justice system capable of addressing the demands of an increasingly digital society.

Keywords: Online Dispute Resolution; Access to Justice; Artificial Intelligence; Digital Justice; E-Mediation.

1 INTRODUCTION

The way justice is administered and delivered is witnessing a significant global change. Traditional court systems, which relied on physical courtrooms, paper filings, and complicated rules, are increasingly feeling the pressure. In many areas, systemic delays, exorbitant legal costs, and physical distance create hurdles that make it difficult for ordinary people to seek timely justice. To tackle these challenges, technology is stepping in not just as a helpful tool, but as a transformative force that can redefine how disputes are resolved. One of the standout innovations in this space is Online Dispute Resolution (ODR), which represents a major shift towards making justice more accessible and less tied to physical locations.

ODR includes a broad range of dispute resolution procedures, such as arbitration, conciliation, mediation, and negotiation, which are supported by automated decision-support systems, digital tools, and telecommunications platforms. ODR has quickly spread into larger civil, commercial, consumer, and

[*] Assistant Professor at Anangpuria Group of Institutions, Faridabad, Haryana, India.
[**] Assistant Professor at Anangpuria Law School, Faridabad, Haryana, India.

family law realms. It was first created to handle high-volume, low-value e-commerce disputes where traditional court procedures were functionally impracticable.

The degree to which ODR and related cutting-edge technologies improve access to justice while guaranteeing procedural justice, openness, and constitutional protections is examined in this paper. This study assesses legislative frameworks, court rulings, policy papers, and scholarly literature using a doctrinal and analytical methodology. The e-Courts Project, virtual courts, e-filing systems, and NITI Aayog's policy measures supporting ODR are among the recent advances in India that are highlighted.

In the end, this article makes the case that although ODR has enormous potential to democratize access to justice, technology integration cannot be seen as a cure-all. Complex normative issues, such as the digital divide, cybersecurity threats, algorithmic prejudice, and worries about consent and enforceability, are brought about by the growth of digital dispute resolution. Technology must be included in a well-balanced legal framework that gives judicial oversight, procedural justice, and institutional integrity first priority in order to stop ODR from producing new types of digital exclusion. In order to create a long-lasting hybrid dispute-resolution architecture, ODR should supplement rather than entirely replace conventional justice delivery methods.

2 THEORETICAL FOUNDATIONS AND CONCEPTUAL ARCHITECTURE OF ODR

2.1 Defining Online Dispute Resolution (ODR)

It is crucial to define the conceptual bounds of ODR in order to assess its legal effectiveness. ODR is a separate field that uses technology as an active structural participant rather than just Alternative Dispute Resolution (ADR) done online. Technology was famously envisioned by Ethan Katsh and Janet Rifkin as the "fourth party" in conflict resolution.[1] The claimant, the respondent, and an impartial third party (mediator or arbitrator) are the three parties involved in traditional ADR. Technology functions as a fourth party in ODR, actively influencing communication dynamics, overseeing procedures, evaluating data, and enabling settlement possibilities.

ODR mechanisms come in a variety of technological forms:

1. *Low-Tech ODR:* Electronic document sharing, audio-visual teleconferencing, and email are examples of facilitative communication tools.
2. *Medium-Tech ODR:* Integrated digital platforms that provide asynchronous communications, structured negotiation workflows, and automated intake.
3. *High-Tech ODR:* Sophisticated systems that assess dispute metrics, forecast results, or automate decision-making using artificial intelligence (AI), machine learning algorithms, natural language processing (NLP), and smart contracts.

2.2 Access to Justice: From Physical Presence to Functional Access

Historically, "access to justice" has been understood to mean having physical access to courts and legal counsel. But legal experts like Richard Susskind contend that a functional redefinition of access to justice

[1] ETHAN KATSH & JANET RIFKIN, ONLINE DISPUTE RESOLUTION: RESOLVING CONFLICTS IN CYBERSPACE (2001).

is necessary.[2] The fair, timely, and reasonably priced settlement of their legal complaints is what citizens need, not actual courtrooms.

According to Marc Galanter, traditional litigation frequently fails the functional access test because "repeat players" have a structural advantage over "one-shooters."[3] To defeat individual or economically disadvantaged claimants, well-resourced corporate or institutional litigants take advantage of procedural delays, technical legality, and financial fatigue. By streamlining processes, reducing transaction costs, doing away with travel restrictions, and cutting down on time delays, ODR levels the playing field and operationalizes the fundamental rights protected by constitutional frameworks.

3 The Indian Context: Legislative Frameworks and Policy Initiatives

When it comes to incorporating technology into the administration of justice, India offers an interesting case study. With tens of millions of cases languishing in trial courts and High Courts, Indian institutions are using technology more and more to expedite the resolution of disputes.

3.1 The e-Courts Project and Virtual Courts

Through the e-Courts Integrated Mission Mode Project, the Supreme Court of India's e-Committee established the fundamental framework for technology-enabled justice in India.[4]

The goals of Phases I and II were to digitize court documents, develop standard case management systems, set up e-filing portals, and implement teleconferencing capabilities throughout the country's judicial complexes.

The introduction of Virtual Courts for some high-volume offenses, like traffic infractions and minor summary disputes, was a significant turning point. With no requirement for the litigant or attorney to be present in person, these virtual courts handle all court processes online, from the sending of summonses to the online plea of guilty and the payment of fines through digital payment channels.

3.2 Legislative and Statutory Alignment

Although India does not currently have a single, cohesive statute known as the "ODR Act," the legal legitimacy of ODR is supported by a number of current statutes:

1. *The Information Technology Act, 2000 (IT Act):* Sections 4 and 5 of the IT Act provide digital signatures and electronic records the same legal status as paper documents.[5] Online dispute resolution agreements are supported by Section 10A, which expressly acknowledges the legality of transactions made electronically.
2. According to Section 7(4)(b) of the Arbitration and Conciliation Act, 1996, an arbitration agreement may be reached by letter, telex, telegram, or "other means of telecommunication which provide a record of the agreement." Electronic arbitration agreements are easily accommodated by this statutory provision.[6]

---

[2] Richard Susskind, Online Courts and the Future of Justice (2019).

[3] Marc Galanter, *Why the 'Haves' Come Out Ahead: Speculations on the Limits of Legal Change*, 9 L. & Soc'y Rev. 95 (1974).

[4] e-Committee, Supreme Court of India, The e-Courts Integrated Mission Mode Project, Policy and Action Plan Document for Phase I & Phase II.

[5] The Information Technology Act, 2000, §§ 4, 5, 10A.

[6] The Arbitration and Conciliation Act, 1996, § 7(4)(b).

3. *The Indian Evidence Act, 1872/Bharatiya Sakshya Adhiniyam, 2023:* Section 65B of the Evidence Act (and its modern counterpart) / Section 63 of the Bharatiya Sakshya Adhiniyam, 2023 establishes stringent standards for certifying digital evidence and regulates the admissibility of electronic data in court.[7]
4. *The Code of Civil Procedure, 1908 (CPC):* Section 89 of the CPC gives courts explicit statutory authority to refer parties to out-of-court dispute resolution procedures such as mediation, arbitration, conciliation, and judicial settlement.[8]

3.3 Judicial Endorsement of Digital Proceedings

The Indian judiciary has been actively pushing the limits of justice enabled by technology. The Supreme Court ruled in State of Maharashtra v. Praful B. Desai (2003) that video conferencing is a "procedure established by law" for recording evidence under the Code of Criminal Procedure, proving that physical presence is not a necessary condition for legal proceedings.[9]

In Re: Guidelines for Court Functioning through Video Conferencing During COVID-19 Pandemic (2020), the Supreme Court used its extraordinary constitutional powers under Article 142 to issue binding directives permitting the widespread adoption of virtual hearings across all Indian judicial forums during the COVID-19 pandemic.[10]

3.4 Policy Push: NITI Aayog's ODR Strategy

The groundbreaking paper Designing the Future of Dispute Resolution: The ODR Policy Plan for India was released by NITI Aayog in 2021.[11] ODR should be used as the standard procedure for settling low- and medium-value disputes between government agencies, public sector projects, e-commerce sites, and financial service providers, according to the research. Prioritizing online mediation and conciliation over binding electronic arbitration or traditional litigation, the policy framework promotes a phased deployment paradigm.

4 KEY ADVANTAGES OF ODR IN JUSTICE DELIVERY

Compared to conventional, physical litigation approaches, the institutional transition toward ODR offers the following concrete advantages:

4.1 Mitigation of Geographical and Cost Barriers

By its very nature, traditional litigation penalizes parties according to their socioeconomic status and place of residence. In order to attend court proceedings in urban regions, claimants who live in remote or rural locations must suffer substantial travel costs and lost wages. Litigants can participate in proceedings from their personal devices or nearby digital service centers thanks to ODR, which eliminates the need for geographical location.

---

[7] The Indian Evidence Act, 1872, § 65B; *see also* The Bharatiya Sakshya Adhiniyam, 2023, § 63.
[8] The Code of Civil Procedure, 1908, § 89.
[9] State of Maharashtra v. Praful B. Desai, (2003) 4 SCC 601.
[10] *In re* Guidelines for Court Functioning through Video Conferencing During COVID-19 Pandemic, (2020) 6 SCC 686.
[11] NITI AAYOG, DESIGNING THE FUTURE OF DISPUTE RESOLUTION: THE ODR POLICY PLAN FOR INDIA (2021).

ODR also significantly reduces administrative expenses. For both private litigants and the state, the total cost of resolving disputes is decreased by doing away with paper documents, court rental spaces, and ongoing in-person court appearances.

4.2 Reduction of Procedural Backlogs

ODR offers a scalable solution for institutional backlogs by optimizing dispute resolution procedures. Dispute platforms can process thousands of claims at once because of automated case intake, algorithmic document sorting, and asynchronous interactions. ODR enables prompt, structured remedies without taxing court dockets in sectors including consumer complaints, auto accident claims, credit card defaults, and small business invoice disputes.

4.3 Asynchronous Communication and Flexibility

The requirement for synchronous participation—all parties, attorneys, and judges must be present in the same physical location at precisely the same time—is a structural limitation of traditional court cases. Asynchronous communication modules are commonly used by ODR platforms. In order to accommodate people with rigorous work, caregiving, or physical accessibility restrictions, litigants may submit evidence, make arguments, and evaluate settlement conditions whenever it is convenient for them.

5 Critical Challenges and Systemic Safeguards

Despite its transformational potential, incorporating technology into the administration of justice presents significant logistical, technological, and constitutional problems. These problems run the risk of replacing conventional obstacles to justice with digital ones if they are not resolved.

5.1 The Digital Divide and Socio-Economic Exclusion

The digital divide is the main threat to ODR deployment in emerging countries like India. Rural, economic, and demographic segments all have unequal access to digital infrastructure. Marginalized groups may be effectively prevented from taking part in ODR by problems like erratic internet connectivity, restricted access to internet-capable devices, and low levels of functional technology literacy.

Digitally excluded parties may experience default judgments or feel pressured into unfavorable settlements if a dispute resolution procedure is moved entirely online without offline options. Instead of acting as a constrictive barrier, justice systems must make sure that digital channels continue to be an extra route to justice.

5.2 Algorithmic Bias and the "Black Box" Problem

Algorithmic fairness is a major challenge as ODR systems incorporate AI capabilities for risk assessment, document analysis, and outcome prediction. Legal datasets from the past are used to train machine learning algorithms. The AI tool runs the risk of codifying and magnifying historical prejudices, cultural biases, or structural inequities in these datasets under the pretense of scientific objectivity.

Furthermore, sophisticated machine learning systems function as "black boxes," meaning that human operators are unable to audit or explain the exact logical processes leading to a particular result.

Fundamental natural justice principles, particularly the entitlement to a reasoned, clear decision and audi alteram partem (hear the other side), are compromised by this lack of transparency.

5.3 Data Privacy, Confidentiality, and Cybersecurity

Sensitive personal information, confidential correspondence, financial records, and proprietary business documents are frequently involved in dispute resolution. ODR platforms are appealing targets for illegal spying, data breaches, and cyberattacks.

ODR service providers are required by international standards like the General Data Protection Regulation (GDPR) and data protection regimes like India's Digital Personal Data Protection Act, 2023 (DPDP Act) to strictly implement end-to-end encryption, stringent access controls, and transparent data retention policies.[12]

A fundamental component of mediation, confidentiality needs to be safeguarded against technical flaws, illegal recording, and the secondary commercial monetization of dispute data.

5.4 Procedural Due Process and Enforceability

The enforceability of settlements and awards made via digital methods is a significant legal obstacle in ODR. Even if statutory frameworks accept electronic signatures and agreements, it is still challenging to validate the absence of coercion, ensure informed consent, and confirm party identity in asynchronous, virtual contexts.

Platforms must use multi-factor authentication, verifiable digital signatures, and transparent digital audit trails in order for ODR results to have legal weight. Online mediation solutions must also be readily convertible into legally binding agreements or decrees under the relevant mediation statutes.

6 THE FUTURE HORIZON: TOWARDS A BALANCED, HYBRID JUSTICE FRAMEWORK

A balanced framework is required to fully utilize technology in the administration of justice while protecting constitutional rights. Justice policy should strive for an integrated, hybrid design rather than trying to completely replace traditional courts.

6.1 Multi-Tiered Dispute Resolution Design

A multi-tiered strategy based on party dynamics and dispute complexity should be used in future justice systems:

1. *Tier I (Automated/Self-Service):* Automated negotiation platforms and guided self-help tools are used to resolve low-value, high-volume consumer or transactional disputes.
2. *Tier II (Technology-Assisted E-Mediation):* Using secure ODR platforms, impartial human mediators settle medium-value business, family, or employment disputes.
3. *Tier III (Hybrid Judicial/Arbitral Determination):* Traditional judges or arbitrators use virtual hearings and e-filing technologies when necessary to settle complex, high-value, or innovative constitutional legal concerns.

[12] The Digital Personal Data Protection Act, 2023; *see also* Regulation (EU) 2016/679 (General Data Protection Regulation).

6.2 Institutional Standards, Certification, and Ethical Guidelines

Regulatory agencies must set stringent operational guidelines to guarantee procedural integrity among private and public ODR service providers. These guidelines ought to require:

- Algorithms that are independently audited or open-source to guarantee impartiality.
- Strict adherence to data minimization rules and strong cybersecurity processes.
- Required "human-in-the-loop" safeguards that guarantee human inspection and appeal of automated or algorithmic outcomes.
- Compliance with accessibility standards for people with impairments, such as screen reader compatibility and user-friendly UI/UX design.

6.3 Public Digital Infrastructure and Kiosks

Governments should set up physical access points with high-speed internet, secure devices, and qualified technical facilitators, such as legal aid centers or digital kiosks at local administrative offices, to close the digital divide. This guarantees that citizens who are digitally illiterate or from low-income backgrounds can access ODR platforms with help, maintaining fair access to justice for all groups.

7 Conclusion

A major change in the administration and provision of justice is represented by online dispute resolution and new legal technologies. ODR offers an efficient way to democratize legal remedies and lessen the load on existing courts by reducing structural barriers associated with geographic distance, procedural delay, and exorbitant litigation expenses.

But technology is a tool, not a perfect answer. Without structural protections, ODR's indiscriminate proliferation runs the risk of escalating digital exclusion, jeopardizing data privacy, and introducing systematic prejudice through unreliable computational tools. Access to justice is guaranteed by the constitution, which calls for procedural justice, equity, and institutional integrity in addition to speed and convenience.

Therefore, a balanced, hybrid framework is the key to delivering justice in the future. Legal systems can create a contemporary, accessible, and inclusive justice system that successfully satisfies the needs of a contemporary digital society by integrating technology with strict judicial control, institutional norms, strong data protection, and focused public infrastructure.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.16

CHAPTER 16

Artificial Intelligence, Online Dispute Resolution and the Future of Justice Delivery: Opportunities, Risks and Regulatory Challenges

DR. YASHASVI SHARMA[*] AND DR. TEJASVI SHARMA[**]

The fast-paced development of Artificial Intelligence (AI) and digital technologies is reshaping the methods of handling disputes and providing new avenues for delivering justice. Online Dispute Resolution (ODR), which was initially devised as a technological tool for Alternative Dispute Resolution (ADR) procedures, has become one of the mechanisms that enables the solution of disputes outside of the courtroom. The implementation of AI capabilities in the sphere of ODR through the use of automated negotiation, dispute categorization, analysis of documents, legal research, prediction of outcomes, and decision-making has further expanded the capabilities of ODR. The main question arises whether innovation may enable better accessibility and efficiency of justice without undermining its principles of fairness, transparency, and accountability.

In this paper, the topic of the transformational potential of AI-based ODR in the process of justice delivery will be discussed with particular attention paid to the Indian legal system. The role of AI-based dispute resolution in cost reduction, overcoming geographical barriers, speeding up the process, enabling multilingualism, and assisting mediators, arbitrators, and disputing parties will be examined. Simultaneously, the issues with algorithmic approach to dispute resolution such as algorithmic bias, lack of explanation, data privacy issues, cybersecurity threats, unequal digital access, lack of accountability for AI results, and excessive automation will be considered. Furthermore, the compatibility of the AI-driven dispute resolution process with the constitutional and procedural principles of equality, access to justice, natural justice, impartiality and requirement of fair and reasonable procedure will be assessed.

The study will also consider the sufficiency of the legal and regulatory framework of the ODR, data protection and artificial intelligence in India and draw comparative lessons on the matter from international practices. It will be argued that AI should not be regarded as a replacement of courts, judges, mediators, and arbitrators but only as an assistive technology working within the human-centered framework of justice delivery. Some suggestions will be made in regard to increased algorithmic transparency, independent auditing, data protection, human control, challenge of AI-assisted decisions, platform accountability and inclusiveness of digital infrastructure. Finally, it will be concluded that the future of AI-enabled justice depends not just on the level of technological development, but on the ability of law and institutions to preserve the precedence of the core principles of fairness, dignity, equality and access to justice.

Keywords: Artificial Intelligence, Online Dispute Resolution, Access to Justice, Digital Justice, Algorithmic Bias, Natural Justice, Data Protection, AI

1 INTRODUCTION

The nexus between technology and justice in terms of ODR is changing conventional ways of dealing with disputes. ODR is the use of computer software in Alternative Dispute Resolution (ADR) processes including negotiation, mediation, and arbitration, thereby broadening its scope from online commercial

[*] Assistant Professor at Rajdhani Law College, Jaipur, Rajasthan, India.
[**] Assistant Professor at Kanoria School of Law for Women, Jaipur, Rajasthan, India.

dispute resolution to other places and institutions.[1] The use of Artificial Intelligence (AI) further augments the ODR process since this enables document analysis, pattern recognition, and communication.[2]

These, however, present serious questions regarding ethics and legality of the matter at hand. Despite the aims of ODR being greater efficiency and convenience, the concerns about fairness, transparency, and accountability cannot be overlooked.[3] ODR should be guided by the same principles established by the UN Commission on International Trade Law, such as due process and transparency, while taking into account the risks of creating bias within the ODR system through the use of AI to make decisions due to the "black-box" issue of using AI in this case.[4]

Data protection, confidentiality and cybersecurity issues are important, as ODR systems process sensitive data. India's Digital Personal Data Protection Act, 2023 provides a legal framework for digital data processing which in itself highlights the necessity of confidentiality to retain the concept of ADR processes.[5]

However, the digital divide is a factor to consider when technology is not equally accessible to different groups and could affect the effectiveness of ODR in resolving disputes.[6] Unequal access to technology also poses challenges to the inclusivity of justice access in India. The Mediation Act, 2023 includes certain provisions on online mediation, showing legal acceptance of technology in the process of dispute resolution in Indian legislation.

The conversation is about the right use of AI technology in the dispute resolution process: not only its ability to work but the way it is used.[7] Ideally, there should be a combination of human control and advanced technology in order to create the digital justice system with emphasis on constitutional values as well.[8]

The upcoming research will elaborate on this topic, examining the possibility of using AI technology for the dispute resolution process in order to achieve a more efficient and fair justice system.

2 AI AND ODR: UNDERSTANDING THE CONCEPT

The use of technology in dispute resolution has resulted in a completely new perspective regarding the methods that can be employed in managing, preventing and resolving conflicts. As an attempt to introduce alternatives to litigation procedures, ADR included methods like negotiation, mediation and arbitration, but ODR brought technology to this area of law. With the advent of AI, the scope of using technology in ODR became even more sophisticated as technology evolved from being a method of

---

[1] K. Mania, *Online Dispute Resolution: The Future of Justice*, 1 INT'L COMP. JURIS. 76–86 (2015), https://doi.org/10.1016/j.icj.2015.10.006.
[2] Mappingadr, *Artificial Intelligence and the Future of Online Dispute Resolution in India*, MAPPING ADR (Apr. 17, 2024), https://jgu.edu.in/mappingADR/artificial-intelligence-and-the-future-of-online-dispute-resolution-in-india/.
[3] P. J. Naidu & B. N. S. Sriram, *Online Dispute Resolution (ODR) in India: Opportunities, Challenges, and Future Prospects*, INT'L J. LEGAL SCI. & INNOVATION (2024), https://ijlsi.com/article/view/online-dispute-resolution-odr-in-india-opportunities-challenges-and-future-prospects#fn-22.
[4] N. N. T. Vy, *AI Implementation in ODR: A Game-Changer or a Troublemaker of Data Protection*, 8 VIETNAMESE J. LEGAL SCI. 1–24 (2023), https://doi.org/10.2478/vjls-2023-0001.
[5] A. Gupta & Dr. A. Dixit, *Data Privacy and Implications for India's Online Dispute Resolution Economy*, 2025 INT'L J. ECON. PRAC. & THEORIES (IJEPT) 129–30, https://ijapt.org.
[6] Mappingadr, *supra* note 2.
[7] Darin Thompson, *Creating New Pathways to Justice Using Simple Artificial Intelligence and Online Dispute Resolution*, 2015 INT'L J. ONLINE DISP. RESOL. 1 (Osgoode Legal Stud. Rsch. Paper No. 27/2015, June 30, 2015), SSRN, https://ssrn.com/abstract=2696499.
[8] Mediate.com, *The Role of AI in Alternative Dispute Resolution: Mediation and Arbitration*, MEDIATE.COM (Mar. 14, 2025), https://mediate.com/news/the-role-of-ai-in-alternative-dispute-resolution-mediation-and-arbitration/.

communication into a more complex system of dispute resolution based on algorithms and information. It is important to comprehend the connection between ODR and AI in order to analyze the future prospects of delivering justice.[9]

2.1 Concept of AI

AI refers to the concept of technological developments in relation to machines, which are used to perform tasks that otherwise required human cognitive abilities such as learning, reasoning, problem solving, perceiving, language processing and decision making.[10] Unlike normal computer programmes where operations are usually done on pre-programmed instructions, modern AI may be capable of finding patterns in large amounts of data and providing output based on the available data and model. Some of the technologies that contribute to the growing abilities of AI include machine learning, natural language processing, deep learning and generative AI.[11]

The connection between AI and the law sector is due to the ability of AI to deal with large amounts of data faster than any human individual can do. The legal process requires a lot of documents such as laws, court decisions, agreements, pleadings and evidence.[12] Through AI, these data can be sorted and analysed; thus, patterns can be found and summaries can be made or predictions provided. Therefore, AI has been seen not just as a management tool but also as a decision support tool in the legal and dispute resolution processes.[13]

2.2 Concept of ODR

Online Dispute Resolution is a general term used to describe the resolution of disputes using information and communications technology.[14] Online Dispute Resolution has significant connections to ADR but is not limited to just taking ADR online. Technology can affect the way in which parties communicate, exchange information, negotiate, mediate, and where necessary, resolve their disputes either in a binding or non-binding way.

UNCITRAL's Technical Notes on Online Dispute Resolution define ODR as a method for settling disputes using an online platform.[15] The Technical Notes identify negotiation, facilitation, and adjudication as some stages of an ODR process.[16] Thus, the main element of ODR is not only the physical

---

[9] Carrie J. Menkel-Meadow, *Is ODR ADR? Reflections of an ADR Founder from 15th ODR Conference, the Hague, the Netherlands, 22-23 May 2016*, 3 INT'L J. ONLINE DISP. RESOL. 4–7 (2016) (UC Irvine Sch. of L. Rsch. Paper No. 2017-01, Jan. 4, 2017), SSRN, https://ssrn.com/abstract=2893919.

[10] Homero Gil de Zúñiga, Manuel Goyanes & Timilehin Durotoye, *A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research*, POL. COMMC'N 18 (2023), 10.1080/10584609.2023.2290497.

[11] Ijraset, *Artificial Intelligence (AI): Evolution, Methodologies, and Applications*, IJRASET, https://www.ijraset.com/research-paper/artificial-intelligence-ai-evolution-methodologies-and-applications.

[12] Amjad Mohammad Badah, Camellia Najeh Khalaf, Fatima Jamil Dwaikat & Naji AlQbailat, *The Usage of Artificial Intelligence in Legal Translation: Bridging the Gap Between Law and Language*, 16 AMPERSAND 100248 (2026), ISSN 2215-0390, https://doi.org/10.1016/j.amper.2025.100248, https://www.sciencedirect.com/science/article/pii/S2215039025000323.

[13] A. Bhaskar, S. Kumar, A. Uniyal, A. N. Tripathi, R. H. Zaidi, A. Gairola, S. Z. Qurashi & A. Mittal, *Introduction to the White Paper*, WHITE PAPER (2025), https://cdnbbsr.s3waas.gov.in/s3ec0490f1f4972d133619a60c30f3559e/uploads/2025/11/2025112244.pdf.

[14] ETHAN KATSH & JANET RIFKIN, ONLINE DISPUTE RESOLUTION: RESOLVING CONFLICTS IN CYBERSPACE (2001).

[15] *Understanding ODR, the Future of Dispute Resolution*, RICH. J.L. & TECH. (Mar. 4, 2022), https://jolt.richmond.edu/2022/03/04/understanding-odr-the-future-of-dispute-resolution/.

[16] UNCITRAL, TECHNICAL NOTES ON ONLINE DISPUTE RESOLUTION ¶¶ 1–4 (2016).

absence of the parties in a courtroom but the application of technology in the dispute resolution process itself.[17]

In terms of the development of ODR, it may be viewed as the next stage of evolution in dispute resolution techniques. Conventional litigation implies that the court is an essential element of dispute resolution, while in the case of ADR, individuals have alternative dispute resolution options which could be more flexible, consensual and efficient. ODR technique is built on the idea of using digital technologies to improve dispute resolution procedure.[18]

One could illustrate the evolution as follows:

Litigation → ADR → Digital ADR → ODR → AI-ODR

It is necessary to note that the change from ADR to ODR does not mean that the principles of negotiation, mediation or arbitration could be changed. In other words, the role of technology here is limited to changing the environment in which these processes should occur. While face-to-face communication was previously used, now it could be conducted through video conferencing, electronic communications or special digital platforms. Individuals are able to exchange documents electronically, conduct hearings remotely and record agreements and settlements electronically.

In this context, ODR is about more than just the "online transfer" of ADR; rather, it is about the opportunity to redesign the dispute-resolution process based on the strengths of technology. The NITI Aayog ODR Policy Plan for India acknowledges that ODR can help provide an accessible, effective, and affordable method for dispute resolution, and sees technology as a means for transforming the way disputes are resolved.[19]

ODR covers various dispute resolution methods such as online negotiation, mediation, and arbitration. In a simple online negotiation, parties can communicate using a digital platform to try reaching a settlement. Mediation online will mean the facilitation of communication between parties through digital means with a third party acting as a neutral.[20] Online arbitration can entail the presentation and examination of evidence and arguments online, culminating in the rendering of an award by an arbitrator based on the governing rules of law and procedure.

Online Dispute Resolution (ODR) applies digital communication and computational tools to negotiation, mediation, arbitration, and related processes.[21] It emerged from disputes generated by electronic commerce, where conventional courts were slow and expensive relative to the value of many claims.[22] In India, extensive case pendency, limited judicial capacity and unequal geographic access have encouraged increasing interest in technology-assisted resolution, which reflects not merely administrative

---

[17] Carrie J. Menkel-Meadow, *Is ODR ADR? Reflections of an ADR Founder from 15th ODR Conference, the Hague, the Netherlands, 22-23 May 2016*, 3 INT'L J. ONLINE DISP. RESOL. 4–7 (2016) (UC Irvine Sch. of L. Rsch. Paper No. 2017-01, Jan. 4, 2017), SSRN, https://ssrn.com/abstract=2893919.

[18] M. C. R. Ballesteros & J. L. G. Ávila, *Online Dispute Resolution Platforms (ODR): A Legal and Technical Perspective*, 4 L. & BUS. 28–38 (2024), https://doi.org/10.2478/law-2024-0006.

[19] NITI AAYOG, DESIGNING THE FUTURE OF DISPUTE RESOLUTION: THE ODR POLICY PLAN FOR INDIA (2021), https://www.niti.gov.in/node/687.

[20] D. Sulistianingsih, A. a. R. Lembang, Y. P. Adhi & M. S. Prabowo, *Online Dispute Resolution: Does the System Actually Enhance the Mediation Framework?*, 9 COGENT SOC. SCI. (2023), https://doi.org/10.1080/23311886.2023.2206348.

[21] J. Lande, *When AI Comes to the Table: How Tech Tools Will Change ADR*, UNIVERSITY OF MISSOURI SCHOOL OF LAW SCHOLARSHIP REPOSITORY, https://scholarship.law.missouri.edu/facpubs/1267/.

[22] D. Carneiro, P. Novais, F. Andrade et al., *Online Dispute Resolution: An Artificial Intelligence Perspective*, 41 ARTIF. INTELL. REV. 211–40 (2014), https://doi.org/10.1007/s10462-011-9305-z.

modernization but a broader shift toward justice delivery that can operate beyond the physical courtroom.[23] AI may support this shift by assisting with dispute classification, information management, negotiation and procedural guidance, while leaving adjudicatory authority subject to human and legal oversight.[24] The central question is therefore whether technological efficiency can be reconciled with fairness, transparency, consent and meaningful access to justice.[25]

3 OPPORTUNITIES PRESENTED BY AI-DRIVEN ODR IN DISPUTE RESOLUTION

The integration of Artificial Intelligence (AI) into Online Dispute Resolution (ODR) offers transformative opportunities to traditional dispute resolution methods. ODR improves accessibility and efficiency by minimizing challenges posed by geographical boundaries and different procedures, whereas AI adds to the effectiveness through automation, data analysis, and intelligent support.[26] The key opportunities offered by AI-based ODR are:

1. Increased Justice Accessibility: The AI-driven ODR system removes accessibility barriers through remote participation and maneuvering through complex processes. This will help in overcoming accessibility concerns, especially for individuals with limited resources and mobility challenges.[27]
2. Reducing Costs in the Process of Dispute Settlement: Automation of routine and boring tasks in the process of dispute settlement through the use of AI leads to lower costs of the litigation process, particularly in cases of small value disputes. There will be fewer legal consultations required in such cases.[28]
3. Expedited Dispute Resolution: The use of AI makes document scrutiny and communication easier and enables a shift towards the constant process involving data.[29]
4. AI-assisted Negotiations: Based on the analysis of positions, interests, and offers, AI technology can enhance negotiations by proposing some ways to solve disputes based on common interests of parties participating in negotiations while maintaining voluntary nature of negotiations.[30]
5. Helping Mediators and other Dispute Resolution Professionals: AI serves as an assistant to professionals in structuring and organizing the information, highlighting positions and producing documents in order to enable them to pay attention to the relational side of mediation.

---

[23] Pitamber Yadav, *Digitizing Justice: The Case for Dedicated Online Dispute Resolution Legislation in India* (May 22, 2024), SSRN, https://ssrn.com/abstract=4840167, http://dx.doi.org/10.2139/ssrn.4840167.

[24] Leah Wing, *Artificial Intelligence and Online Dispute Resolution Systems Design: Lack of/Access to Justice Magnified*, 4 INT'L J. ONLINE DISP. RESOL. 16–20 (2017), SSRN (Sept. 30, 2017), https://ssrn.com/abstract=3665409.

[25] D. Singh & K. Devi, *Online Dispute Resolution (ODR): A Paradigm Shift in Access to Justice*, 7 INT'L J. FOR MULTIDISCIPLINARY RSCH. (2025), https://doi.org/10.36948/ijfmr.2025.v07i02.40938.

[26] N. N. T. Vy, *AI Implementation in ODR: A Game-Changer or a Troublemaker of Data Protection*, 8 VIETNAMESE J. LEGAL SCI. 1–24 (2023).

[27] Dr. S. S. Khatana & Dr. S. Choudhary, *Role of Artificial Intelligence in Enhancing Access to Justice in India - A Critical Analysis*, 23 LEX LOCALIS - J. LOCAL SELF-GOV'T 1770–78 (2025), https://doi.org/10.52152/qytz3640.

[28] Agus Agus, Sudirman Sudirman & Ahmad Rustan, *The Use of Artificial Intelligence in Dispute Resolution Through Arbitration: The Potential and Challenges*, SASI 570–78 (2023), 10.47268/sasi.v29i3.1393.

[29] Yatharth Garg, *AI and Alternative Dispute Resolution (ADR): Automating Arbitration and Mediation*, INDIAAI (2024), https://indiaai.gov.in/article/ai-and-alternative-dispute-resolution-adr-automating-arbitration-and-mediation.

[30] J. Zeleznikow, *Can Artificial Intelligence and Online Dispute Resolution Enhance Efficiency and Effectiveness in Courts*, 8 INT'L J. FOR CT. ADMIN. 30 (2017), https://doi.org/10.18352/ijca.223.

6. Intelligent Dispute Classification and Triage: AI is able to categorize disputes by different characteristics such as the nature of the case and its complexity and send them to appropriate settlements right at the beginning of the process.[31]

Beyond all this, AI ODR also improves the decision-making processes in legal disputes through prediction of the possible outcomes and their impact on settlements but not replacing legal decision-making. Moreover, it takes care of multilingual issues in Indian linguistic environment with the help of natural language processing for the purposes of ODR. ODR is helpful in cross-border disputes in the sense that it permits remote participation and the use of AI for translation and legal material interpretation. In addition, ODR with the help of AI can help take off some burden from courts as it resolves the suitable disputes outside the court and thereby helps create a multi-tier justice system.[32] The data from AI may help improve the dispute resolution process, moving towards prevention from reactive actions. The user-friendly AI system will help make the process more convenient.

In brief, AI-driven ODR is going to help enhance justice delivery by making it more accessible, cheaper, effective, and inclusive.

4 RISKS AND CHALLENGES OF AI-DRIVEN ODR

Incorporation of AI into ODR provides several advantages in terms of speed, access, and efficiency; however, it also raises numerous issues on the legal, ethical, and institutional levels.[33] Unlike traditional ODRs, in AI-enabled ODR there may be certain issues regarding the classification, evaluation, and resolution of disputes, because of the risk of algorithmic bias that arises due to biased data.[34] In such a way, the application of artificial intelligence technology in dispute resolution may cause discrimination unintentionally, which is why it is important to ensure that algorithms are impartial.[35]

The problem of the "black box" that arises due to the complexity of the algorithm's decision-making process can cause additional difficulties.[36] The participants of a dispute may have difficulties in understanding why some decisions were taken, thus questioning the issue of procedural justice as the parties have to be able to submit their arguments and understand how they are being analyzed.[37]

Over-automation of the process of ODR is one more risk, as it may remove human judgment that takes into account emotional elements of the disputes. The problem of attribution of responsibility is also complicated due to the involvement of multiple entities in the process of making an error.[38]

---

[31] J. Zeleznikow, *Using Artificial Intelligence to Provide Intelligent Dispute Resolution Support*, 30 GROUP DECISION & NEGOT. 789–812 (2021), https://doi.org/10.1007/s10726-021-09734-1.

[32] Vidhikarya, *The Role of AI in Online Dispute Resolution Opportunities and Challenges* (Jan. 2, 2026), https://www.vidhikarya.com/legal-blog/the-role-of-ai-in-online-dispute-resolution-opportunities-and-challenges.

[33] Robert Buckland, *AI, Judges and Judgement: Setting the Scene* (2023), https://www.hks.harvard.edu/centers/mrcbg/publications/awp/awp220 (last visited June 30, 2026).

[34] A. A. Khaliq, S. S. Ahmed & S. Zafar, *Artificial Intelligence in ODR Platforms: Ethical and Legal Concerns* (Feb. 9, 2026), https://assajournal.com/index.php/36/article/view/1501.

[35] *Id.*

[36] N. N. T. Vy, *AI Implementation in ODR: A Game-Changer or a Troublemaker of Data Protection*, 8 VIETNAMESE J. LEGAL SCI. 1–24 (2023), https://doi.org/10.2478/vjls-2023-0001.

[37] J. Remis, *Ethics in Online Dispute Resolution: Balancing Technology and Human Judgment*, ODR.COM (May 22, 2025), https://odr.com/ethics-in-online-dispute-resolution-balancing-technology-and-human-judgment/.

[38] G. P. N & R. V. KT, *A Novel Pathway to Justice Using Artificial Intelligence and Humans Oversight in AI-Driven Alternative Dispute Resolution*, INT'L J. LEGAL SCI. & INNOVATION (2024), https://ijlsi.com/article/view/a-novel-pathway-to-justice-using-artificial-intelligence-and-humans-oversight-in-ai-driven-alternative-dispute-resolution.

Handling of sensitive information in ODR also requires appropriate privacy and cybersecurity solutions in order to address the possible threats connected to data leakage and unauthorized access.[39] In addition, equal accessibility of AI-enabled ODR must be guaranteed since lack of knowledge about the technology and limited resources may deepen the existing gap in dispute resolution.[40]

Language differences may cause problems in terms of misunderstandings and communication in legal language. There is always a risk of faulty output of AI; therefore, AI input should be considered a probability, not certainty. It is important to have a regulated AI-ODR system which would provide legal protection.

A well-regulated AI-ODR system that would include legal guarantees is needed. This system needs to include automation and human involvement in order to guarantee that the principles of justice will be kept and that efficiency would add value to justice principles, but not replace them.

5 Current Legal Framework for ODR in India and Comparative Analysis

There are a number of bases for technology-mediated dispute resolution in India's legal system, even if it does not constitute an entire ODR code. The Information Technology Act, 2000, and the Consumer Protection Act, 2019, are among the legislation that have been found to support electronic transactions and consumer-recognition via the current legislative and procedural structure.[41] However, these scattered statutes fail to cover matters such as platform accreditation, accountability for algorithms, data protection, informed consent, or redress for technological malfunction.[42] The lack of dedicated legislation leaves gaps regarding the enforceability, confidentiality, neutrality criteria, and the nature of the interface between technological support and decision-making powers.[43] An integrated ODR code must retain procedural flexibility and set minimum standards for the entire spectrum of ODR service providers[44] subjected to digital proceedings. The Mediation Act, 2023, is also a relevant piece of legislation regarding mediation institutions.[45]

These measures permit technological assistance, but the legal status of automated recommendations, the allocation of responsibility between platforms and neutrals, and the reviewability of digitally generated outcomes remain insufficiently defined.[46] This fragmentation is particularly significant where ODR processes involve vulnerable consumers or self-represented parties, because procedural convenience cannot substitute for consent, impartiality, and an effective opportunity to challenge error.[47] Dedicated legislation should therefore connect recognition of electronic processes with enforceable standards for transparency,

---

[39] *Id.*

[40] Leah Wing, *Artificial Intelligence and Online Dispute Resolution Systems Design: Lack of/Access to Justice Magnified*, 4 INT'L J. ONLINE DISP. RESOL. 16–20 (2017), SSRN (Sept. 30, 2017), https://ssrn.com/abstract=3665409.

[41] S. G. Taddesse, V. Bhanwar & S. Verma, *Online Dispute Resolution in the Field of Consumer Protection in India*, 4 J. DIGIT. TECHS. & L. 73–97 (2026), https://doi.org/10.21202/jdtl.2026.4.

[42] *Id.*

[43] O. A. Shamov, *From Ethical Principles to Certified Practice: A Tiered Regulatory Model for Artificial Intelligence in Online Dispute Resolution*, 3 LEGAL POSITION 10–14 (2025), https://doi.org/10.32782/2521-6473.2025-3.2.

[44] *Id.*

[45] *Id.* at 10.

[46] Pitamber Yadav, *Digitizing Justice: The Case for Dedicated Online Dispute Resolution Legislation in India* (May 22, 2024), SSRN, https://ssrn.com/abstract=4840167, http://dx.doi.org/10.2139/ssrn.4840167.

[47] Leah Wing, *Artificial Intelligence and Online Dispute Resolution Systems Design: Lack of/Access to Justice Magnified*, 4 INT'L J. ONLINE DISP. RESOL. 16–20 (2017), SSRN (Sept. 30, 2017), https://ssrn.com/abstract=3665409.

accessibility, data governance, and human oversight.[48] A dedicated statute could also establish a tiered system of certification, requiring stricter scrutiny where AI performs functions that materially influence settlement or adjudicatory outcomes.[49] Such legislation should mandate accessible explanations, auditability, complaint mechanisms, and meaningful human review, while preserving party autonomy and ensuring that technological efficiency does not displace procedural justice.[50]

This would create a clearer basis for the integration of ODR with India's existing mediation and court-connected institutions.[51] A statutory framework should also distinguish facilitative tools from systems that truly determine outcomes, so that accountability is apportioned according to the level of automated influence.[52] It should require ODR providers to keep records sufficient for independent review, protect confidential communications and provide accessible alternatives where parties cannot participate digitally.[53] These safeguards would make technological innovation consistent with equality of arms and the continuing authority of human decision-makers.[54]

Specifically, the regulation of AI is beginning to impact legal policy, especially in the domain of Online Dispute Resolution (ODR). While ODR is appreciated for its ability to support negotiation and mediation, it also poses diverse problems of bias, responsibility and procedural justice when AI is implicated in ODR. At present, there is no global regulatory approach to AI-ODR. The European Union has adopted a risk-based legislative approach by enacting the EU Artificial Intelligence Act while the UK adopts the principles-based approach and the US adopts the risk-management approach. India falls somewhere in between these approaches as it does not have an AI-ODR statute but provides a legal ground for it through policies and statutes such as the Mediation Act 2023 and the Digital Personal Data Protection Act 2023 which regulates the consent for and processing of personal data used in AI-based ODR mechanisms. Apart from these, Arbitration and Conciliation Act, 1996 allows online arbitration and enforcement of electronic awards through its technologically neutral language, when read with IT Act, 2000[55]; Information Technology Act, 2000 recognizes the credibility of electronic documents and signatures as digital evidence[56] and Bharatiya Sakshya Adhiniyam, 2023 also recognizes digital evidence[57]; Consumer Protection Act, 2019 stipulates electronic filing and virtual proceedings for customer complaints.[58]

India however has not one but several legal and policy instruments along with the ODR policy plan drafted by the NITI Aayog that talks about the technological capabilities of ODR but lacks a regulatory regime regarding the hazards caused by AI and therefore its regulation framework is permissive but inadequate.

---

[48] O. A. Shamov, *From Ethical Principles to Certified Practice: A Tiered Regulatory Model for Artificial Intelligence in Online Dispute Resolution*, 3 LEGAL POSITION 10–14 (2025), https://doi.org/10.32782/2521-6473.2025-3.2.

[49] *Id.*

[50] *Id.*

[51] Y. Choung, *Application of AI in Online Dispute Resolution in India*, 23 L. REV. 57–75 (2023), https://doi.org/10.57057/lawreview.2023.09.23.3.57.

[52] J. Zeleznikow, *Using Artificial Intelligence to Provide Intelligent Dispute Resolution Support*, 30 GROUP DECISION & NEGOT. 789–812 (2021), https://doi.org/10.1007/s10726-021-09734-1.

[53] *Id.* at 17.

[54] J. Zeleznikow, *Can Artificial Intelligence and Online Dispute Resolution Enhance Efficiency and Effectiveness in Courts*, 8 INT'L J. FOR CT. ADMIN. 30 (2017), https://doi.org/10.18352/ijca.223.

[55] Arbitration and Conciliation Act, 1996, § 36.

[56] Information Technology Act, 2000, §§ 3, 4, 5.

[57] Bharatiya Sakshya Adhiniyam, 2023, §§ 62, 63.

[58] Consumer Protection Act, 2019, § 35.

Conversely, the approach taken by the EU[59] defines high-risk artificial intelligence technologies with stringent requirements in terms of transparency, risk management and human interaction.[60] Accessibility and procedural fairness are major considerations in the Online Procedure Rules of the UK, where it is stated that the application of technology must not hinder justice.[61] In the US, it is regulation and standard-setting that matter along with continuous risk management.[62]

Some important lessons for India from these comparative examples would include creating a bespoke AI-ODR approach that will include risk-based approach of the EU, consideration of accessibility by the UK, and continuous risk management of the US. Principles of India's regulatory framework could be the following: classification of AI technologies according to their legal effects, human involvement, transparency, and digital inclusion.

This core dilemma revolves around deciding the amount of control which is left in the hands of AI. Maintaining an optimal balance between innovation and essential measures could assist in building a successful AI-powered ODR system in India, which would not ignore important legal aspects of law. In the end, a hybrid strategy that incorporates the learning experiences from the EU, UK and US will be most suitable for India.

5.1 Current ODR Systems Operative in India

SAMA Platform employs AI technology to evaluate disputes and allocate cases to experts and achieves an average resolution time of below 45 days for numerous disputes related to financial matters and employment. The platform was able to solve 10,000 cases in six months. Presolv360 solves disputes before litigation through NLP interpretation of contracts and AI prediction of the process of negotiation. In E-Lok Adalats of Gujarat, Maharashtra, and Rajasthan, mass disputes related to traffic fines and small debts are sorted out with AI scheduling and biometric authentication techniques.[63] Apart from these, the MSME portal also has its own ODR system.[64]

6 FUTURE OF AI-ENABLED JUSTICE

The future of justice in AI is the development of judicial processes due to the implementation of digital and intelligent systems. AI is used in legal research, case management, and online dispute resolution (ODR) and experiments with various functions. However, it is necessary to organize the implementation of AI to retain fundamental values such as fairness, transparency, and human control over decisions.[65]

---

[59] Regulation (EU) 2024/1689, https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng/.

[60] C. Novelli, F. Casolari, P. Hacker, G. Spedicato & L. Floridi, *Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity*, 55 COMPUT. L. & SEC. REV. 106066 (2024), https://doi.org/10.1016/j.clsr.2024.106066.

[61] Online Procedure Rule Committee, *Online Procedure Rules 2026*, GOV.UK (Aug. 26, 2026), https://www.gov.uk/government/publications/online-procedure-rules-2026.

[62] E. TABASSI, ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (NIST Trustworthy and Responsible AI, 2023), https://doi.org/10.6028/NIST.AI.100-1, https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225 (last visited Sept. 11, 2026).

[63] Record of Law, *Integration of Artificial Intelligence in Online Dispute Resolution (ODR)*, RECORD OF LAW (Feb. 21, 2026), https://recordoflaw.in/integration-of-artificial-intelligence-in-online-dispute-resolution-odr/.

[64] Ministry of Micro, Small and Medium Enterprises, Government of India, *MSME Online Dispute Resolution (ODR) Portal*, https://odr.msme.gov.in/.

[65] S. Ravishankar, *The Role of Artificial Intelligence in Arbitration: Prospects for India*, 7 INT'L J. FOR MULTIDISCIPLINARY RSCH. (2025), https://doi.org/10.36948/ijfmr.2025.v07i02.41005.

Thus, the process of the evolution from digital to intelligent justice shows the potential of using AI in automated processes including document summarization and procedural support, which results in better accessibility of justice.[66] The human-in-the-loop model is vital to make sure that the decision on legal questions still lies with humans and is accountable. It is possible to use the analysis of previous trends and predict decisions thanks to the use of predictive analytics; however, there is a danger of bias reinforcement.

In the future, ODR platforms will enable automated resolution of disputes through human oversight. Justice personalization will increase users' satisfaction, but the use of AI should take into account the risk of discrimination.[67] And it is crucial to make sure that everybody has access to the digital world and there is no digital divide in particular in those countries that are still developing.

Furthermore, the application of AI in the legal process should be transparent for both parties in order to understand what kind of role AI plays in that process.[68] It is essential to develop the accountability framework when artificial intelligence technologies fail, and the responsibilities of each party should be determined. Lawyers should get new skills in AI technologies and ethics.

Finally, the future of justice should become such a paradigm in which AI is used as the supplementary technology but cannot replace the competences of humans.[69] The main principle is that AI can change the nature of justice but should not affect the basic principles of justice and create a human-centered environment.

7 RECOMMENDATIONS AND SUGGESTIONS

The juncture of artificial intelligence (AI) and ODR will require a careful regulation to stimulate innovations in order to secure justice. The recommendations include:

1. Developing a special regulatory regime for AI in ODR.
2. Risk assessment approach to categorize the use of AI technology in terms of its impact on human rights.
3. Inclusion of human participation in critical cases.
4. Guarantee of transparency and clarity about the use of AI.
5. Protection of data privacy.
6. Performing independent evaluations of AI technologies with higher risks.
7. Protecting vulnerable users against digitalization through ODR.
8. Encouraging AI literacy among lawyers.
9. Establishing a review process for AI decisions.
10. Stimulating responsible innovations that comply with the constitution.

---

[66] *AI in International ODR: Reshaping Global Justice*, NEXLAW BLOG (May 13, 2026), https://www.nexlaw.ai/blog/global-odr-ai-disputes/.

[67] H. Alessa, *The Role of Artificial Intelligence in Online Dispute Resolution: A Brief and Critical Overview*, 31 INFO. & COMMC'NS TECH. L. 319–42 (2022), https://doi.org/10.1080/13600834.2022.2088060.

[68] *Id.*

[69] Michael Araszkiewicz et al., *The Role of New Information Technologies in Alternative Dispute Resolution of Divorce Disputes*, 1 EUR. DISP. RESOL. J. 549, 552 (2014).

8 Conclusion

There is great potential for Artificial Intelligence (AI) in facilitating the delivery of justice as well as improving access to justice through ODR services. Such improvements could be seen in terms of reduced delay, reduced costs and increased efficiency of process, especially in case of low value disputes and those involving cross-border claims.[70] But there are also some difficulties that arise from such an approach – namely algorithmic bias, transparency questions, privacy concerns, digital divide and many others. Jurisdictions such as the EU, the UK and the US face similar issues and have started dealing with this problem, offering useful experience for India. India is starting to recognize the need for online mediation and ODR policy making, yet additional safeguards are needed with regard to transparency, oversight and accountability. The main question here is the level of authority that is granted to AI, which should not become the sole decision-maker. The future of AI-assisted justice is therefore one in which the future can be said to lie in human-centered and enhanced justice. In this kind of situation, the AI will undertake all the computational, analysis and administration functions, while the judges, mediators, arbitrators and other professionals will remain responsible for interpreting, exercising discretion and making the final decisions.

* * *

[70] Dr. S. S. Khatana & Dr. S. Choudhary, *Role of Artificial Intelligence in Enhancing Access to Justice in India - A Critical Analysis*, 23 Lex Localis - J. Local Self-Gov't 1770–78 (2025), https://doi.org/10.52152/qytz3640.

PART IV

Privacy, Data Protection and Speech Online

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.17

CHAPTER 17

The Data Protection Regime in India: Constitutional Privacy to Statutory Compliances

ENAKSHI GUHA*

In India, the data protection landscape has evolved from a constitutional principle into a comprehensive statutory and regulatory framework anchored in the recent Digital Personal Data Protection Act, 2023 (DPDP Act) and DPDP Rules, 2025, and parallel cybersecurity mandates under the Information Technology Act, 2000 and CERT-In Directions, 2022. Relying on the jurisprudential precedents of the data protection architecture, the recognition of right to privacy as a fundamental right intrinsic to Article 21 of the Indian Constitution, is analysed in the paper. The analysis details the four core rights conferred on data principals: the right to access, right to correction and erasure, right to grievance redressal, and right to nominate (Section 14), alongside the corresponding obligations of data fiduciaries, including notice, consent, legitimate uses, accuracy, security safeguards, personal data breach intimation (with a detailed report to the Board within 72 hours under rule 7 of the DPDP Rules, 2025), retention limits, and grievance redressal mechanisms provided in the Act.

The paper further examines cross-border data transfer provisions, noting that transfers are permitted except to countries or territories that the Central Government restricts by notification, and that no such notification has been issued as of July 2026. Until notified, organizations must rely on reasonable security safeguards and contractual protections, while monitoring for potential restrictions on sensitive categories (health, financial, biometric data). This paper provides a rigorous doctrinal and practical analysis of the emerging data protection, privacy, synthesising constitutional foundations, statutory obligations, enforcement mechanisms, and operational challenges facing organizations as the DPDP framework enters phased operational force. The paper further examines the core architecture and the ground level application of the Act, including definitions of Data Principal, Data Fiduciary, Data Processor, and the yet-to-be-notified category of Significant Data Fiduciary (SDF), which will trigger enhanced obligations such as data protection officer appointments, audits, and impact assessments.

Finally, the paper identifies five critical compliance challenges in practice: (1) consent fatigue from repeated user requests; (2) legacy data collected for broad or undefined purposes; (3) AI training on personal data without explicit statutory guidance; (4) SME and startup burden in meeting SDF-level obligations; and (5) CERT-In vs. DPDP reporting tension. Mitigation strategies, including layered notices, data mapping, shared compliance services, and integrated incident response frameworks, are proposed.

Keywords: Digital Personal Data Protection Act 2023, data privacy and CERT-In Directions, data fiduciary obligations, cross-border data transfers, proportionality test and constitutional foundations.

1 INTRODUCTION TO THE TOPIC

In India, the trajectory of data protection law represents a paradigmatic shift from constitutional recognition to statutory operationalisation. The Hon'ble Supreme Court of India in its landmark judgment in Justice K.S. Puttaswamy (Retd.) vs. Union of India,[1] established privacy as a fundamental right under Article 21, Constitution of India, laying the doctrinal foundation for comprehensive data protection

* Practicing Advocate.

[1] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

legislation. Years later, the Digital Personal Data Protection Act, 2023 (DPDP Act) emerged as the nation's first dedicated data protection statute, received Presidential assent on 11th August, 2023, and entered phased commencement with the notification of DPDP Rules, 2025 on 13th November, 2025.[2]

The mentioned landmark judgment articulated a three-fold proportionality test for evaluating privacy intrusions: 1) legality, requiring valid statutory authorisation, 2) legitimate state aim of sovereignty, security and public order, and 3) proportionality, ensuring a rational nexus between the objects and the means adopted.[3] This constitutional framework has governed subsequent privacy-related litigations, including Aadhaar challenges, surveillance disputes and data localisation debates. However, until the enactment of the Act, India lacked a comprehensive statutory regime governing personal data processing by private and public entities, relying instead on the Information Technology Act, 2000 and the SPDI Rules, 2011, which proved inadequate for addressing contemporary data processing practices. The DPDP Act operationalises this doctrine pronounced by the judgment by establishing rights-based obligations for data fiduciaries, enforcement mechanisms through the Data Protection Board of India, and penalties up to Rs. 250 crore for security failures. Later, the phased implementation reflects a calibrated approach to compliance, allowing organizations time to align data processing practices with statutory requirements.[4]

The paper pursues three primary objectives: 1) to examine the constitutional foundations of data protection in India, 2) to analyse how the core provisions of the DPDP Act – notice, consent, data principal rights, fiduciary obligations, breach notifications, and security safeguards – operate in practice for Indian organizations across sectors, and 3) to suggest a practical compliance framework integrating DPDP Act obligations with parallel cybersecurity mandates under CERT-In Directions addressing overlaps, tensions, and operational challenges. The scope of the paper is limited to the ground-level application of the Act within India focusing on data privacy and data protection obligations for domestic data fiduciaries, the judicial precedents interpreting privacy rights post-Puttaswamy judgment, existing research on compliance challenges, enforcement mechanisms, and sector-specific impacts, and the examination of cross-border data transfer restrictions (pending government notifications); it does not extend to technical cybersecurity standards beyond CERT-In's baseline requirements.

2 PATHWAY OF CONSTITUTIONAL FOUNDATIONS

The constitutional bedrock of the nation's data protection regime is the unanimous judgment pronounced by the Hon'ble Supreme Court of India in *Justice K. S. Puttaswamy (Retd.) vs. Union of India*.[5] This decision recognized the right to privacy as a fundamental right intrinsic to Article 21 and protected under Article 14 and Article 19 of the Constitution.

The Privacy holding: The court rejected the earlier opinions given in *M.P. Sharma vs. Satish Chandra*[6] and *Kharak Singh vs. State of Uttar Pradesh*[7] that privacy was not a standalone fundamental right, and further held that privacy is an inalienable aspect of human dignity and a precondition for exercising other fundamental freedoms. The judgment categorically identified multiple dimensions of privacy, including spatial

[2] Digital Personal Data Protection Act, 2023.
[3] *Puttaswamy Case Analysis: Right to Privacy*, THE LAW ACADEMY (Aug. 20, 2026), https://thelawcademy.in/law_resources/puttaswamy-case-analysis-right-to-privacy/.
[4] *The DPDP Act 2023 & DPDP Rules 2025, Explained*, VRATEX (May 28, 2026), https://www.vratex.com/dpdp-act.
[5] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[6] M.P. Sharma v. Satish Chandra, AIR 1954 SC 300.
[7] Kharak Singh v. State of Uttar Pradesh, AIR 1963 SC 1295.

privacy (protection of physical spaces), decisional privacy (autonomy over personal choices), and informational privacy (control over collection, use and disclosure of personal data). This informational privacy directly underpins the consent-based frameworks, data principal rights, privacy and fiduciary obligations as provided under the DPDP Act.

The three-fold proportionality test: The court established that any state action interfering with privacy must satisfy a three-fold proportionality test, which has become the governing standard for evaluating all privacy-related legislation and executive actions.

TABLE 1. *The three-fold proportionality test in Puttaswamy*

LIMB	REQUIREMENT	JUDICIAL STANDARD
Legality	Valid statutory authorisation	Executive action alone is insufficient; a clear, precise law enacted by Parliament or State Legislature must exist. The law must not be vague or overbroad.
Legitimate Aim	Proper state objective	The restriction must pursue a legitimate state interest: national security, crime prevention, welfare delivery, public order, or protection of other fundamental rights.
Proportionality	Rational nexus	There must be a rational nexus between the objects and the means adopted to achieve them; the Aadhaar judgment, (2019) 1 SCC 1, later added that the measure must be the least restrictive available and balanced.

In the case, Justice D.Y. Chandrachud held that an invasion of privacy must meet the three-fold requirement of legality, need (a legitimate state aim) and proportionality, which ensures a rational nexus between the objects and the means adopted to achieve them.[8]

Application to data protection: The court had observed that "Privacy includes at its core the preservation of personal intimacies, the sanctity of family life, marriage, procreation, the home and sexual orientation. Privacy also connotes a right to be left alone." In its interpretation, the famous Puttaswamy judgment explicitly recognized that informational privacy requires safeguards against unauthorized data collection, processing, and disclosure. Thus, the holding of the case directly structures the Act as follows:[9]

1. Consent requirements under section 6, DPDP Act: Data principals must exercise control over data processing through free, specific, informed, and unambiguous consent.
2. Purpose limitation under section 5, DPDP Act: Data fiduciaries must provide clear notice specifying the purposes for which data will be processed.
3. Retention limits under section 8(7), DPDP Act: Data must be deleted when the processing purpose is exhausted, preventing indefinite retention.
4. Security safeguards under section 8(5), DPDP Act: Fiduciaries must implement reasonable technical and organizational measures to protect data from breaches.

Post-Puttaswamy judgment: Subsequent to the Puttaswamy judgment, many other judgments have refined and applied the proportionality test to diverse contexts.[10]

[8] *Justice K.S. Puttaswamy (Retd.) v. Union of India (2017)*, RECORD OF LAW (Aug. 20, 2026), https://recordoflaw.in/justice-k-s-puttaswamy-retd-v-union-of-india-2017-6/.

[9] *Complete DPDP Act 2023 Guide — Citizen and Business Reference 2026*, RIGHT TO INFORMATION WIKI (Aug. 22, 2026), https://rightoinformation.wiki/dpdp-act-2023-complete-guide.

[10] *Justice K.S. Puttaswamy v. Union of India*, LIBERTY LEXICON (Apr. 25, 2026), https://www.libertylexicon.in/blog/case-commentary-10/justice-k-s-puttaswamy-v-union-of-india-16.

1. Privacy challenges in regard to Aadhaar in 2018-2023: In the Aadhaar judgment,[11] the constitution bench upheld the constitutionality of Aadhaar for welfare delivery and its mandatory linking with PAN for income tax filing, but struck down mandatory linking with bank accounts as a disproportionate intrusion and held that Aadhaar could not be made mandatory for school admissions, education being an entitlement and not a subsidy, benefit or service. The court emphasized data minimisation, i.e., only necessary data shall be collected for specific purposes.
2. Surveillance and privacy: Many high courts have applied the same foundations to scrutinise mass surveillance programmes. In Vinit Kumar vs. Central Bureau of Investigation,[12] the Bombay High Court held that orders for the interception of telephone calls must observe the statutory safeguards and satisfy the proportionality test, and set aside orders that did not.
3. Compelled disclosure and biometric data: In *Ritesh Sinha vs. State of Uttar Pradesh*,[13] the Hon'ble Supreme Court held that, until Parliament legislates on the point, a Judicial Magistrate may order a person to give a sample of his voice for the purpose of investigation, the right to privacy not being absolute and yielding to a compelling public interest.

Continuing legacy: The foundation established by the Puttaswamy judgment has transformed the constitutional law by elevating privacy from a common law right to a fundamental right enforceable against the state and, indirectly, the private entities through statutory frameworks like the DPDP Act; establishing proportionality as the analytical framework for evaluating privacy intrusions, influencing not only data protection but also surveillance laws, criminal procedure, and administrative law; and informing legislative design as the exemptions under the Act for government instrumentalities under section 17 and legitimate uses under section 7 are explicitly framed around the legitimate categories aimed at sovereignty, security and public order. However, the broad exemptions under section 17 risk undermining the proportionality requirement by allowing the executive to exempt itself from compliance without robust parliamentary oversight or judicial review. This stress between constitutional doctrine and statutory implementation remains a critical area for future judicial scrutiny.

3 STRUCTURIZATION OF THE DPDP ACT, 2023

The DPDP Act, 2023 represents one of the nation's first comprehensive statutory frameworks governing the processing of personal data by both public and private entities. The Act received Presidential assent on 11 August 2023 and was published in the Gazette of India, marking the culmination of a decade-long legislative journey that began with the Justice A.P. Shah Committee Report (2012) and included multiple draft bills (2018, 2019, 2022).

Phase Commencement: Following the parliamentary passage in August 2023, the Act entered a phased commencement schedule to allow organizations time to align their data processing practices with statutory requirements. The DPDP Rules, 2025[14] were notified on 13 November 2025, operationalising delegated provisions under section 40 of the Act. The MeitY Secretary has confirmed no extension to these

[11] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2019) 1 SCC 1.
[12] Vinit Kumar v. Cent. Bureau of Investigation, 2019 SCC OnLine Bom 3155.
[13] Ritesh Sinha v. State of Uttar Pradesh, (2019) 8 SCC 1.
[14] Ministry of Electronics and Information Technology, Digital Personal Data Protection Rules, 2025, Gazette of India (Nov. 13, 2025).

deadlines, urging startups and enterprises to begin compliance preparations immediately. As of August 2026, organizations face a firm compliance deadline of May 2027 for core obligations.[15]

TABLE 2. *Phased commencement of the DPDP Act, 2023*

PHASE	COMMENCEMENT DATE	PROVISIONS IN FORCE
Phase 1	13 November 2025	Definitions (Section 2), Data Protection Board (Sections 18–26), rule-making powers (Section 40), Section 44(3) RTI amendment.
Phase 2	13 November 2026	Consent Manager registration (Sections 6(9) and 27(1)(d)).
Phase 3	13 May 2027	Full compliance with consent, notice, security and data rights obligations (Sections 3–17, except Section 6(9)), the Board's inquiry, appeal and penalty provisions (Sections 27–34, except Section 27(1)(d)), and Sections 36, 37 and 44(2).

Key concepts under the Act: The Act's operational framework rests on four foundational definitions provided under section 2.[16]

1. Data principal under section 2(j), DPDP Act: The individual to whom personal data relates, including the parents or lawful guardian of a child and the lawful guardian of a person with disability acting on her behalf.
2. Data fiduciary under section 2(i), DPDP Act: Any person (individual, company, organization, or government body) who, alone or in conjunction with others, determines the purpose and means of processing personal data. This definition mirrors the concept of 'data controller' under the General Data Protection Regulation, which is a comprehensive data privacy law passed by the European Union that sets strict rules for collecting and processing personal data. It imposes primary compliance responsibility on fiduciaries, irrespective of any agreement to the contrary.
3. Data processor under section 2(k), DPDP Act: A person who processes personal data on behalf of a data fiduciary under valid contractual arrangement. Processors are not directly liable under the Act, but fiduciaries remain responsible for ensuring processor compliance through contractual safeguards.
4. Significant data fiduciary (SDF) under section 2(z), DPDP Act: A category yet to be notified by the Central Government, expected to include entities that process large volumes of sensitive personal data or pose higher risks to data principals. Once it is notified, SDFs will face additional obligations of appointing a Data Protection Officer, based in India, appointing an independent data auditor, conducting periodic data protection impact assessments and additional compliance requirements as prescribed by DPDP Rules, 2025.

[15] Centre for Applied Data Protection, *DPDP Act Implementation Status 2026, Complete Tracker*, CADP (Mar. 9, 2026), https://cadp.in/resources/guides/dpdp-implementation-tracker/.
[16] Centre for Applied Data Protection, *DPDP Act Key Provisions Explained: A Detailed Analysis*, CADP (Jan. 30, 2026), https://cadp.in/resources/articles/dpdp-act-key-provisions-explained/.

Territorial applicability: Section 3, DPDP Act establishes an extraterritorial reach so that processing outside India is covered where it is in connection with offering goods or services to data principals in India, as with foreign platforms such as Meta, Google and Amazon serving Indian users.

Personal data over sensitive private data: The Act defines personal data broadly as "any data about an individual who is identifiable by or in relation to such data" as provided under section 2(t). Unlike the General Data Protection Regulation, the DPDP Act does not create a separate category of 'sensitive private data' with enhanced protection. However, the SDF designation, once notified, may impose stricter obligations for processing certain data categories. Furthermore, the Act applies only to personal data collected in digital form or digitised afterwards (section 3(a)); section 3(c) excludes personal data processed by an individual for any personal or domestic purpose and personal data made publicly available by the data principal or under a legal obligation; and processing necessary for research, archiving or statistical purposes is exempted under section 17(2)(b), DPDP Act, where the personal data is not used to take any decision specific to a data principal and the processing follows the standards prescribed in the Second Schedule to the DPDP Rules, 2025.

The structurization of the Act reflects a rights-based approach balanced with enforcement mechanisms and flexibility for state interests. However, the definitional framework introduces several departures from international models like the following:

1. No explicit data minimisation principle: Unlike Article 5(1)(c) of the General Data Protection Regulation, the DPDP Act states no general data minimisation principle; section 6(1), however, expressly limits consent to such personal data as is necessary for the specified purpose.
2. Broad government exemptions: Section 17 of the Act allows the Central Government to exempt any instrumentality of the State from its provisions, raising serious proportionality concerns.

4 RIGHTS AND OBLIGATIONS UNDER THE ACT

The Act establishes a rights-based framework that confers enforceable entitlements on data principals while imposing corresponding obligations on data fiduciaries. The core provisions governing data principal rights under sections 11–14 and fiduciary obligations under sections 5–8, alongside government exemptions under section 17 and legitimate uses under section 7 are discussed below.[17]

Data principal rights: The Act confers several rights on data principals, enforceable against all data fiduciaries processing their personal data. The primary four rights are as follows:

1. Right to access under section 11, DPDP Act: Upon request, data principals may obtain a summary of personal data being processed, processing activities undertaken such as collection methods, usage patterns, disclosures, etc., and identities of all data fiduciaries and processors with whom data was shared, along with descriptions of data shared, from the data fiduciary. Organizations must maintain their data processing inventories and implement mechanisms to respond to access requests within a reasonable timeframe. Failure to comply may attract penalties up to Rs. 50 crores under section 33 read with the Schedule. However, legacy systems and fragmented data storage across departments may complicate comprehensive access responses, particularly for large organizations with decentralised data infrastructures.

[17] *Id.*

2. Right to correction and erasure under section 12, DPDP Act: Data principals may demand correction of inaccurate or misleading data, completion of incomplete data, updating of outdated data and erasure of personal data when the processing purpose is exhausted, which is also subject to statutory retention obligations under other general laws of the nation. Fiduciaries must establish data quality assurance processes and verify the accuracy of data before using it for decision-making or disclosing it to any third parties. For erasure requests, organizations must balance DPDP obligations against retention requirements under sectoral laws. However, the Act does not specify any timeline for correction or erasure compliance, creating uncertainty for organizations managing high-volume requests.
3. Right to grievance redressal under section 13, DPDP Act: Every data fiduciary must publish their contact details for grievance communication, respond to grievances within a period not exceeding 90 days (rule 14(3), DPDP Rules, 2025); a data principal who remains dissatisfied may then complain to the Data Protection Board, having first exhausted this remedy (section 13(3)). Organizations must designate a grievance officer and implement ticketing systems to track response timelines. Non-compliance may attract penalties up to Rs. 50 crores.
4. Right to nominate under section 14, DPDP Act: Data principals may nominate another individual to exercise their rights in the event of death or incapacity. This provision addressed the post-mortem privacy gap in the existing law, allowing a nominee to exercise the deceased or incapacitated data principal's rights under the Act, such as access, correction and erasure. Platforms must implement nomination mechanisms and verify nominee credentials before granting access.

Data fiduciary obligations: Data fiduciaries bear primary responsibility for compliance, irrespective of any agreement to the contrary or failure of the data principal to carry out duties. The primary obligations are as follows:[18]

1. Notice under section 5, DPDP Act: Before collecting personal data, fiduciaries must provide a clear, concise, accessible notice containing an itemised description of the personal data, the specified purpose of processing with a description of the goods, services or uses it enables, the means by which the data principal may withdraw consent and exercise her rights, and the manner in which she may complain to the Board (section 5(1) and rule 3). Privacy policies must be layered, available in multiple languages, and presented before consent is obtained. Pop-up notices, just-in-time prompts, and granular preference centres are recommended for digital platforms. Organizations with legacy data collected before the commencement of the Act must give each data principal a notice under section 5(2) as soon as it is reasonably practicable, and may continue processing until she withdraws her consent; fresh consent is not required.
2. Consent under section 6, DPDP Act: Consent must be free, specific, informed, unambiguous, and capable of withdrawal. Consent management platforms (CMPs) must enable granular consent (for example, separate toggles for marketing, analytics, third-party sharing) and one-click withdrawal. Pre-ticked boxes, dark patterns, and consent fatigue (excessive pop-ups) are non-compliant. Consent fatigue, user exhaustion from repeated consent requests, may lead to 'click-through' behaviour, undermining the informed consent requirement. Mitigation strategies include layered notices, contextual consent (just-in-time prompts), and consolidated preference dashboards.

[18] *Compliance with DPDP Act*, ECON. TIMES (Aug. 10, 2026), https://economictimes.indiatimes.com/topic/compliance-with-dpdp-act.

3. Legitimate uses without consent under section 7, DPDP Act: Processing is permitted without consent for narrow, specified purposes like the specified purpose for which the data principal has voluntarily provided her data; State subsidies, benefits, services, certificates, licences and permits; State functions under law and the sovereignty, integrity and security of India; disasters and breakdown of public order; compliance with orders, judgments, statutory obligations, life-threatening situations where consent cannot be obtained, and processing necessary for employment-related activities. However, the broad phrasing of 'state functions' and 'public order' raises concerns about potential overreach, particularly given the section 17 exemptions for government instrumentalities.
4. Accuracy and completeness under section 8(3), DPDP Act: Where personal data is likely to be used to make a decision that affects the data principal, or to be disclosed to another data fiduciary, the fiduciary must ensure its completeness, accuracy and consistency. Organizations must implement data validation checks at collection points and periodic audits to identify and correct inaccuracies. Automated decision-making systems (AI/ML models) must be trained on accurate, representative datasets to avoid discriminatory outcomes.
5. Security safeguards under section 8(5), DPDP Act: Fiduciaries must implement reasonable security safeguards (technical and organizational measures) to prevent unauthorised access, acquisition, disclosure, use, loss, destruction, or damage. The term 'reasonable' is not defined in the Act, but rule 6 of the DPDP Rules, 2025 prescribes the minimum safeguards: encryption, obfuscation, masking or tokenisation; access controls; logging and monitoring; data backups; one-year retention of logs; and security clauses in processor contracts. Minimum safeguards may include encryption (in transit and at rest), access controls (role-based permissions, multi-factor authentication), regular vulnerability assessments and penetration testing, and incident response plans and breach notification procedures.
6. Breach notification under section 8(6), DPDP Act: Upon discovering a personal data breach, fiduciaries must intimate the breach, without delay, to each affected data principal and to the Data Protection Board, and must give the Board a detailed report within 72 hours of becoming aware of it. Organizations must establish incident response playbooks integrating 6-hour reporting obligation under CERT-In with the 72-hour detailed report to the Board under rule 7(2)(b) of the DPDP Rules, 2025. The intimation must describe the breach (its nature, extent, timing and location), its likely impact and consequences, the mitigation measures taken, and the safety steps data principals can take. The 72-hour deadline may be insufficient for organizations to identify all affected principals, particularly in complex breaches involving multiple systems.
7. Retention limitation under section 8(7), DPDP Act: Personal data must be deleted when the processing purpose is exhausted, unless retention is required by law. Organizations must implement data lifecycle management policies with automated deletion triggers in cases of account closure or statutory retention period expiry. This provision does not conflict with the 180-day log retention mandate under CERT-In, because section 8(7) does not apply where retention is necessary for compliance with any law; rule 6(1)(e) of the DPDP Rules, 2025 itself requires logs to be kept for one year.

Government exemptions under section 17, DPDP Act: The Central Government may exempt any instrumentality of the State from the provisions of the Act in the interest of sovereignty and integrity of India, security of the State, friendly relations with foreign States, maintenance of public order, and preventing

incitement to cognizable offence relating to the abovementioned. These exemptions extend to processing by the Central Government of personal data furnished by such instrumentalities. However, the broad, undefined terms risk enabling mass surveillance and data collection without adequate safeguards, potentially violating the three-fold proportionality test.

Legitimate uses by the State under section 7, DPDP Act: Processing is permitted without consent for state functions under law, including welfare delivery such as subsidies, pensions and healthcare, tax collection, law enforcement and crime prevention and national security and intelligence operations. However, these require paramount judicial oversight and the presence of independent authorization mechanisms aligning to proportionality concerns.

5 INSTITUTIONAL INFRASTRUCTURE AND ENFORCEABILITY

The Act establishes a centralised enforcement mechanism through the Data Protection Board of India, supported by a tiered penalty structure and appellate review.[19]

Data Protection Board of India (DPB): The DPB was established in Phase 1 as the primary adjudicating body for complaints, disputes, and penalty determination under section 18 of the Act. The Board shall consist of a Chairperson and such number of other Members as the Central Government may notify under section 19(1). The Board adjudicates complaints from data principals regarding violation of the above discussed rights, breach notification failures by fiduciaries, security safeguard deficiencies, consent and notice violations and references from the Central Government or a State Government or on the direction of a court. For proper adjudication, the Board is vested with the power to summon parties and examine witnesses under oath, impose penalties up to Rs. 250 crores, direct cessation of processing activities, and issue binding directions (section 27(2)) and interim orders (section 28(10)).

However, as of July 2026, no appointment of a chairperson or members had been notified, and the Board was not yet operational. This raises serious concerns regarding enforcement capacity and adjudicatory timelines once complaints begin flowing starting Phase 3 in May 2027.

Penalty structure under section 33 and the Schedule, DPDP Act: The Act prescribes tiered penalties considering which the DPB must determine the penalties. The Board must have regard to the nature, gravity and duration of the breach, the type and nature of the personal data affected, the repetitive nature of the breach, whether the person realised a gain or avoided a loss, any mitigating action and its timeliness and effectiveness, whether the penalty is proportionate and effective, and the likely impact of the penalty on the person (section 33(2)).

TABLE 3. *Maximum penalties under the DPDP Act, 2023*

Violation	**Maximum Penalty**
Failure to maintain reasonable security safeguards (Section 8(5))	₹250 crore
Failure to notify data breach to DPB/principals (Section 8(6))	₹200 crore
Breach of the additional obligations in relation to children (Section 9)	₹200 crore

[19] Centre for Applied Data Protection, *DPDP Act 2023 Full Text — Interactive Reference*, CADP (Mar. 6, 2026), https://cadp.in/resources/official-texts/dpdp-act-2023/.

Violation	Maximum Penalty
Breach of the additional obligations of a Significant Data Fiduciary (Section 10)	₹150 crore
Breach of any other provision, including notice and consent (Sections 5–6), data principal rights (Sections 11–14) and grievance redressal within 90 days (Section 13)	₹50 crore

The enforcement process includes a complaint by a data principal, an intimation of a personal data breach by the data fiduciary, a reference by the Central Government or a State Government, or a direction of a court (section 27(1)); the Act gives the Board no suo motu power. This is followed by inquiry and summary proceedings by the DPB. Thereafter, the DPB passes an adjudication order of penalty imposition and/or compliance directives. The order is appealable to the Appellate Tribunal, the Telecom Disputes Settlement and Appellate Tribunal, within sixty days from the date of receipt of the order under section 29 of the Act, and a further appeal lies to the Supreme Court. However, there still persist enforcement gaps as the DPB remains unstaffed which creates uncertainty about operational readiness. Delays in appointments may result in a compliance backlog and undermine deterrence.

6 CYBERSECURITY COMPLIANCES UNDER CERT-IN DIRECTIONS, 2022

Parallel to the DPDP Act, the Indian Computer Emergency Response Team (CERT-In) issued directions under section 70B of the Information Technology Act, 2000 on 28th April 2022, effective from 28th June 2022, imposing stringent cybersecurity obligations on organizations operating in India. Non-compliance with the CERT-In directions is punishable under section 70B (7) of the IT Act, 2000, with imprisonment up to 1 year, or fine up to Rs. 1 crore, or both.[20]

TABLE 4. *Principal obligations under the CERT-In Directions, 2022*

OBLIGATION	REQUIREMENT	APPLICABILITY
6-Hour Incident Reporting	Report specified cyber incidents to CERT-In within 6 hours of detection	Service providers, intermediaries, data centres, body corporates, government organizations.
180-Day Log Retention	Maintain logs of all ICT systems for a rolling 180-day period within India	All covered entities.
NTP Synchronisation	Synchronise clocks to NIC/NPL time sources (National Informatics Centre / National Physical Laboratory)	All covered entities.
5-Year Subscriber Data Retention	Data centres, VPS, cloud and VPN service providers must keep validated subscriber registration details for 5 years after the registration ends; virtual asset service providers, exchanges and custodian wallet providers must keep KYC records and financial transaction records for 5 years	Data centres, VPS, cloud and VPN providers; virtual asset service providers, exchanges and custodian wallet providers.

[20] Indian Computer Emergency Response Team, Directions under Section 70B(6) of the Information Technology Act, 2000, Gazette of India (Apr. 28, 2022).

Annexure I provides for twenty categories of reportable incidents which include malware, ransomware, phishing attacks, unauthorized access, data breaches, DDoS attacks, website defacement, cryptojacking, IoT compromises, critical infrastructure incidents and so on.

The reporting mechanism undertakes a contact-based system where reports must be filed with CERT-In (https://www.cert-in.org.in/ - contacts include their email: incident@cert-in.org.in).

7 Overlap and Tension: DPDP Act Versus CERT-In Directions

Organizations operating in India face parallel compliance obligations under DPDP Act and CERT-In Directions, creating overlaps, tensions, and operational complexities.[21]

Dual breach notification duties: A practical compliance strategy is required. This may be presented as:

1. 0–6 hours: Immediate triage, assess incident scope, report to CERT-In;
2. 6–72 hours: Investigate breach, identify affected principals, intimate the DPB and each affected data principal without delay, and prepare the detailed report to the DPB;
3. Within 72 hours: Submit the detailed breach report to the DPB.

However, the 6-hour CERT-In deadline may precede full breach investigation, requiring organizations to report with incomplete information and submit supplementary updates later.

Log retention conflict: Organizations must retain logs for 180 days minimum under CERT-In Directions, even if the DPDP purpose is exhausted earlier, unless a specific DPDP exemption applies (for example, retention necessary for compliance with law under section 8(7)). This creates a statutory override scenario where sectoral mandate under CERT-In prevails over general retention limit under DPDP Act. Practical mitigation may include implementation of segregated log storage with access controls, ensuring logs are retained for 180 days but not used for unrelated processing purposes (for example, profiling, marketing, etc.).

NTP Synchronisation and Data Accuracy: NTP synchronisation requirement under CERT-In (clocks must align with NIC/NPL time sources) supports accuracy obligation under section 8(3), DPDP Act by ensuring timestamps on data processing activities are reliable and auditable. This alignment reduces compliance friction.

Subscriber Data Retention (VPN/Cloud Providers): 5-year subscriber data retention mandate under CERT-In for VPN providers and cloud services conflicts with purpose-based retention limit under section 8(7), DPDP Act. However, the same VPN/cloud providers must keep validated subscriber registration details for 5 years under CERT-In, and section 8(7), DPDP Act, permits retention necessary for compliance with any law for the time being in force, exempting such processing from the general retention limit.

Suggested integrated Compliance Framework: Organizations, thus, should adopt an integrated cybersecurity and data protection compliance framework which may include a unified incident response playbook addressing both CERT-In (6-hour) and DPDP (72-hour) reporting, data mapping inventory identifying which data categories trigger CERT-In vs. DPDP obligations, retention schedules reconciling 180-day/5-year mandates under CERT-In with purpose-based limits under DPDP Act, and vendor contracts

[21] *CERT-In Directions & ISO 27001: The 6-Hour Mapping*, TCSA (July 11, 2026), https://www.tcsa.in/frameworks/iso-27001/for-cert-in.

ensuring processors comply with both regimes (for example, cloud providers keeping subscriber registration details for 5 years while implementing DPDP security safeguards).

8 CROSS-BORDER TRANSFER AND EMERGING CHALLENGES

Cross-Border data transfers: Under Section 16 of the DPDP Act, personal data may be transferred to any country or territory outside India except one that the Central Government restricts by notification; section 16(2) preserves any other law that provides a higher degree of protection or restriction, and rule 15 of the DPDP Rules, 2025 subjects transfers to such requirements as the Central Government may, by general or special order, specify for making personal data available to a foreign State or to any person or entity under the control of, or any agency of, such a State. As of July 2026, the Central Government has not notified any restricted country or territory. Until notifications are issued organizations may transfer data subject to reasonable security safeguards under section 8(5), DPDP Act and contractual protections (data processing agreements, standard contractual clauses), transfers should be purpose-specific and documented in privacy notices under section 5, DPDP Act, and sensitive data categories like health, financial, and biometric should be transferred with heightened caution until the Government's position on them is clear. Various industry expectations and speculations suggest the government will leave transfers to jurisdictions such as the European Union member states, the United Kingdom, Singapore and Japan unrestricted, while restricting transfers to countries with weak data protection regimes or mass surveillance laws. Multinational organizations are suggested to closely monitor MeitY notifications and prepare for potential data localisation requirements for critical sectors like healthcare, banking, telecom, and so on.

Emerging compliance challenges: Some of the legal impediments and implications concerning the compliances are discussed below:

1. Consent fatigue: Users face repeated consent requests from multiple platforms, leading to 'click-through' behaviour that undermines the informed consent requirement under section 6, DPDP Act. This may be mitigated by layered notices, just-in-time consent prompts, consolidated preference dashboards, and granular toggles like separate consent for marketing, analytics, third-party sharing.
2. Legacy data and purpose limitation: Organizations hold historical data collected for broad or undefined purposes, which may not meet the requirement that personal data be processed for a lawful purpose (section 4) stated in the notice as the specified purpose (sections 5 and 2(za)), DPDP Act. This may be mitigated by data mapping exercises, purpose re-justification documentation, deletion of obsolete records, re-consent campaigns for active users.
3. Artificial Intelligence training and automated processing: The DPDP Act does not explicitly address AI/ML model training on personal data. Processing for AI training may exceed original consent purposes, triggering fresh consent or legitimate use requirements. However, the risk includes automated decision-making like credit scoring, hiring, insurance underwriting, which must comply with section 8(3), DPDP Act, (completeness, accuracy and consistency), requiring organizations to audit AI models for bias and ensure explainability.
4. SME and startup burden: Smaller entities lack resources for data protection officers, audits, impact assessments (once SDF rules are notified), and legal compliance teams. This may be mitigated by shared compliance services (third-party DPOs), regulatory sandboxes, government guidance documents, industry associations providing template policies and contracts.

5. Reporting tensions between CERT-In Directions and DPDP Act: The 6-hour CERT-In deadline conflicts with the need to identify affected principals for DPDP intimation without delay and for the detailed report to the DPB within 72 hours. The best practice is to maintain incident response playbooks with parallel reporting tracks, pre-drafted notification templates, and escalation matrices ensuring both deadlines are met.

9 Conclusion

The Digital Personal Data Protection Act, 2023, read with the DPDP Rules, 2025 and CERT-In Directions, 2022, establishes a comprehensive, risk-based, and enforceable data protection regime in India. The Act operationalises the three-fold proportionality test, established in the famous Justice K.S. Puttaswamy (Retd.) vs. Union of India, interpreting constitutional right to privacy into a statutory obligation for data fiduciaries and enforceable entitlements for data principals.

However, given the present scenario of the implementation, organizations might face a complex compliance landscape characterized by overlapping mandates like that of DPDP breach notification versus CERT-In incident reporting, conflicting retention requirements like 180-day logs versus purpose-based deletion, and evolving enforcement capacity like the non-functional Data Protection Board. The phased implementation schedule provides a calibrated timeline for alignment, but the firm deadlines and substantial penalties, up to ₹250 crore, demand immediate action.

Critical challenges still persist: consent fatigue undermining informed consent, legacy data requiring purpose re-justification, AI training on personal data without explicit statutory guidance, and SME resource constraints in meeting compliance obligations. Furthermore, the broad government exemptions under Section 17 risk violating the proportionality requirement unless accompanied by robust parliamentary oversight and judicial review.

However, the era of accountability for data fiduciaries has begun. As the Data Protection Board becomes fully operational and penalties take effect, organizations must prioritise data mapping, notice revision, consent mechanisms, security audits, and integrated incident response frameworks to achieve compliance. The established three-fold proportionality test remains the constitutional touchstone for balancing individual autonomy with legitimate state and commercial interests, ensuring that India's data protection framework evolves in harmony with technological innovation and fundamental rights.

* * *

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) ·
Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.18

CHAPTER 18

From Privacy to Practice: Evaluating the Implementation of India's Digital Personal Data Protection Act

DR. MALIYA[*] AND SANTOSH KUMAR[**]

India's Digital Personal Data Protection Act 2023 (DPDP Act) is the country's first comprehensive cross-sectoral framework for digital personal data. Its significance lies in whether it can translate the constitutional right to privacy recognised in Justice K.S. Puttaswamy v. Union of India into effective institutional and organisational safeguards. This paper evaluates that translation at a critical transitional moment. The final DPDP Rules were notified and the Data Protection Board of India was legally established in November 2025, but most substantive duties and enforcement powers follow an eighteen-month commencement period. Using doctrinal and institutional analysis, the paper assesses the Act's consent architecture, rights, duties of data fiduciaries, children's-data regime, breach response, state exemptions, regulatory design and organisational implementation.

It argues that the Act provides a usable compliance framework but only partially embodies the constitutional conception of informational privacy. Its strengths include plain-language notice, purpose limitation, security safeguards, breach notification and potentially substantial penalties. Its weaknesses include broad state-processing grounds, limited rights, reliance on executive rulemaking, the absence of a general special-category framework and questions concerning the independence and capacity of the Data Protection Board. Effective implementation will depend on moving beyond formal consent toward demonstrable accountability: data inventories, purpose governance, privacy engineering, processor oversight, risk assessment, accessible rights channels and transparent enforcement. The paper proposes a phased implementation model and concludes that the success of the DPDP regime should be measured not by the number of consent notices or penalties, but by whether it reduces informational power asymmetries and makes lawful, fair and secure data practices routine across the Indian digital economy and the State.

Keywords: DPDP Act; privacy; data protection; India; Data Protection Board; compliance; digital governance

1 INTRODUCTION

India's digital transformation has created extensive public and private data infrastructure. Everyday life now depends on Aadhaar-enabled services, digital payments, e-commerce, telecommunications, health platforms, educational technology and data-driven welfare systems. This expansion has economic and administrative value, but it also concentrates informational power. Personal data can be reused, aggregated, inferred and stored in ways that individuals may neither understand nor control. Data breaches, manipulative consent interfaces, behavioural profiling and function creep show why confidentiality alone cannot protect privacy. Effective protection requires rules for the entire data lifecycle and institutions capable of enforcing them.

[*] Assistant Professor at School of Law, NIILM University, Kaithal, Haryana, India.
[**] Research Scholar at School of Law, NIILM University, Kaithal, Haryana, India.

The constitutional starting point is Justice K.S. Puttaswamy (Retd.) v Union of India, in which a nine-judge bench unanimously held that privacy is a fundamental right grounded in dignity, liberty and autonomy. The judgment did not enact a complete privacy code. It required intrusions on privacy to have a legal basis, pursue a legitimate aim, use proportionate means and include safeguards against abuse. It also recognised informational privacy as essential to individual freedom in a networked society and anticipated a data-protection framework that balances individual interests with legitimate uses of data.[1] The Digital Personal Data Protection Act 2023 (DPDP Act) is Parliament's principal response to that constitutional requirement.[2]

This paper asks whether the Act translates constitutional privacy into meaningful institutional and organisational protection. The question is practical. Enacting a statute does not by itself change how a retailer collects phone numbers, an employer monitors staff, a platform profiles children, a processor secures databases or a public authority combines welfare records. Implementation depends on regulatory capacity, organisational routines, accessible rights and remedies and credible enforcement incentives. The paper argues that the DPDP Act provides a credible minimum architecture but remains an incomplete constitutional translation. Its effectiveness will depend on whether the implementation phase creates demonstrable accountability or produces little more than notices and consent records.

2 METHOD AND EVALUATIVE FRAMEWORK

This study uses doctrinal and institutional analysis. It interprets the DPDP Act, the final Digital Personal Data Protection Rules 2025 (DPDP Rules), the commencement and institutional notifications, Puttaswamy and the Aadhaar judgment. It then considers how the resulting obligations would operate within data-processing organisations. The EU General Data Protection Regulation (GDPR) and the OECD privacy principles are used as comparators, not as models that India must reproduce, but as sources of established implementation tools such as independent supervision, privacy-management programmes and risk assessment.[3]

The analysis uses four criteria. Constitutional fidelity asks whether the regime reflects legality, necessity, proportionality, dignity and protection against arbitrary informational power. Institutional effectiveness concerns the independence, competence, accessibility and remedial capacity of the Data Protection Board. Organisational operationality asks whether the statutory duties can be translated into governance, engineering, contracting and incident-response practices. Distributive fairness considers whether safeguards work for children, people with limited literacy, workers, people who depend on essential services and those dealing with the State. Together, these criteria distinguish formal compliance from substantive privacy protection.

This is necessarily an early implementation study. On 13 November 2025, the Government notified the final Rules, commenced specified institutional provisions, legally established the Board and set a phased commencement schedule.[4] The principal provisions on lawful processing, consent, fiduciary

[1] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

[2] Digital Personal Data Protection Act, 2023.

[3] Regulation (EU) 2016/679 (General Data Protection Regulation), especially arts. 5, 12–22, 25, 32–35, 51–52; OECD, RECOMMENDATION OF THE COUNCIL CONCERNING GUIDELINES GOVERNING THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA (2013 revision).

[4] Digital Personal Data Protection Rules, 2025, G.S.R. 846(E) (Nov. 13, 2025), as corrected; Commencement Notification, G.S.R. 843(E) (Nov. 13, 2025); Notification Establishing the Data Protection Board of India, G.S.R. 844(E) (Nov. 13, 2025); Notification on the Size of the Data Protection Board of India, G.S.R. 845(E) (Nov. 13, 2025).

obligations, individual rights, penalties and Board procedure are due to commence after an eighteen-month transition, in May 2027. MeitY began recruiting the Chairperson and members in May 2026, but no appointment notification was found in the public record by the end of August 2026.[5] The analysis therefore addresses readiness, institutional design and foreseeable implementation problems rather than mature enforcement outcomes.

3 Constitutional Privacy and Statutory Data Protection

Puttaswamy treated privacy as more than secrecy. It protects decisional autonomy, bodily integrity and control over the disclosure of personal information. Its proportionality framework requires a valid law, a legitimate state purpose and proportionate means, supported by safeguards against abuse.[6] The Aadhaar litigation likewise shows that constitutional assessment must examine the architecture of a data system, including data minimisation, retention, authentication records, security and oversight. A formally lawful system may still become constitutionally excessive.[7]

The DPDP Act uses distinct terminology. The person to whom personal data relates is the 'Data Principal'; the person who determines the purpose and means of processing is the 'Data Fiduciary'; and an entity that processes data on a fiduciary's behalf is a 'Data Processor'. The Act applies to digital personal data processed in India and to processing outside India connected with offering goods or services to Data Principals in India. It covers data collected digitally and data collected offline that is later digitised. Personal or domestic processing falls outside the Act and publicly available data is excluded in specified circumstances.[8] This broad, technology-neutral scope is practical, but the exclusion for publicly available data may leave individuals unprotected when such data is aggregated or used for profiling.

Processing must serve a lawful purpose and rest on consent or one of the statutory 'certain legitimate uses'. Consent must be free, specific, informed, unconditional and unambiguous, expressed through clear affirmative action and limited to the personal data necessary for the specified purpose. Withdrawing consent must be as easy as giving it.[9] The final Rules require a standalone, comprehensible notice that identifies the data and purpose, explains the relevant goods or services and describes withdrawal, rights and grievance procedures.[10] These requirements can discipline opaque privacy policies. They do not, however, eliminate consent fatigue, take-it-or-leave-it services or manipulative interfaces. Constitutional autonomy requires more than clicking 'agree' when a person lacks bargaining power or understanding.

4 Rights, Duties and the Accountability Gap

The Act gives Data Principals rights to obtain information about processing, seek correction and erasure, pursue grievances and nominate another person to exercise rights after death or incapacity. These rights include access to a summary of the personal data being processed and the identities of other fiduciaries and processors with which the data has been shared. The Act also imposes duties on Data Principals,

[5] Ministry of Electronics and Information Technology, Appointment to the Post of Chairperson and Other Members in the Data Protection Board of India, F No 2(1)/2026-Pers.I (May 6, 2026).

[6] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1; *see also* Anuradha Bhasin v. Union of India, (2020) 3 SCC 637 and Internet & Mobile Ass'n of India v. Reserve Bank of India, (2020) 10 SCC 274 on proportionality.

[7] K.S. Puttaswamy (Aadhaar) v. Union of India, (2019) 1 SCC 1.

[8] Digital Personal Data Protection Act, 2023, §§ 2–3.

[9] *Id.* §§ 4, 6–7.

[10] Digital Personal Data Protection Rules, 2025, r. 3.

including duties not to impersonate another person, suppress specified material information or submit false or frivolous grievances. Breach of those duties may attract a penalty.[11]

The catalogue is narrower than that found in many comprehensive data-protection regimes. The Act creates no freestanding rights to data portability, objection to processing or restriction of processing and no general protection against decisions based solely on automated processing. It also creates no general category of sensitive personal data subject to enhanced safeguards.[12] Health, biometric, financial, sexual-orientation and political-affiliation data therefore receive the Act's general protection unless sectoral law or the duties of a Significant Data Fiduciary add further safeguards. A technology-neutral framework is simpler, but it can treat materially different risks alike. A leaked restaurant preference and a leaked genomic record are both personal data, but their consequences differ sharply.

A Data Fiduciary remains responsible for processing carried out on its behalf. Its core duties include ensuring accuracy when data is used for a decision or disclosed, applying reasonable security safeguards, notifying personal-data breaches, erasing data when consent is withdrawn or the purpose is exhausted, publishing contact information and providing an effective grievance mechanism.[13] The final Rules translate the security duty into more concrete controls, including contractual safeguards for processors, access controls, encryption or comparable protection, backups, monitoring and retention of relevant logs.[14] These minimum controls are useful, but 'reasonable' security must remain contextual. A checklist can establish a floor; the level of protection above that floor should reflect risk, scale, sensitivity and likely harm.

5 BREACH RESPONSE AND ORGANISATIONAL PREPAREDNESS

Breach notification will test how the regime performs under pressure. The final Rules require a fiduciary to notify each affected Data Principal promptly and in clear language, describing the breach, its likely consequences, mitigation and protective measures and a contact point. The fiduciary must also notify the Board promptly and provide a fuller report within seventy-two hours unless the Board permits more time.[15] This approach is broader than the GDPR's risk-based rule for notifying affected individuals.[16] Universal notice may improve transparency, but without guidance that distinguishes material communications from routine technical events, it may also produce over-notification and notification fatigue.

An organisation cannot meet this duty by drafting a template after an incident. It needs asset inventories, data classification, detection capabilities, escalation thresholds, decision logs, forensic-preservation procedures, communication workflows and rehearsed coordination among security, privacy, legal, communications and senior management teams. Processor contracts should specify notification times, access to evidence, cooperation duties and responsibility for response. Outsourcing infrastructure does not outsource accountability: the Data Fiduciary remains responsible for processing carried out on its behalf.[17]

These operational demands are especially difficult for micro, small and medium enterprises, start-ups, public bodies and organisations with legacy systems. Risk-based implementation support can include

[11] Digital Personal Data Protection Act, 2023, §§ 11–15 and Schedule.
[12] *Compare* Regulation (EU) 2016/679 (General Data Protection Regulation), arts. 9, 15–22, *with* the rights and obligations in the Digital Personal Data Protection Act, 2023.
[13] Digital Personal Data Protection Act, 2023, § 8.
[14] Digital Personal Data Protection Rules, 2025, r. 6.
[15] Digital Personal Data Protection Act, 2023, § 8(6); Digital Personal Data Protection Rules, 2025, r. 7.
[16] Regulation (EU) 2016/679 (General Data Protection Regulation), arts. 33–34.
[17] Digital Personal Data Protection Act, 2023, § 8(1)–(2).

model notices, standard contractual clauses, sectoral codes, breach exercises, shared security services and plain-language guidance from the Board. Enforcement should distinguish organisations that acted in good faith and maintained demonstrable governance from those that ignored known risks or shifted responsibility to others. Penalties matter, but predictable standards and technical assistance are also necessary to build capacity.

6 Children, Consent Managers and Significant Data Fiduciaries

Processing a child's data generally requires verifiable parental consent. The Act also restricts processing likely to harm a child's well-being, behavioural monitoring or tracking of children and targeted advertising directed at them, subject to prescribed exceptions.[18] Treating everyone under eighteen as a child is protective but operationally difficult. Age assurance may itself require intrusive data collection and parental control may conflict with adolescents' growing autonomy, particularly in health, counselling and educational services. Implementation should use proportionate age-assurance methods, collect only the minimum evidence, avoid persistent identity databases and account for differences in risk across services.

Consent Managers are intended to provide interoperable means to give, manage, review and withdraw consent. If trusted, they could reduce fragmented consent interfaces and make withdrawal effective. The Rules prescribe eligibility, financial, technical, governance and fiduciary requirements.[19] Yet an intermediary that manages consent may itself become a concentrated source of behavioural data and cyber risk. Consent Managers will either strengthen individual control or add another layer to the data economy, depending on registration standards, conflict-of-interest controls, auditability, interoperability and limits on secondary use.

The Government may designate Significant Data Fiduciaries by considering factors such as the volume and sensitivity of data processed, risks to Data Principals' rights, sovereignty, electoral democracy, security and public order. A designated entity must appoint a Data Protection Officer in India, engage an independent data auditor, conduct periodic data-protection impact assessments and audits and comply with additional prescribed measures.[20] This risk-tiered model sensibly links additional duties to systemic influence. Its legitimacy will depend on clear criteria, reasoned designation and a fair opportunity to challenge arbitrary selection. Impact assessments should inform actual decisions to deploy, redesign or stop high-risk processing and meaningful summaries should show affected people and the Board that necessity, alternatives and risks were genuinely considered.

7 State Processing Exemptions and Constitutional Proportionality

State processing is the most important test of constitutional fidelity. Section 7 permits specified uses without consent, including the provision of subsidies, benefits, services, certificates, licences or permits in defined circumstances; the performance of state functions under law or pursuant to judgments; and responses to medical emergencies and disasters. The Act also permits exemptions by Central Government notification and contains exemptions connected with sovereignty, integrity, security, public order, investigation and related grounds.[21]

---

[18] Digital Personal Data Protection Act, 2023, § 9; Digital Personal Data Protection Rules, 2025, rr. 10, 12, and Fourth Schedule.
[19] Digital Personal Data Protection Act, 2023, §§ 2(g), 6(7)–(9); Digital Personal Data Protection Rules, 2025, r. 4 and First Schedule.
[20] Digital Personal Data Protection Act, 2023, § 10.
[21] *Id.* §§ 7, 17.

Some flexibility is necessary because government cannot obtain bespoke consent for every legitimate public function. The constitutional question is whether the legal basis, purpose, scope, retention, sharing and safeguards are necessary and proportionate. Broad statutory language can allow welfare databases to migrate into policing, profiling or exclusion. The Rules' standards for state service delivery, including lawful processing, purpose limitation, necessity, accuracy, retention controls, security, notice and accountability, are useful.[22] They should be supported by published data inventories and purpose statements, access logs and independent audits.

The relationship between privacy and transparency also requires attention. The DPDP Act amended section 8(1)(j) of the Right to Information Act 2005, which governs the exemption for personal information.[23] Privacy law should prevent disclosure that causes unjustified harm without becoming a shield against accountability for public officials, public spending or regulatory action. Public-interest balancing, proportionality and severability remain important. The central constitutional danger is asymmetry: demanding the narrowest consent and compliance from the private sector while allowing the broadest discretion to the State. Puttaswamy is especially important where state databases affect access to essential services and rights.

8 DATA PROTECTION BOARD: INDEPENDENCE, CAPACITY AND REMEDY

The Data Protection Board connects rights on paper with enforcement in practice. It is a statutory body corporate, designed to operate digitally, with powers to inquire into non-compliance and breaches, order urgent mitigation, accept voluntary undertakings and impose monetary penalties. Decisions may be appealed to the Appellate Tribunal.[24] The Government legally established the Board in the National Capital Region and fixed its size in November 2025. Recruitment documents issued in 2026 contemplated a Chairperson and four members, but no completed appointments were identified in the public record by the end of August 2026.[25] In addition, the provisions conferring the Board's principal functions and procedures are subject to the later commencement schedule. Legal existence must therefore be distinguished from operational capacity.

The Board's independence is a central design concern. The Central Government appoints members through the statutory selection process; members serve two-year terms and may be reappointed; and the executive retains substantial rulemaking and administrative influence.[26] By comparison, GDPR Articles 51 and 52 require supervisory authorities to act with complete independence and to receive adequate human, technical and financial resources.[27] India need not copy the European model for its Board to be constitutionally valid. Even so, perceived dependence may weaken trust when the Board hears complaints about government data systems. Longer and staggered terms, published selection criteria and recusals, an independent secretariat and a separately disclosed budget would strengthen credibility.

Capacity is equally important. A Chairperson and four members may face complaints and violations across one of the world's largest digital populations. A digital-by-design process can widen access, but it

[22] Digital Personal Data Protection Rules, 2025, r. 5 and Second Schedule.
[23] Digital Personal Data Protection Act, 2023, § 44(3), *amending* Right to Information Act, 2005, § 8(1)(j).
[24] Digital Personal Data Protection Act, 2023, §§ 18, 27–29, 32–33.
[25] Notification Establishing the Data Protection Board of India, G.S.R. 844(E) (Nov. 13, 2025); Notification on the Size of the Data Protection Board of India, G.S.R. 845(E) (Nov. 13, 2025); Ministry of Electronics and Information Technology, Appointment to the Post of Chairperson and Other Members in the Data Protection Board of India, F No 2(1)/2026-Pers.I (May 6, 2026).
[26] Digital Personal Data Protection Act, 2023, §§ 19–20.
[27] Regulation (EU) 2016/679 (General Data Protection Regulation), arts. 51–52.

can also exclude people with poor connectivity, limited language options, disabilities or limited legal literacy. The Board should provide multilingual and accessible interfaces, assisted filing, reasoned and searchable decisions, predictable timelines and procedures for consolidating systemic cases. Effective remedies should include mitigation and corrective directions rather than treating monetary penalties as the only outcome. Annual statistics on complaints, sectors, disposal times, breaches, penalties, undertakings and compliance would allow Parliament and the public to assess performance.

9 From Legal Text to Organisational Practice

A sound implementation programme begins with governance, not a privacy notice. Boards and senior managers should establish ownership, risk appetite and the evidence required to demonstrate compliance. Every organisation should maintain records of processing that identify data types, sources, purposes, systems, locations, recipients, processors, retention periods and legal bases. Without such a map, an organisation cannot reliably answer access requests, delete data or assess breaches and unlawful use. Product approval and change-management procedures should enforce purpose limitation so that a new analytical use does not bypass the original notice.

Consent must be designed into and enforced through the interface. Organisations should separate purposes, eliminate pre-ticked boxes, retain notice versions, record affirmative action and propagate withdrawal to processors and downstream systems. Erasure should extend through production systems, archives and derived datasets, subject to documented legal-retention exceptions. Rights portals should use proportionate authentication and provide human escalation. Privacy by design should incorporate access control, data minimisation, encryption, tokenisation, safe testing data, logging and secure deletion throughout the development lifecycle.[28] For AI systems, organisations should document training-data provenance, inference risks, purpose compatibility and human review even though the Act creates no express right concerning automated decisions.

Supply chains are another channel through which accountability operates. Vendor due diligence should examine security, sub-processing, data location, breach history, deletion and audit rights. Contracts should reflect operational reality and be tested against evidence rather than questionnaires alone. Significant Data Fiduciaries should connect impact assessments and audits to decisions to deploy, redesign, restrict or abandon high-risk processing. The OECD privacy-management approach is useful because it treats accountability as a programme proportionate to an organisation's structure, scale, volume and sensitivity, supported by risk assessment and incident planning.[29] Substantive compliance requires an organisation to explain why it processes data, how it minimises risk and what it did when controls failed.

10 Comparative Lessons Without Regulatory Copying

Comparison is useful, but India's framework is deliberately more minimalist than the GDPR. The GDPR provides multiple lawful bases, special-category protections, extensive processor duties, rights to portability and objection, safeguards for automated decision-making, impact assessments for high-risk

[28] European Data Protection Board, Guidelines 4/2019 on Article 25 Data Protection by Design and by Default (final version 2020).

[29] OECD, Recommendation of the Council Concerning Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data (2013 revision); OECD, Going Digital Guide to Data Governance Policy Making (2022).

processing and independent supervisory authorities.[30] The DPDP Act relies more heavily on consent and specified legitimate uses, delegated rulemaking and additional duties for notified Significant Data Fiduciaries. Simplicity may reduce compliance complexity, but it can also leave protection dependent on future designation and interpretation.

Three comparative lessons are especially relevant. First, regulatory independence and resources are functional requirements, not institutional decoration. Second, consent should be supplemented by risk-based accountability because people cannot negotiate their way out of surveillance or insecure infrastructure. Third, guidance and enforcement should be transparent. European experience also shows that broad rights may accomplish little without an empowered, well-resourced and efficient regulator. India should not assume that a shorter statute is easier to enforce or that a high maximum penalty necessarily produces deterrence.

India can develop a model suited to its scale, linguistic diversity and digital public infrastructure. Consent Managers could become distinctive rights-enabling institutions; digital Board proceedings could reduce geographic barriers; and interoperable technical standards could improve withdrawal and grievance procedures. These benefits depend on privacy-enhancing, auditable systems that are accessible beyond smartphone-centred, English-language services. Regulation that supports innovation must still make rights effective and difficult to evade.

11 RECOMMENDATIONS FOR THE IMPLEMENTATION PHASE

The phased commencement period should be treated as a regulatory build-out period. MeitY and the Board should publish a common implementation calendar, model notices, breach-reporting formats, standards for rights requests and sector-specific guidance. The Government should also publish transparent risk criteria for designating Significant Data Fiduciaries. Before the substantive duties commence, the Board should have trained legal, technical and investigative staff and tested digital and assisted-filing channels. Appointment and budget information should be public and decisions should be searchable and available in major Indian languages.

Organisations should prioritise demonstrable controls: executive accountability, processing inventories, lawful-purpose review, consent records, retention schedules, rights workflows, processor governance, incident exercises and performance metrics. Organisations engaged in high-risk processing should conduct impact assessments even before formal designation. Public entities should publish data-purpose registers and audit access to population-scale datasets. Children's services should use proportionate age assurance and prevent behavioural advertising by design. Exemptions should be limited, reasoned and time-bound, while model documents and technical support should help MSMEs comply.

Implementation should be assessed by outcomes. Useful indicators include the time taken to fulfil rights requests, the proportion of withdrawn consents propagated across systems, deletion after a purpose expires, breach-detection and notification times, recurrence of control failures, accessibility of complaint mechanisms, remedial orders and sectoral patterns of harm. Civil-society organisations, researchers and consumer groups should have access to anonymised enforcement data. After the first enforcement cycle, parliamentary review should consider whether the exemptions, Board structure and catalogue of rights remain adequate.

[30] Regulation (EU) 2016/679 (General Data Protection Regulation), especially arts. 5, 9, 12–22, 25, 28, 32–35, 51–52.

12 Conclusion

The Digital Personal Data Protection Act 2023 creates India's first general law for digital personal data and gives statutory form to the fundamental right to privacy recognised in Puttaswamy.[31] It requires Data Fiduciaries to process personal data for lawful purposes on the basis of valid consent or another authorised ground, adopt reasonable security safeguards and notify affected individuals and the Data Protection Board of personal-data breaches.[32] A food-delivery application that requests a user's name, address and telephone number should explain why those details are needed and should not seek access to photographs, contacts or location history unless they are necessary for a stated purpose. The DPDP Rules 2025 add operational detail, while phased commencement gives organisations time to adjust systems and internal processes.[33]

Consent alone cannot ensure privacy where people have unequal power. An employee may accept workplace surveillance for fear of losing a job, while a student may accept an educational application's privacy policy because no realistic alternative exists. The absence of a general category for sensitive personal data is also significant: disclosure of medical, biometric or financial information can cause far greater harm than disclosure of an ordinary shopping preference. Similarly, the Act's limited protections concerning profiling and automated decision-making may leave people without a clear statutory remedy when an algorithm rejects an application for credit, employment or insurance without an adequate explanation or human review.

The Act should therefore be understood as a foundation for India's privacy system rather than a complete achievement. Effective protection will require an independent, adequately staffed and accessible Data Protection Board, responsible organisational conduct and constitutional limits on state processing. Collecting identity data to administer welfare benefits may serve a legitimate public function; indefinitely linking it with medical, travel and communications data may be disproportionate. The regime's success should be measured not by the number of consent notices or penalties, but by whether people can exercise their rights easily and whether organisations and public authorities routinely collect only necessary data, protect it and erase it when it is no longer required.

* * *

[31] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1; Digital Personal Data Protection Act, 2023.
[32] Digital Personal Data Protection Act, 2023, §§ 4, 6–8.
[33] Digital Personal Data Protection Rules, 2025, G.S.R. 846(E) (Nov. 13, 2025), as corrected; Commencement Notification, G.S.R. 843(E) (Nov. 13, 2025).

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.19

CHAPTER 19

"You Can Reset a Password, But Not Your Biometrics": The Problem of Irreversible Biometric Data

AAKRITI JAIN* AND SHREYA AGARWALA**

Traditionally, passwords have been used as the first-line guardian of digital identity, but they have proven rather fragile, as users tend to use either predictable passwords that can be easily hacked or passwords that are difficult to remember. These consistent drawbacks have pushed us towards developing new methods of verification that are much faster, convenient and safer to use. Biometric authentication (fingerprinting, facial recognition, iris scanning, voice recognition) is the first thing that springs to mind in this scenario. While there have been numerous studies on the effectiveness and convenience of biometrics in the past, one important issue that needs to be answered is whether the existing legal framework is prepared for the unique challenge presented by irreversibility. A legal, doctrinal and comparative approach has been applied in this paper in analysing data protection laws, judicial decisions and guidelines from selected jurisdictions to assess whether the current remedy framework sufficiently accounts for the enduring nature of biometric damage. It argues that the issue of irreversibility requires a unique approach to regulate, i.e., an approach based on proactive precautionary measures instead of reactive ones and ends with suggestions for how this could be done.

Keywords: Biometric authentication, irreversibility, data protection law, privacy, biometric damage, digital identity

1 INTRODUCTION

A password is a matter of fact about what an individual remembers while the biometric identifier is a matter of what an individual is. The difference may be somewhat academic, but it serves as the point of departure for this paper. Passwords can never be right as they can only work or fail. Indeed, passwords are constantly failing and are being reset. Breach alerts typically conclude with an instruction to reset the password, no matter how often that process has already been undergone by the victim. In contrast, once the fingerprint, the iris pattern, the facial geometry or the voiceprint has been captured and stored, it can never be changed again in one's lifetime, and possibly much longer than that.

The ease of use of biometrics for authentication is not under contention here and is already widely accepted as it is faster than entering a password and it is impossible to forget in the same sense as a PIN can be and most importantly it protects against the most rudimentary form of credential sharing and phishing attacks. Indeed, governments have invested so heavily in national identity frameworks based on biometrics, like India's Aadhaar program, the largest of its kind in the world. Likewise, private sector platforms have integrated fingerprinting and facial recognition into the basic experience of mobile device use. The issue that has been lagging behind all of this is the legal understanding of what happens once the template itself is breached. Data protection laws were built upon an architecture that assumes that

* Student at Unitedworld School of Law, Karnavati University, Gandhinagar, Gujarat, India.
** Student at Unitedworld School of Law, Karnavati University, Gandhinagar, Gujarat, India.

any given data could be easily renewed, superseded or become obsolete through the simple process of issuing a new one. Biometric data is an exception to this rule.

2 RESEARCH METHODOLOGY

2.1 Research Questions and Methodology

The two guiding research questions for the analysis include:

- firstly, whether the irreversibility of biometric identifiers raises a novel legal risk that cannot be sufficiently covered by traditional breach and remedy approaches; and
- secondly, whether biometric-specific preventive measures ought to be required under Indian data protection law, instead of imposing onerous horizontal obligations as to all personal data.

The paper uses a comparative method based on doctrinal research and compares Indian constitutional privacy law to the Digital Personal Data Protection system, along with the EU special-category approach and Illinois biometrics legislation.[1]

2.2 Research Gap and Contribution

Past literature related to biometrics often tends to be centred on themes like surveillance, consent, accuracy, discrimination and cybersecurity. However, in this paper, the gap has been identified in terms of doctrine, specifically the clash between the nature of the biometric identifier's permanence and the remedy-oriented approach that takes for granted containment of the compromised information. The primary contribution of this research is the formulation of an "irreversibility principle", which holds that when the data principal cannot reasonably replace the compromised biometric identifier, the duty of protection should rest with the entity that decides to capture, store, and use that identifier.[2]

3 IRREVERSIBILITY AS A UNIQUE LEGAL HARM

3.1 The Unique Character of Compromise of a Biometric Identifier

Earlier, laws designed for data protection have been constructed on the presumption derived from the fields of financial and identification security, that of replacing an identifier in case of its breach. For instance, a credit card number, once revealed, is cancelled and replaced; similarly, passwords, once revealed, are reset by the system and even Social Security and Aadhaar numbers, which are very hard to recover once leaked can still be flagged, monitored and eventually reissued and/or supplemented with further means of verification. Biometric identifiers fall outside of the scope of this paradigm. A person only has a finite set of ten fingerprints, two irises, one face and one voice. In case the template of any of those is compromised, a replacement cannot be made as in the case of passwords and the same identifier will keep being used indefinitely even for the rest of a person's lifetime and across different systems unaware of this breach.

[1] Digital Personal Data Protection Act, 2023, §§ 4–7; Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 9, 2016 O.J. (L 119) 1, 38–39 (2016).

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council, arts. 5, 32, 2016 O.J. (L 119) 1, 35, 51–52 (2016); Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/5(c), 15(a)–(b) (2024).

This is the exact rationale behind the passage of the Biometric Information Privacy Act (BIPA) by the Illinois legislature in 2008 which is one of the very first and most significant biometrics-specific laws ever passed in the world. The findings in the law itself recognise the distinction between biometric identifiers and other unique identifiers because of the simple reason that while a Social Security number can be altered after it is compromised, there is no way for a biometric identifier to be altered since it is biologically determined.[3]

The doctrinal importance of this legislative holding is that it changes the nature of the injury. In any traditional data breach claim, the act of exposure is wrongful, while the actual injury arises from its misuse. This principle is codified in the standing doctrine used by courts of law across many jurisdictions, including the United States, in determining the viability of any data breach claims when plaintiffs usually must establish an appreciable risk of the future misuse of their personal information, rather than its actual exposure. The Illinois courts' approach to interpreting BIPA is quite different, however, since the very act of collecting biometric data without notifying the individual in writing about his/her rights and obtaining his/her written consent constitutes a violation of the privacy right. The body of BIPA case law on this topic culminates in Rogers v. BNSF Railway Co. and similar lawsuits in which the court has ruled against a rail operator for its reckless or intentional violation of BIPA by fingerprinting thousands of truckers without written, informed consent.[4]

3.2 Irreversibility Accumulates over Time and Across Other Systems

The second aspect that makes the breach of biometrics stand out from a regular data breach is the ability of the compromise to accumulate. While a password can no longer be used in any other place after a reset, if the user takes care of this problem, a compromised biometric template could potentially have found its way into multiple systems: a mobile phone, a company security system, a countrywide identification system, customer identification at a bank, an investigative database in a state police force, all of which would independently collect the same physical trait from the person.

The cumulative quality of such a problem is precisely the reason why the argument developed within this paper separates the issue of irreversibility from a question of degree. An approach which is geared towards addressing the possibility that, for example, an email address might be collected for the purpose of spamming cannot be properly balanced to address the problem of an irreversibility that is cumulative, cross-situational and cannot be fixed once achieved. Any legal mechanism which attempts to respond mainly through post-fact notification and monetary compensation will necessarily treat the wrong as a transitory rather than permanent one.

4 Comparative Analysis

4.1 European Union: Biometric Data as a Special Category

The General Data Protection Regulation (GDPR) is perhaps the most complex regulation of the three regimes that have been discussed in this paper since it categorises biometric data in an entirely unique manner. According to Article 9 of the GDPR, biometric data used for uniquely identifying a natural person belongs to a closed list of "special categories" of personal data. Along with information about the

[3] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/5(c) (2024).
[4] Rogers v. BNSF Ry. Co., 680 F. Supp. 3d 1027, 1032–34 (N.D. Ill. 2023).

health status of a person, genetic data, racial or ethnic origin of the individual and some other categories of personal information, which present a particularly high risk in case of processing, biometric data is subject to certain restrictions. Processing biometric data is prohibited by default and allowed in specific instances described in Article 9(2).[5]

The doctrine of a special-category approach, as far as the irreversibility argument put forward here is concerned, consists of the fact that it introduces an element of asymmetry into the very beginnings of the regulatory regime instead of depending only on the remedy stage. The data controller wanting to process biometric data according to the requirements of the GDPR has to comply with a higher standard of compliance before the very possibility of collecting it and the very reason why a higher standard is required lies precisely in the increased danger posed to the individual by the processing of such data and a danger which as has been suggested earlier is essentially irreversibility-based. While this model alone does not resolve the irreversibility dilemma since a special category can be broken through and the danger would remain unresolved nonetheless, it does something that reactive measures cannot. It structurally discourages collecting biometric data in the first place, since what has never been collected cannot be irreversibly damaged.

4.2 The United States: Sectoral Differences, Partially Remedied by State Law

In the United States, there is no broad federal data protection legislation like the GDPR, nor does biometric data receive any form of federal protection. Instead, a select few states, including Illinois, Texas, and Washington State (with the latter two lacking a private right of action) have their own independent biometric privacy laws. BIPA's structure is instructive precisely because it appears to have been drafted with irreversibility as its central premise rather than an incidental concern. Under the statute, the private entity must have a publicly accessible written policy on retention and destruction in place before collection, obtain informed, written consent before collection and destroy the biometric data after the initial purpose of its collection is served or within a statutory time period, whichever is sooner.[6] Most importantly, BIPA provides a private right of action with an already determined sum of money for each breach, thus eliminating the proof-of-loss issue in many instances.

The potential magnitude of exposure that this liability model could create for the non-compliant organisation has been illustrated through Rogers v. BNSF Railway Co., wherein a jury awarded damages against the defendant for thousands of statutory violations committed due to the unwitting fingerprinting of truckers as they entered its rail yards, yielding an initial judgment of $228 million that was modified based on subsequent guidance from the Illinois Supreme Court on the accumulation and discretionary nature of statutory damages. Regardless of any assessment regarding the magnitude of this exposure, the liability model deserves consideration on its own merits. BIPA does not require proof that the compromised fingerprint subsequently led to a fraud perpetrated upon the person. Rather, the liability is in place due to the fact that the individual's inability to ever change their compromised fingerprint creates the complete harm itself. This is exactly the kind of front-loaded and non-discretionary form of liability that a regime sensitive to irreversibility needs, despite the limited geographical scope of the statute and the fact that few American states provide similar protections.[7]

[5] Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 9, 2016 O.J. (L 119) 1, 38–39 (2016).
[6] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15(a)–(b) (2024).
[7] Rogers v. BNSF Ry. Co., 680 F. Supp. 3d 1027, 1032–34 (N.D. Ill. 2023).

4.3 India: Constitutional Acknowledgement Without a Statutory Special Category

India offers the best example for this paper's claim due to the unusual discrepancy between constitutional law and statutory provisions. The 2017 Supreme Court judgment made in a case of nine judges through Justice K.S. Puttaswamy v. Union of India declared that the right to privacy is a fundamental right and it constitutes an integral part of the Right to Life and Personal Liberty guaranteed under Article 21 of the Constitution according to the constitutional law theory.[8] On the other hand, the 2018 five-judge bench case of Justice K.S. Puttaswamy (Ret'd) v. Union of India (Aadhaar), where the constitutionality of India's national biometric ID scheme based on the proportionality test was considered, stated the objective of Aadhaar as delivering welfare benefits as justifiable, and the scheme thus passed the proportionality test. Nonetheless, Justice D.Y. Chandrachud's dissenting opinion is doctrinally relevant in this context insofar as he has regarded the privacy right of an individual with respect to his biometric data as particularly important due to the permanence and association with bodily integrity of the data in question, in view of the fact that the compromise of the biometric database cannot be corrected as other kinds of compromised information can.[9]

4.3.1 Digital Personal Data Protection Act, 2023: Protection of Irreversible Biometric Identifiers

The Digital Personal Data Protection Act, 2023 (DPDP Act), which was assented to by the President of India in August 2023, and whose Rules were notified on 13 November 2025 and which provides for gradual enforcement, including many of its main operative provisions eighteen months after notification, has intentionally diverged from the special category approach of the GDPR and previous Indian draft data protection laws. The DPDP Act does not have any special category of sensitive personal data and biometric data is governed by precisely the same horizontal standards of notice, consent or one of nine other exceptions to consent, purpose limitation and security as the person's name and address.[10] The Act does not refer to biometric data at all.

This is not a mere question of drafting differences, but a real regulatory gap. In the light of the DPDP Act regime, any organisation which processes fingerprints or face data, whether for purposes of work attendance or building entry, stands subject to more or less similar threshold obligations as compared to those organisations that process email IDs for a newsletter, even though the November 2025 Rules have imposed special obligations such as the appointment of Data Protection Officers and periodic audits on Significant Data Fiduciaries, which are organisations that qualify as such because of the nature or scale of processing, though in practice these will include biometric processors too.[11]

The end result is thus of a state whose Supreme Court has acknowledged, in terms of constitutional theory, the special vulnerability associated with biometric information, yet whose main data protection law has failed to incorporate this insight into a unique rule for collection, a deadline for destruction, or a template protection requirement akin to those contained in BIPA as a matter of private enforcement, and/or in the GDPR as an exception category.[12] It is in this way, the present paper argues, that the gap

[8] Justice K.S. Puttaswamy (Ret'd) v. Union of India, (2017) 10 SCC 1, 260–62; INDIA CONST. art. 21.
[9] K.S. Puttaswamy (Ret'd) v. Union of India (Aadhaar), (2019) 1 SCC 1, 1026–27, 1400–01 (Chandrachud, J., dissenting).
[10] Digital Personal Data Protection Act, 2023, §§ 4–7.
[11] *Id.* § 10.
[12] Digital Personal Data Protection Act, 2023, §§ 4–8; Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15; Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 9, 2016 O.J. (L 119) 1, 38–39 (2016).

between the rationale of the Aadhaar dissent and the design of the DPDP Act can be most clearly articulated.

Thus, the fundamental problem is not the lack of privacy laws in India, but the lack of a regulatory trigger for biometric data which is distinct from other types of personal information. Even in the presence of a horizontal regime governing the collection and use of biometric data, it will fail to account for the fact that the risk associated with an error is very high since there is no way for the data subject to get a replacement face, fingerprint, iris, or voice. India should thus distinguish regular personal data and data whose misuse results in an irreversibly changed status with respect to the safety of the data subject.

4.3.2 Right to Information Act, 2005: Transparency Versus Biometric Privacy

The Right to Information Act of 2005 becomes equally significant in relation to the governance of biometric information of individuals that is held by public authorities. Since biometric information needs to have strict privacy protection because of the fact that once the fingerprint, facial geometry, iris pattern or voiceprint becomes compromised, it cannot usually be replaced, the process of collecting, processing and storing of such information needs to remain transparent. Therefore, in the context of biometric databases, the RTI framework can become a complementary tool, which will allow citizens to obtain information about the policies and procedures of handling, retention and securing of biometric information, as well as about the mechanisms of its governance. However, the right to information is not an unlimited right to access any kind of personal or sensitive information. The exemptions listed in Section 8(1)(j) of the RTI Act, especially those relating to security issues and information that involves personal privacy, help balance governmental transparency with the protection of individuals from unnecessary disclosure.[13] This distinction becomes even more important for biometric databases like the Aadhaar database: public institutions need to be answerable about how biometric data is managed, without making individuals' biometric identifiers publicly available. In this respect, the RTI can complement the precautionary principle that has been suggested in this paper.

5 FROM REACTIVE REMEDY TO PROACTIVE DESIGN

5.1 Comparative Synthesis

The comparison gives rise to regulatory instincts: the GDPR increases the level of lawfulness of data processing, BIPA transforms duties specific to biometrics into legal obligations protected through a private remedy and India's DPDP regime offers a horizontal baseline but fails to address the particularity of the biometric issue.[14] For India, the lesson is not to emulate either system in toto; instead, it needs to integrate the notion of front-end cautioning alongside the recognition that capturing biometric data illegally is in itself an actionable wrong.

This comparative analysis leads to a more focused proposition than the general one regarding stricter penalties. The issue is not that the penalties are insufficient but the problem is that the entire framework of remedies – notice of breach, mitigation and compensation tied to actual loss – assumes a harm which is finite and, at least in part, reversible. If the harm is irreversible, by definition it is not something which

[13] Right to Information Act, 2005, § 8(1)(j).

[14] Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 9, 2016 O.J. (L 119) 1, 38 (2016); Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15, 20 (2024); Digital Personal Data Protection Act, 2023, §§ 4–8.

can be reversed: by the time the legal system intervenes in response to the breach, it is already too late for this type of data, regardless of whether the breach occurred accidentally or on purpose. This is not to argue that biometric authentication should not be used as its benefits in comparison to passwords are undeniable. The point, however, is the need to reframe biometric regulation around precaution rather than remedy.[15]

5.2 Template Protection Is Mandatory

One issue raised frequently in the technical literature regarding security issues relating to biometric systems is the difference between storing a raw biometric image and storing an irreversible and non-reversible cryptographic template based on that raw image. Techniques classified under 'cancellable biometrics' and 'biometric template protection' were introduced specifically so that, if compromised, the stored template can be invalidated and a new template can be created by applying a new transformation to the underlying physical feature, which creates an analogous ability to reset the biometric to the same degree as an ordinary credential. There are standards documents produced by the US National Institute of Standards and Technology outlining this very category of technique for over ten years, yet very few data protection laws mandate its use.[16]

A precautionary approach to regulation would demand by statute that an organisation using biometrics for identification store only biometric templates which cannot be reversed and which have been created using such processes, but not using any raw image or matchable biometric feature sets and would make the legitimacy of the collection process dependent upon demonstrable adherence to a recognised technical standard.

5.3 Purpose-Based and Time-Limited Retention as a General Rule by Default

The provision under BIPA that a private entity must have a written retention and destruction policy before the collection of biometric information and the deletion of biometric information after the satisfaction of the intended purpose or the expiry of a statutory period (whichever is earlier) is a prudent design that should be considered for adoption outside of Illinois and beyond the private sector. While the principle of limitation of processing on grounds of purpose under the DPDP Act limits data usage after the intended purpose is achieved, there is no maximum limit for biometrics and the practice of implementation follows the complaint-based approach of the DPDP Act.[17]

A ceiling on retention specific to biometric identifiers, such as a hard-coded maximum duration unless there is a valid ongoing consent or statutory mandate, would lower the number of vulnerable records at any one time and therefore, the scope of any potential breach incident, which is precisely what the precautionary approach would seek to mitigate, since the harm resulting from its triggering cannot subsequently be mitigated.

---

[15] Rosenbach v. Six Flags Ent. Corp., 129 N.E.3d 1197, 1206–07 (Ill. 2019); Cothron v. White Castle Sys., Inc., 2023 IL 128004, ¶¶ 20–27 (Ill. 2023).

[16] NIST SPECIAL PUBLICATION 800-63B § 5.2.3.

[17] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15(a) (2024); Digital Personal Data Protection Act, 2023, §§ 8, 12.

5.4 Non-Contingent Liability Calibrated to Incentivise Collection Prevention and Not Misuse

The most instructive element of the American experience, when viewed from a structural perspective, is that a liability regime requiring a plaintiff to demonstrate the misuse, financial harm, or risk of future identity theft of a biometric data identifier before awarding any form of compensation asks too much of the injured party. The difficulty inherent in such a requirement derives from the reality that misuse of a captured biometric may take place several years down the road via entirely unconnected means. Thus, while Illinois' Biometric Information Privacy Act's model of statutory liquidated damages, based on unlawful capture rather than actual misuse, does not solve the problem of irreversibility, it recognises this difficulty by making it an intrinsic part of its liability formula: a biometric data capturer takes the risk that capturing such data without a lawful process will come at a cost.[18]

India is an example of a jurisdiction that currently relies on the administration of penalties via the Data Protection Board in the DPDP Act and not the creation of an individual enforceable private right of action in order to enforce the Act.[19] It would be necessary for this jurisdiction to balance the cost of creating the private right of action against the precautionary principle of incentive alignment.

5.5 Consent as a Threshold Gate and Not as a Liability Shield

Lastly, it is essential for a precautionary approach to avoid seeing the act of obtaining consent as a sufficient response to the problem of irreversibility. The consent mechanisms within all three regimes mentioned earlier serve their purpose quite effectively as a gate into collecting biometric information; however, they fail to solve the question of how to deal with the consequences of the data becoming vulnerable without any further involvement of the data principal in the process. The person who consented to having his fingerprint scanned for gaining entry to a facility did not consent to the risk of that information becoming public due to the vendor's breach of security in another context. It follows that in a precautionary regime, consent is only necessary but not sufficient: where collection is legal in terms of the justification for collecting data, there should be safeguards both technological and in terms of retaining data regardless of whether or not the consent is legally obtained.[20]

TABLE 1. *Treatment of biometric data under the GDPR, BIPA and the DPDP Act*

Issue	**GDPR**	**BIPA**	**India (DPDP)**
Regulatory Treatment	Special Category/ High Bar	Biometric-Specific Law	Horizontal Personal Data Law
Collection	Prohibited unless Article 9 condition satisfied	Written notice and informed consent	Lawful purposes within DPDP regime
Retention	Principle of storage limitation	Written retention/destruction policy	Broad principles; no specific biometric retention limit

[18] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/20 (2024); Cothron v. White Castle Sys., Inc., 2023 IL 128004, ¶¶ 20–27 (Ill. 2023).

[19] Digital Personal Data Protection Act, 2023, §§ 27–33.

[20] Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 9(2)(a), 2016 O.J. (L 119) 1, 38–39 (2016); Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15(b) (2024); Digital Personal Data Protection Act, 2023, § 6.

Issue	GDPR	BIPA	India (DPDP)
Remedy	Administrative/judicial remedies	Private cause of action and statutory damages	Admin enforcement through Board regime
Basic Message	Prevent unnecessary data collection	Illegal data collection must have independent consequences	Add biometric-specific duties

6 OBJECTIONS AND LIMITATIONS

Five objections are posed against the proposal:

- First, there could be improvement in the security and ease of access through biometric authentication and restrictive measures might drive organisations to go for less secure methods. The regulation would thus focus on the conditions of use instead of prohibiting biometrics.
- Second, no method of template protection will make biometric authentication completely safe. It is rather an issue of risk reduction and legal management and not technology.[21]
- Third, compulsory template protection architecture, fixed retention limits and independent certification requirements will incur unavoidable compliance costs, which will unfairly burden small data fiduciaries and new entrants in the market and may even cement the position of incumbent firms who can better bear these compliance costs compared to small firms, which could offer services such as biometric attendance or access control solutions. In any event, this is an unavoidable trade-off rather than an argument against adopting a precautionary design, since a differentiated compliance regime, depending on either the amount of processing or the sensitivity of the use case, as in the GDPR's proportionality principle, would alleviate some of the burden.
- Fourth, detaching recovery from actual harm results in disproportionate litigation. The ruling in Cothron v. White Castle System, Inc., by the Supreme Court of Illinois, whereby the Illinois Supreme Court ruled that there is a violation of the BIPA every time biometric data is captured and transmitted, rather than once per collection, turned normal timekeeping into a potential basis for statutory liability and led to the Illinois legislature capping damages for such violations shortly thereafter. In adopting the non-contingent approach of BIPA, an alternative jurisdiction needs to incorporate the fine-tuning that the Illinois legislature failed to include in the statute, either by setting statutory damages on the basis of one accrual of liability per incident for each data subject, or by instituting a proportionality rule.[22]
- Fifth, making the consent gate tougher will create conflict for those who rely on biometric authentication systems, especially those based on Aadhaar to gain access to various welfare benefits and essential services and for whom refusal of consent does not actually represent a viable alternative. A precautionary principle intended to shield people from irremediable damage should not be applied in such a way that exacerbates the exclusion of the very group it intends to help. Such conflict can be better handled using the other parts of the above proposal, namely retention and deletion and liability provisions without recourse to consent procedures.

[21] NIST SPECIAL PUBLICATION 800-63B § 5.2.3.

[22] Cothron v. White Castle Sys., Inc., 2023 IL 128004, ¶¶ 1, 20–27 (Ill. 2023).

7 IMPLEMENTATION

An irreversibility-sensitive system would regulate the circumstances under which biometrics can be used and not ban biometrics. These regulations embody the normative principle contained in the paper in actionable legal obligations:

7.1 Necessity Analysis Before Collection

The data fiduciary must state why biometric authentication is necessary, why less invasive identifiers are insufficient and an alternative that was considered. Processing of high-risk information requires a documented risk analysis.[23]

7.2 Template-First System Architecture

Raw biometric images must not be stored whenever possible when authentication may be performed using protected templates. Compliance must be independently verifiable.[24]

7.3 Retention and Deletion Control Mechanisms

Each biometric system must have a defined purpose, maximum retention period, deletion controls and an established process for retention exceptions where required by law.[25]

7.4 Irreversible Breach Response Mechanism

The biometric breach response should not only include notification but also contain containment, forensic examination, downstream processor identification, enhanced authentication methods and appropriate reporting to regulators.[26]

7.5 Remedies and Accountability

Unlawful collection or retention of biometrics should be able to attract an appropriate remedy regardless of whether downstream identity theft occurs in all cases. There must be an established means through which people can complain.[27]

7.6 Technical Standards Independent of the Legislation

The regulator should establish or adopt minimum standards for the protection of templates, encryption, access controls, key management, proof of destruction and auditing of processes.[28]

---

[23] Digital Personal Data Protection Act, 2023, §§ 4–8; Regulation (EU) 2016/679 of the European Parliament and of the Council, arts. 5, 25, 35, 2016 O.J. (L 119) 1, 35, 39, 56 (2016).
[24] NIST SPECIAL PUBLICATION 800-63B § 5.2.3.
[25] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15(a) (2024); Regulation (EU) 2016/679 of the European Parliament and of the Council, art. 5(1)(e), 2016 O.J. (L 119) 1, 35 (2016).
[26] Regulation (EU) 2016/679 of the European Parliament and of the Council, arts. 33–34, 2016 O.J. (L 119) 1, 51–52 (2016); Digital Personal Data Protection Act, 2023, § 8(5)–(6).
[27] Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/20 (2022); Cothron v. White Castle Sys., Inc., 2023 IL 128004, ¶¶ 20–27 (Ill. 2023).
[28] NIST SPECIAL PUBLICATION 800-63B § 5.2.3.

8 Conclusion

Biometric authentication addresses a genuine vulnerability associated with passwords, but this paper does not dispute the use of biometrics. What the paper disputes is that the laws constructed to govern the use of biometrics were, for the most part, constructed on an entirely different set of assumptions about data – namely data which could be reissued, substituted and managed after a breach. The special category approach of the EU increases the bar to collection of such data, but no more. Among the various models proposed in the sectoral approach to biometrics regulation, the American one, illustrated by BIPA in Illinois, represents the most precautionary one because liability is triggered simply upon the fact of unconsented collection as opposed to demonstrable damage downstream, but still suffers from the geographic and institutional fragmentation. India provides the most vivid example of the chasm highlighted throughout this paper: the biometric-specific constitutional jurisprudence, especially that of Justice Chandrachud in his dissenting opinion in the Aadhaar case, which recognises biometric data as uniquely and irreversibly vulnerable data, coexisting with a newly implemented statutory regime which fails to incorporate this understanding into a biometric-specific standard of collection and retention. This can be achieved without entirely discarding the principle-based horizontal approach established in the DPDP Act, and only requires adding biometric-specific protections, in terms of template protection standards and strict ceilings on retention, as well as a more precautionary liability rule. Any rule of law which only reacts after irreversible damage has occurred fails to provide protection for the individual because it is reactive in nature.[29]

* * *

[29] NIST Special Publication 800-63B, *supra* note 16, § 5.2.3; Regulation (EU) 2016/679, art. 9; Biometric Information Privacy Act, 740 Ill. Comp. Stat. 14/15, 20 (2022); K.S. Puttaswamy (Ret'd) v. Union of India (Aadhaar), (2019) 1 SCC 1, 1026–27 (Chandrachud, J., dissenting).

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.20

CHAPTER 20

Right to be Forgotten vs. Right to Information: Finding the Digital Balance

DR. GARGI BHATT*

The rapid expansion of digital platforms has fundamentally changed the way information is created, accessed, stored and circulated. Information once published online can remain searchable and accessible for years, often long after its relevance has diminished. In this context, the emerging Right to be Forgotten (RTBF) seeks to protect individuals from the continuing digital consequences of outdated, irrelevant or disproportionately harmful personal information. At the same time, the Right to Information (RTI) serves as an important instrument of transparency, accountability and informed democratic participation. The increasing overlap between these two rights raises a difficult question: how can individual privacy and dignity be protected without undermining society's legitimate interest in accessing information?

This paper examines the evolving relationship between the RTBF and the right to information, particularly within the Indian constitutional and legal framework. It explores the tension between privacy, reputation and informational autonomy on one hand, and freedom of expression, transparency and public interest on the other. The paper also considers the challenges posed by search engines, social media platforms and the permanence of digital records, where removing information from one platform may not necessarily erase its wider digital footprint.

The study argues that neither right should be treated as absolute. A balanced approach requires contextual assessment based on factors such as the nature and sensitivity of the information. The paper concludes by proposing the need for clearer legal standards and transparent mechanisms that can reconcile privacy with the democratic value of access to information in an increasingly digital society.

Keywords: Right to be Forgotten; Right to Information; Privacy; Digital Rights; Data Protection; Public Interest

1 INTRODUCTION

The digital environment has altered the relationship between information, time and identity. In the analogue world, information about an individual was dispersed across archives, newspapers, registries and personal recollection. Its practical accessibility was limited by geography, cost, institutional fragmentation and the ordinary processes of human forgetting. The internet has displaced these limitations. Information can now be copied indefinitely, indexed globally, retrieved instantaneously and recombined with other data to construct enduring profiles of individuals. The past is no longer merely remembered; it is continuously searchable.

This transformation has generated a conflict between two important but potentially competing interests. On one side lies the right to be forgotten, understood broadly as an individual's ability to seek the removal, erasure, de-indexing, anonymization or reduced accessibility of personal information. It is

* Assistant Professor & Head, Training & Placement Office at DME Law School, Delhi Metropolitan Education, Noida, Uttar Pradesh, India.

connected with privacy, dignity, autonomy, reputation and the possibility of personal development or social reintegration. On the other side lies the right to information, including freedom of expression, freedom of the press, the public's right to know and access to matters of legitimate public concern. Democratic societies require the circulation of information, including information that may be uncomfortable, embarrassing or critical of powerful individuals.

The conflict is not limited to obviously false or unlawful material. It frequently concerns information that was lawfully published, substantially accurate and once newsworthy, but which has become disproportionately harmful because search engines make it permanently prominent. A criminal conviction, bankruptcy, family dispute, professional disciplinary proceeding or youthful indiscretion may remain technically available for decades. The question is therefore not simply whether the information is true. It is whether continued, name-based prominence remains justified in light of changed circumstances.

European law has generally treated privacy and freedom of expression as rights of equal normative importance. Neither is automatically superior; the outcome depends upon the circumstances of the case.[1] The United States, by contrast, has been more resistant to a general right to be forgotten because of the First Amendment's strong protection of truthful publication and public access to information.[2] India occupies a developing middle position. The Supreme Court has constitutionally recognized privacy, including decisional autonomy and informational control, while the Digital Personal Data Protection Act 2023 creates a statutory framework whose relationship with freedom of expression and archival access remains unsettled.[3]

This chapter argues that the appropriate response is neither an absolute right to erase the past nor an unrestricted right of indefinite digital exposure. A principled, context-sensitive balance is required. That balance should distinguish between deletion and de-indexing, public and private information, public officials and ordinary individuals, current and historical relevance, truth and falsity, serious harm and ordinary embarrassment, and the interests of the data subject and those of the wider public. It should also recognize that the institutional location of decision-making matters. A system that leaves fundamental-rights adjudication primarily to search engines risks creating opaque private jurisprudence. The proper objective is therefore a framework in which individual protection is effective, public access is preserved where justified, and decisions are transparent, reviewable and proportionate.

2 CONCEPTUAL FOUNDATIONS

2.1 The Right to Be Forgotten

The expression "right to be forgotten" is rhetorically powerful but legally imprecise. It may describe several distinct remedies:

a. Deletion of personal data from the original source;

b. Erasure by a controller or processor;

---

[1] S. Kulk & Frederik Zuiderveen Borgesius, *Privacy, Freedom of Expression, and the Right to Be Forgotten in Europe*, *in* THE CAMBRIDGE HANDBOOK OF CONSUMER PRIVACY 301–20 (E. Selinger, J. Polonetsky & O. Tene eds., 2018), https://doi.org/10.1017/9781316831960.018.

[2] Amy Gajda, *Privacy, Press, and the Right to Be Forgotten in the United States*, 93 WASH. L. REV. 201 (2018), https://digitalcommons.law.uw.edu/wlr/vol93/iss1/5.

[3] Gursimar Kaur, *The Right to Be Forgotten in India: Balancing Privacy Rights and Freedom of Expression in the Digital Era*, 9 ICONIC RSCH. & ENG'G J. 2493–97 (2026), https://doi.org/10.64388/irev9i12-1719150.

c. De-indexing of a webpage from searches based on an individual's name;
d. Anonymization of the individual in an archival publication;
e. Restriction of processing or reduced prominence; and
f. Correction or contextualization of inaccurate or incomplete material.

These remedies have different effects. Deletion suppresses the source itself. De-indexing leaves the source available to those who know its location or search by other terms, but removes the ordinary name-based route to discovery. Anonymization preserves the historical record while reducing the connection between the event and the individual's present identity. Contextualization may correct the problem without suppressing access.

The modern debate emerged from older data-protection principles: purpose limitation, storage limitation, relevance, accuracy and the prohibition of processing that becomes excessive or unnecessary. The right to be forgotten is therefore not necessarily a wholly new right. It can be understood as an application of established privacy principles to a technological environment in which data persist and remain accessible long after the purpose of publication has changed.[4]

Its moral foundation is the idea that individuals require a degree of control over the presentation of their identity. Privacy is not limited to secrecy. A person may have a legitimate interest in controlling the circumstances in which information is collected, aggregated, interpreted and made salient. A public tweet, old newspaper report or court record may individually reveal little, yet become deeply intrusive when indexed, ranked and combined with other data. The central harm may arise not from the existence of information but from its perpetual accessibility and the inferences generated by its technical organization.

This is the "digital memory" problem. Digital networks eliminate the natural decay of information. An event may remain searchable without regard to time, context or the changed character of the person involved. The internet also collapses distinctions between different audiences. Information originally directed to a local readership may become globally available; an old report intended for contemporaneous consumption may become a permanent identity marker; and material that was once difficult to locate may be placed at the top of a search-results page.

The right to be forgotten should therefore be understood as a claim to temporal and contextual integrity rather than literal oblivion. It expresses the proposition that the legal significance of information may change over time. The continued processing of accurate data may become disproportionate when the information is no longer necessary for the original purpose or when its present effects are substantially different from those reasonably contemplated at the time of publication.[5]

At the same time, privacy cannot be treated as an automatic consequence of publication. Individuals may voluntarily disclose information, participate in public affairs or become subjects of legitimate public reporting. Nor does public availability eliminate all privacy interests. Scholarly analysis of public information has shown that privacy may concern control over the manner, aggregation and inference of information, rather than merely the secrecy of isolated facts.[6] The fact that data are accessible does not conclusively answer whether their continued processing is fair.

---

[4] G. Sartor, *The Right to Be Forgotten in the Draft Data Protection Regulation*, 5 INT'L DATA PRIVACY L. 64–72 (2014), https://doi.org/10.1093/idpl/ipu030.

[5] I. Iglezakis, *The Right to Be Forgotten in the Google Spain Case (Case C-131/12): A Clear Victory for Data Protection or an Obstacle for the Internet?*, SSRN ELEC. J. (2014), https://doi.org/10.2139/ssrn.2472323.

[6] B. Rumbold & J. Wilson, *Privacy Rights and Public Information*, 27 J. POL. PHIL. 3–25 (2018), https://doi.org/10.1111/jopp.12158.

2.2 The Right to Information and Freedom of Expression

The right to information encompasses several related interests. It protects the speaker's liberty to communicate, the media's ability to investigate and report, and the audience's interest in receiving information. In constitutional democracies, access to information also enables political accountability, informed voting, scrutiny of public institutions and the development of individual personality.

Freedom of expression protects more than agreeable or socially beneficial speech. It includes criticism, historical inquiry, investigative journalism and the communication of facts that may harm reputation. A right to information is therefore not exhausted by access to official documents or current news. Historical archives, judicial decisions, public registers and journalistic databases contribute to collective memory and democratic accountability.

The public interest is not identical to public curiosity. Information about a celebrity's private life may attract attention without contributing meaningfully to public debate. Conversely, material concerning corruption, public health, abuse of office, professional misconduct or threats to public safety may retain substantial public value even when it is old. The distinction is difficult because public interest is contextual, not reducible to the number of people interested in the information.

The relationship between privacy and information is also not necessarily zero-sum. Privacy may improve the quality of public discourse by enabling individuals to participate without fear of permanent surveillance. Conversely, excessive secrecy can undermine accountability. A constitutional order must therefore protect both the individual's capacity to develop a private life and society's ability to remember, investigate and criticize.

The underlying issue is one of institutional design. The internet does not merely host speech; it determines visibility. Search engines rank, organize and connect information. Their decisions affect the practical scope of expression even when no webpage is removed. The right to information must consequently include attention to the architecture through which information is found.

3 THE LANDMARK JURISPRUDENCE

3.1 Google Spain

The central European authority *Google Spain SL v Agencia Española de Protección de Datos*[7]*, decided by the Court of Justice of the European Union in 2014, concerned* a newspaper report about a property auction connected with social-security debts. The information had been lawfully published and was not alleged to be inaccurate. The claimant nevertheless argued that searching his name produced results linking him to an old financial difficulty, causing continuing harm.

The Court held that the activity of a search engine constituted processing of personal data and that the operator could be treated as a data controller in respect of that processing. It further held that, under the then-applicable Data Protection Directive, an individual could in appropriate circumstances request the removal of links from results displayed following a search of the person's name. The ruling did not require deletion of the source material from the newspaper's archive. It concerned the association generated by the search engine between the individual's name and the underlying information.

[7] Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317 (May 13, 2014).

The judgment's significance lies in its recognition that search engines alter the practical meaning of publication. The newspaper's archive and the search-results page were not treated as functionally equivalent. Search engines made it possible to obtain a comprehensive profile of an individual through a single name query. The Court consequently focused not only on the legality of the original publication but also on the continuing effects of automated indexing and name-based accessibility.

The Court adopted a balancing approach. The data subject's interests in privacy and protection of personal data had to be weighed against the interests of internet users and the public in accessing information. Relevant considerations included the nature of the information, its sensitivity, the role played by the individual in public life, the passage of time, the accuracy and continuing relevance of the material, and the contribution of the information to a matter of public interest.[8]

Importantly, the Court did not create an unconditional entitlement to removal. It contemplated circumstances in which the public interest in access would prevail. The identity and social role of the data subject mattered. A public official or influential businessperson could reasonably be subject to greater informational scrutiny than a private individual. The age of the information mattered because the public significance of an event may diminish, although time alone does not extinguish public interest. Accuracy mattered because inaccurate or misleading material presents a stronger case for correction or removal. The nature of the harm also mattered, particularly where the result created a seriously distorted or disproportionate picture of the person.

Google Spain nevertheless attracted substantial criticism. First, critics argued that it transformed data-protection law into a mechanism for controlling truthful speech. Secondly, the judgment imposed the initial balancing responsibility on search engines, private corporations whose principal expertise is technological and commercial rather than constitutional adjudication. Thirdly, publishers might not receive effective notice or participation when a link is de-indexed. Fourthly, the process could favour wealthy or powerful applicants able to submit sophisticated requests or threaten litigation. The result might be censorship without a conventional judicial order.

The concern is not merely theoretical. The removal of a link does not always suppress information completely, but it changes visibility. Search engines are important gateways to knowledge, and de-indexing can render information practically inaccessible to ordinary users. The decision therefore creates a form of indirect speech restriction while formally preserving the source publication.

3.2 Article 17 of the GDPR

Article 17 of the General Data Protection Regulation codifies a "right to erasure", also described as the right to be forgotten. It provides for erasure in several circumstances, including where data are no longer necessary for the purposes for which they were collected or processed, where consent is withdrawn and no other legal ground exists, where the individual objects to processing in specified circumstances, or where processing has been unlawful.

The right is not absolute. Article 17 contains express exceptions for freedom of expression and information, compliance with legal obligations, public interest in public health, archiving in the public interest, scientific or historical research, statistical purposes and the establishment, exercise or defence of

[8] P. Jonason, *The Right to Be Forgotten: The Balance Between the Right to Privacy and Freedom of Expression*, 30 Revue Européenne de Droit Public/European Review of Public Law 213–26 (2018).

legal claims. These exceptions reveal that the GDPR does not establish a hierarchy in which privacy automatically defeats information. Rather, it creates a structured proportionality inquiry.

Article 17 also raises questions about the distinction between controllers and publishers. Where an individual requests erasure, the controller may have obligations to take reasonable steps to inform other controllers processing the data. Yet the Regulation leaves difficult questions about hosting services, journalistic archives, search engines and cross-border accessibility. The provision addresses erasure but only partially resolves the underlying issue of online publication and the allocation of responsibility among different intermediaries.[9]

The GDPR's architecture confirms that different remedies should be considered before deletion. If the problem is inaccuracy, correction may be sufficient. If the problem is excessive prominence, de-indexing or anonymization may be preferable. If the information remains necessary for archival or journalistic purposes, continued access may be justified subject to safeguards. The remedy should correspond to the harm.

3.3 The Risk of Privatized Adjudication

A major structural objection is that search engines have become the first-instance decision-makers in disputes involving fundamental rights. They receive the application, gather information, assess the competing interests, decide whether the result should remain visible and communicate the outcome to the applicant. Their decisions may be reviewed by data-protection authorities or courts, but review is usually subsequent, selective and costly.

This has been described as "private jurisprudence". Private actors effectively develop rules governing privacy, data protection and access to information without the procedural protections ordinarily associated with constitutional adjudication.[10] The problem is heightened by secrecy. Search engines may not disclose detailed reasons because of concerns about privacy, fraud, manipulation or commercial confidentiality. A person whose link is removed may not know why; a publisher may not know that the link has disappeared; and the public cannot easily assess the consistency of decisions.

The model also produces asymmetry. Search engines face legal risk if they refuse removal, but may face less immediate risk if they remove a link too readily. This can create incentives toward over-removal. The public's interest in access is diffuse and difficult to represent, whereas the applicant's complaint is concrete and legally actionable. The resulting system may systematically undervalue collective interests.

The answer is not necessarily to exclude intermediaries. A centralized court-only model may be too slow and expensive for the volume of requests. The better approach is to require reasoned decisions, meaningful notice, standardized criteria, independent review, publication of anonymized statistics and escalation of difficult cases to public authorities. Private processing can be administrative; it should not become the final, unreviewable determination of constitutional rights.

---

[9] G. Sartor, *The Right to Be Forgotten in the Draft Data Protection Regulation*, 5 INT'L DATA PRIVACY L. 64–72 (2014), https://doi.org/10.1093/idpl/ipu030.

[10] M. Leiser, *"Private Jurisprudence" and the Right to Be Forgotten Balancing Test*, 39 COMPUT. L. & SEC. REV. 105458 (2020), https://doi.org/10.1016/j.clsr.2020.105458.

4 Comparative Perspectives

4.1 Europe: The GDPR and Strasbourg

The European approach combines data-protection law under the GDPR with privacy and expression principles developed under the European Convention on Human Rights. The Convention's Article 8 protects private and family life, while Article 10 protects freedom of expression. The European Court of Human Rights has generally approached conflicts through proportionality, asking whether the interference is prescribed by law, pursues a legitimate aim and is necessary in a democratic society.

Strasbourg case law emphasizes factors such as the contribution of the publication to a debate of general interest, the notoriety of the person concerned, the subject of the report, the prior conduct of the person, the content, form and consequences of publication, and the severity of the sanction imposed. These criteria are especially important in cases concerning archives and searchability because the original publication may have been lawful while later accessibility creates a new privacy impact.

In the case of *Hurbain v Belgium*[11], the Grand Chamber considered the publication of an old newspaper article identifying a driver involved in a fatal road accident. The Court accepted that continued identification could seriously affect the individual's private life and professional rehabilitation. At the same time, it treated the case as involving a historical journalistic archive rather than a demand for complete deletion. Anonymisation was significant because it could reduce the continuing harm while preserving the archival record. The case illustrates the value of graduated remedies: anonymity may be more proportionate than erasure.[12]

The Strasbourg approach also places emphasis on significant harm. Mere discomfort, embarrassment or the fact that information is unflattering should not automatically justify interference with expression. The applicant should demonstrate a substantial and continuing impact, assessed in light of the nature of the information, the passage of time and the person's circumstances. This is not a requirement that harm be catastrophic, but it discourages the transformation of ordinary reputational dissatisfaction into a right of suppression.

European law therefore contains two related but not identical tendencies. The Court of Justice of the European Union's (CJEU) search-engine jurisprudence focuses on data processing and name-based accessibility. The European Court of Human Rights' (ECtHR) approach concentrates on proportionality, journalistic freedom and the consequences of the restriction. The relationship between the two bodies of law remains complex, and inconsistent standards may produce uncertainty.[13]

4.2 The United States

The United States has no general statutory right to be forgotten comparable to Article 17 GDPR. The First Amendment creates strong constitutional protection for truthful information, particularly where the information concerns public proceedings, matters of public significance or lawfully obtained facts. The Supreme Court has repeatedly been reluctant to permit government restrictions on truthful publication merely because disclosure causes distress or reputational injury.

---

[11] Hurbain v. Belgium, App. No. 57292/16 (Eur. Ct. H.R. Grand Chamber July 4, 2023), https://hudoc.echr.coe.int/eng?i=001-225814.

[12] T. Oganesian, *The Right to Be Forgotten: The European Court of Human Rights in Search of the Necessary Balance*, 12 MEŽDUNARODNOE PRAVOSUDIE 32–56 (2022), https://doi.org/10.21128/2226-2059-2022-1-32-56.

[13] *Id.*

The American position reflects a different constitutional culture. The presumption is that the remedy for harmful or contested speech is often more speech, correction, context or counter-speech rather than suppression. Defamation law generally distinguishes false statements from truthful but embarrassing facts, and public figures face demanding standards when seeking damages. A general right requiring search engines or publishers to remove truthful information would risk prior restraint and content-based regulation.

This does not mean that American law ignores privacy. Intrusion upon seclusion, public disclosure of private facts, appropriation, breach of confidence, contractual obligations and sector-specific data-protection laws may provide protection. The law may also recognize claims involving false light or misuse of private data. But these doctrines do not amount to a general power to require de-indexing of truthful historical information.

The American resistance is especially strong where removal would conceal information about public officials, criminal justice, political conduct or matters of public safety. It also reflects concern that private intermediaries could make decisions with public consequences while avoiding the constitutional constraints applicable to the state. Comparative scholarship accordingly describes the United States as lacking a general common-law right to be forgotten and as facing significant difficulty reconciling such a right with the First Amendment.[14]

Nevertheless, the contrast should not be overstated. Even in the United States, the practical design of search engines raises questions about privacy and informational power. The First Amendment does not necessarily require that every piece of information receive equal prominence forever. Voluntary platform policies, technical measures, privacy-enhancing design and narrowly tailored statutory regulation may protect individuals without authorizing broad governmental censorship.

4.3 India

India's constitutional position has evolved substantially through *Justice K.S. Puttaswamy (Retd.) v Union of India*.[15] The Supreme Court recognised privacy as a constitutionally protected right grounded in dignity, liberty, autonomy and the freedoms guaranteed by Part III of the Constitution. Privacy includes bodily, decisional and informational dimensions. Although the judgment did not settle the complete content of a right to be forgotten, its reasoning provides a constitutional foundation for protecting individuals against unjustified collection, processing and dissemination of personal data.

Indian constitutional law must also account for Article 19(1)(a), which protects freedom of speech and expression, and the recognized importance of press freedom and public access to information. Privacy under Article 21 cannot operate as an automatic injunction against publication. Restrictions on expression must satisfy constitutional standards of legality, legitimate purpose, necessity and proportionality.

Indian courts have occasionally addressed requests for removal or masking of personal information, including in relation to matrimonial disputes, criminal cases and online publication of sensitive personal material. The judicial treatment has been developing rather than uniform. Some decisions have accepted that continued online availability may cause disproportionate harm; others have stressed the importance

[14] M. Federico, *Have We Forgotten Freedom of Information? A Comparative Critical Overview of the Right to Be Forgotten in Europe and the United States*, ARCA (UNIVERSITÀ CA' FOSCARI VENEZIA) (2024).
[15] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.

of open justice, court records and the public character of proceedings. The central difficulty is to determine whether a request concerns legitimate privacy, reputational discomfort, suppression of adverse history or protection against a genuine and continuing injury.

The Digital Personal Data Protection Act 2023 marks an important legislative development. It establishes duties for data fiduciaries, rights for data principals and a framework for processing digital personal data. Its relevance to the right to be forgotten arises through rights relating to erasure or correction and through the regulation of processing. Yet the statute does not simply replicate Article 17 GDPR.[16] Its institutional structure, exemptions, relationship with public authorities and interaction with journalistic and constitutional freedoms require careful interpretation. The Act must be read consistently with constitutional values, particularly privacy, dignity, equality, expression and access to justice.

India should avoid adopting either an unlimited European-style removal power or an absolutist American model. Its social conditions make protection against doxxing, non-consensual intimate imagery, cyber harassment, identity theft and caste- or gender-based targeting especially important. At the same time, historical records, court judgments, investigative journalism and public accountability cannot be casually erased. The Indian framework should therefore prioritise graduated remedies and judicially reviewable standards.

5 The Search for a Balance

5.1 The Applicable Balancing Criteria

A principled balance should begin with the recognition that neither privacy nor information is presumptively dominant. The following criteria should guide decision-making.

5.1.1 Nature and Sensitivity of the Information

Highly sensitive personal data—medical information, sexual life, intimate images, information concerning children, psychological distress or victim status—deserve strong protection. Information relating to public functions, corporate conduct, official decisions or criminal proceedings may attract a stronger public interest. The classification is not absolute, but it establishes the starting point.

5.1.2 Public Interest Rather Than Public Curiosity

The decisive question should be whether access contributes to democratic accountability, public safety, historical understanding, professional integrity or informed public debate. Mere popularity, sensationalism or commercial value should carry limited weight. A person's celebrity may increase public interest, but it should not convert every aspect of private life into legitimate public information.[17]

5.1.3 Role of the Data Subject

Public officeholders, senior professionals exercising public trust, corporate executives and persons who deliberately place themselves at the centre of public controversy may reasonably expect greater

[16] Shubham Saurabh, *The Digital Personal Data Protection Act of 2023: Strengthening Privacy in the Digital Age*, 3 Int'l J.L. Changing World 77–94 (2024), 10.54934/ijlcw.v3i2.84.

[17] Von Hannover v. Germany, App. No. 59320/00 (Eur. Ct. H.R. June 24, 2004), https://hudoc.echr.coe.int/eng?i=001-61853.

scrutiny. Private individuals generally have stronger claims to de-indexing, especially where the information relates to an isolated event rather than a continuing public role.

5.1.4 Passage of Time and Changed Circumstances

Time is central but not dispositive. Old information may lose relevance, particularly where the individual has reformed, completed a sentence, settled a dispute or rebuilt a private life. Conversely, historical information may retain importance where it concerns continuing accountability, patterns of conduct, unresolved wrongdoing or public office. The legal inquiry should ask whether the information's present value justifies its present accessibility.

5.1.5 Accuracy, Completeness and Context

False information should ordinarily be corrected or removed through existing legal remedies. But accuracy alone does not defeat a privacy claim. True information may become misleading when presented without subsequent developments or contextual explanation. Search engines should consider whether the result creates a distorted profile by highlighting a single historical incident.

5.1.6 Seriousness and Causation of Harm

The applicant should identify a substantial, foreseeable and continuing harm. Relevant factors include employment discrimination, harassment, threats, exclusion, psychological injury, family consequences and inability to reintegrate. Harm should be connected to the contested processing rather than merely to the underlying event. The emphasis on significant harm is particularly important in European human-rights analysis.[18]

5.1.7 Availability of Less Restrictive Remedies

Removal should not be the default. Alternatives include correction, annotation, updating, reduced ranking, geo-limited de-indexing, anonymization, suppression of autocomplete suggestions, removal of duplicated copies, or limitation to searches using additional terms. The remedy should be tailored to the source of the injury.

5.2 Public Figures, Private Individuals and the Right to Information

The distinction between public and private persons is not a complete rule, but it is a valuable organising principle. Information about a public official's corruption, a doctor's disciplinary history, a company's environmental violations or a politician's prior statements may remain relevant because the person's public role creates continuing accountability. A private person involved in an isolated incident should ordinarily have a stronger claim once the event has ceased to carry public significance.[19]

However, public-figure status should not be permanent or total. A person may be publicly known in one domain but retain privacy in another. Nor should a person lose all protection merely because they once appeared in the news. The inquiry must remain tied to the particular information and the particular context.

---

[18] T. Oganesian, *The Right to Be Forgotten: The European Court of Human Rights in Search of the Necessary Balance*, 12 MEŽDUNARODNOE PRAVOSUDIE 32–56 (2022), https://doi.org/10.21128/2226-2059-2022-1-32-56.

[19] Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317, ¶ 97 (May 13, 2014).

The right to information also belongs to future users, not only present readers. Journalists, researchers, historians, lawyers and members of the public may have legitimate reasons to consult archives. De-indexing can impair access even where the original source remains technically available. A balance must therefore distinguish between preserving the record and maintaining unlimited discoverability through a person's name.

5.3 Courts, Regulators and Private Intermediaries

The most difficult institutional question is who should perform the balancing exercise. Courts offer independence, procedural safeguards, public reasons and authoritative interpretation. But judicial adjudication of every request is unrealistic. Data-protection authorities can develop expertise and issue guidance, yet may be slow or under-resourced. Search engines can process large volumes and respond rapidly, but they lack democratic legitimacy and may not adequately represent the public interest.[20]

A layered system is preferable:

1. Initial administrative assessment by the intermediary: Using published criteria and a presumption against removal where there is a strong public-interest basis.
2. Notice and reasoned decision: Subject to privacy safeguards and protection against retaliation.
3. Independent review: By a data-protection authority or specialized tribunal for contested or significant cases.
4. Judicial review: Particularly where the matter concerns public officials, journalism, court records, political speech or a novel legal issue.
5. Periodic reconsideration: Because relevance and harm may change over time.

The process should include mechanisms for notifying publishers where appropriate, although notification must not expose applicants to further harm. It should also permit participation by public-interest representatives in cases involving journalism, archives or public records.

The goal is not to eliminate private decision-making but to prevent it from becoming private constitutional law. Search engines should administer a public framework, not invent one in secret.[21] Their decisions should be auditable through anonymised transparency reports, consistent terminology, reason-giving and external oversight.

6 Challenges and Open Questions

6.1 Censorship and Historical Revision

The strongest objection to the right to be forgotten is that it can become a mechanism for censorship. Powerful individuals may seek to conceal misconduct, political actors may attempt to rewrite their histories, and corporations may use privacy claims to suppress investigative reporting.[22] If removal decisions are secret, the public may never know what has disappeared.

---

[20] M. Leiser, *"Private Jurisprudence" and the Right to Be Forgotten Balancing Test*, 39 Comput. L. & Sec. Rev. 105458 (2020), https://doi.org/10.1016/j.clsr.2020.105458.

[21] Didier Liambomba, *The Right of Access to Public Information: Human Rights Issues, Transparency and Good Governance*, 4 Constitutionale 1–28 (2023), 10.25041/constitutionale.v4i1.2601.

[22] P. Amponsah, Investigative Reporting: Its Practice, Boundaries and Implications (2024) (ISBN: 978-9988-9369-0-7).

This danger is especially acute where the information is accurate and concerns public conduct. A legal regime that treats reputational injury as sufficient would convert privacy into a veto over criticism. The law should therefore distinguish dignity-protective remedies from reputation management. The right to be forgotten should not be a general right to curate one's public image.

6.2 Global Enforcement and Territoriality

The internet is transnational, but privacy and speech laws remain territorially organised. A European order may require de-indexing within the European Union, while another jurisdiction may prohibit removal. Global application of a removal order risks exporting one jurisdiction's speech standards to the world. Territorial limitation, however, may be ineffective where users can access foreign versions of a search engine or use technical circumvention.[23]

The CJEU has wrestled with the territorial scope of de-indexing, particularly in cases concerning whether removal should extend beyond European domains. The problem reveals a broader tension between effective privacy protection and respect for constitutional pluralism. A sensible approach should begin with territoriality, permit broader measures only where necessary and proportionate, and avoid requiring platforms to impose one state's balance globally.

6.3 The Right to Information as a Counterweight

The public's right to know is often diffuse and underrepresented in individual removal procedures. A person seeking removal is identifiable and motivated; the journalist, researcher or future citizen who may need the information is unknown. This structural asymmetry favours privacy claims.

Public-interest representation is therefore necessary. Data-protection authorities should issue guidance on categories of information presumptively linked to public interest, including corruption, serious professional misconduct, public-health threats, abuse of public office and threats to vulnerable groups. Courts should be cautious when removal would impair access to judicial records, historical archives or investigative journalism.[24]

The existence of a public record should not automatically defeat privacy. Open justice has limits, and the continued online indexing of a judgment may impose a burden far greater than access to the physical file. But the solution should often be anonymization, redaction or controlled access rather than disappearance.

6.4 Technological Feasibility

Technical systems cannot reliably distinguish public interest from public curiosity. Automated tools may identify names, dates and categories, but proportionality requires contextual judgment. Machine-learning systems may reproduce bias, over-remove controversial content or under-protect marginalized persons. They may also fail to recognize cultural and linguistic differences.

Search engines must additionally manage duplicate webpages, mirrors, cached content, social-media re-posts and alternative search terms. A single de-indexing order may be ineffective if information is

[23] G. Gregorio, *The Transnational Dimension of Data Protection* (2024), https://doi.org/10.1163/27725650-01020006.

[24] S. SINGH, STUDYING THE INTRICACIES OF PRIVACY AND DATA PROTECTION IN INDIA – A CRITICAL ANALYSIS OF THE INDIAN LAWS 118.

replicated across platforms. Conversely, overly broad technical enforcement may suppress unrelated information sharing the same name.

Technological design should therefore support, rather than replace, legal judgment. Systems should provide graduated controls, record decision rationales, permit human escalation and preserve evidence for review. Privacy by design should include minimization, purpose limitation and mechanisms preventing unnecessary repetition of sensitive information.

6.5 Inconsistent Standards

The absence of uniform criteria creates uncertainty for applicants, publishers and platforms. European institutions themselves have developed overlapping but not identical approaches. The CJEU data-protection reasoning and the ECtHR's proportionality analysis may point in different directions. India is developing its constitutional and statutory position, while the United States remains resistant to a general removal right.

Some inconsistency is inevitable because privacy and expression are culturally and constitutionally situated. The aim should not be complete uniformity but principled compatibility. International standards could establish minimum procedural requirements: reasoned decisions, notice, independent review, protection for journalism, consideration of public interest, preference for less restrictive remedies and transparency concerning aggregate outcomes.

6.6 Identity, Equality and Social Reintegration

The right to be forgotten has an important equality dimension. Persistent digital records may disproportionately burden persons from disadvantaged groups, defendants who lacked resources to respond to allegations, victims of abuse, former offenders seeking reintegration and individuals targeted by online harassment. A purely formal commitment to equal access to information may reproduce unequal social consequences because some individuals bear far greater risks from permanent visibility than others.

At the same time, privacy remedies must not privilege elites. If only prominent or wealthy applicants can secure de-indexing, the system will become a tool of reputational control. Accessible procedures, legal aid, transparent criteria and regulatory oversight are essential.[25]

7 Conclusion

The conflict between the right to be forgotten and the right to information is not a contest between a modern privacy right and an outdated conception of free expression. Both respond to genuine features of digital life. Individuals require protection from permanent, contextless and disproportionate exposure. Democratic societies require the preservation and circulation of information, including information that challenges reputations and unsettles the powerful.

The correct legal approach rejects hierarchy. Privacy should not automatically defeat expression, and expression should not automatically defeat privacy. The balance must be contextual and proportionate. It should consider the nature and sensitivity of the information, the public role of the individual, the

[25] M. Leiser, *"Private Jurisprudence" and the Right to Be Forgotten Balancing Test*, 39 Comput. L. & Sec. Rev. 105458 (2020), https://doi.org/10.1016/j.clsr.2020.105458.

contribution of the information to public debate, the passage of time, accuracy, context, the seriousness of harm and the availability of less restrictive remedies.

Five principles should guide the digital balance.

First, the law should distinguish source deletion from search de-indexing, anonymization, correction and contextualization. The least restrictive effective remedy should be preferred.

Secondly, public interest should be distinguished from public curiosity. Information concerning public accountability, safety, corruption, professional integrity and historical understanding should receive strong protection.

Thirdly, private individuals should generally receive greater protection than public officials or persons who voluntarily assume public-facing roles, though no category should be absolute.

Fourthly, decision-making should be transparent, reasoned and independently reviewable. Search engines may perform an initial administrative function, but they should not become the final arbiters of fundamental rights. The problem of private jurisprudence must be addressed through public standards, regulatory supervision and judicial review.[26]

Fifthly, remedies should be temporally sensitive. A refusal to remove information today should not foreclose reconsideration when circumstances change. Conversely, a decision to de-index should not prevent legitimate archival or public-interest access through appropriate channels.

The European experience demonstrates both the necessity and the difficulty of recognizing informational self-determination. The American experience warns against allowing truthful information to be suppressed through broad governmental or private powers. The Indian constitutional framework offers an opportunity to develop a middle path grounded in dignity, autonomy, expression, open justice and proportionality. The Digital Personal Data Protection Act 2023 can contribute to that project if interpreted as part of a constitutional settlement rather than as a self-contained erasure mechanism.

The digital balance will never be achieved by a single rule. It requires institutions capable of distinguishing harm from discomfort, public interest from curiosity, historical preservation from perpetual exposure and privacy protection from censorship. The right to be forgotten should therefore be understood not as a right to erase history, but as a legal claim to prevent the unjustified transformation of a person's past into a permanent and disproportionate feature of their present. The right to information, correspondingly, should not mean that every fact must remain equally prominent forever. The enduring task is to preserve both human dignity and democratic memory.

* * *

[26] *Id.*

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.21

CHAPTER 21

Samvaad: Digital Constitutionalism, Article 19 and the Future of Online Dissent

ANOUSHKA GIRISH[*] AND APOORVA T[**]

Digital constitutionalism has emerged as a framework for examining how constitutional values are preserved in an increasingly platform-driven society. In India, where political discourse has progressively shifted from physical spaces to digital platforms, the constitutional guarantee of freedom of speech and expression under Article 19(1)(a) faces new challenges. While digital media has expanded democratic participation, it has also intensified concerns relating to algorithmic amplification, platform moderation, misinformation, surveillance, internet shutdowns, and criminal regulation.

This paper examines the evolution of Article 19 within digital constitutionalism, focusing on student-led protests and online mobilisation. Using the Delhi student rallies as a contemporary case study, it analyses how digital platforms have transformed dissent by enabling rapid dissemination of information, wider participation, and collective action beyond geographical boundaries. It evaluates whether digital activism strengthens constitutional democracy or creates risks through misinformation, polarisation, digital vigilantism, and targeted attacks on individual dignity. The paper further examines the reasonable restrictions under Article 19(2) and their application to online expression through the principles of legality, necessity, and proportionality. It considers the relationship between Article 19 and the dignity protected under Article 21, particularly where harmful content is rapidly reproduced and accountability becomes difficult. It also examines the growing influence of private platforms and algorithmic systems over political expression. The paper argues that digital constitutionalism must balance freedom, dignity, technological power, and constitutional accountability to preserve meaningful democratic dissent in India's digital age.

1 INTRODUCTION

The public square is no longer confined to streets, university campuses, newspapers, or public meetings. It increasingly exists on screens. Social media platforms have transformed the way individuals communicate, organise political movements, express dissent, and participate in democratic life. A protest that once depended primarily upon physical gathering can now begin with a post, spread through hashtags, and reach thousands within minutes.[1] For student movements in particular, digital platforms have become important spaces for mobilisation, documentation, and political expression. This transformation raises a fundamental constitutional question: how far does the protection of freedom of speech and expression under Article 19(1)(a) extend into the digital public sphere?[2] The Constitution was framed in an era of conventional communication, yet contemporary political discourse is increasingly shaped by social media platforms, algorithms, private content moderation, and instant communication. Online expression can

[*] Student at CMR University, School of Legal Studies, Bengaluru, Karnataka, India.
[**] Assistant Professor at CMR University, School of Legal Studies, Bengaluru, Karnataka, India.

[1] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

[2] INDIA CONST. art. 19, cl. 1(a).

be amplified to enormous audiences, restricted by governmental action, removed by private platforms, or rendered inaccessible through internet shutdowns.[3] At the same time, digital spaces create genuine concerns relating to misinformation, public order, surveillance, privacy, and individual dignity.

The concept of digital constitutionalism provides a framework for examining these developments.[4] It considers how constitutional values such as liberty, dignity, accountability, privacy, and freedom of expression can be preserved when public discourse is increasingly mediated by digital technologies and private platforms. The constitutional question is consequently no longer limited to whether the State restricts speech. It also concerns who controls the digital spaces in which speech takes place and how technological systems influence what individuals can see, express, and circulate. The Indian judiciary has already addressed several dimensions of these concerns. In *Shreya Singhal v. Union of India* (2015), the Supreme Court struck down Section 66A of the Information Technology Act, 2000, reaffirming the constitutional protection of online expression.[5] *K.S. Puttaswamy v. Union of India*[6] recognised privacy as a fundamental right and linked it with autonomy and dignity, while *Anuradha Bhasin v. Union of India*[7] recognised the internet as an important medium through which constitutionally protected freedoms may be exercised. These decisions demonstrate that fundamental rights cannot become weaker merely because their exercise has moved into digital spaces.

The constitutional difficulty becomes particularly visible in the context of student-led dissent. Digital platforms allow students to organise rapidly, document protests, circulate information, and reach audiences beyond geographical boundaries. However, the same platforms can facilitate misinformation, digital vigilantism, harassment, and targeted attacks upon individuals. During the Delhi student rallies, online posts and stories depicting injured students became part of the wider public narrative surrounding the protests. At the same time, instances of provocative and degrading content directed at public figures demonstrated the difficulty of distinguishing protected political expression from expression that may affect another person's dignity and reputation.

This creates a tension between Articles 19 and 21. While Article 19(1)(a) protects political expression and dissent, Article 19(2) permits reasonable restrictions on specified grounds, including public order, security of the State, and incitement to an offence.[8] Article 21, meanwhile, protects life and personal liberty and has been interpreted to include dignity.[9] The constitutional challenge is therefore not whether online expression should be absolutely free, but where the boundary lies between legitimate dissent and unjustifiable harm.

This paper examines that boundary through the framework of digital constitutionalism, focusing particularly on student-led online mobilisation and the Delhi student rallies. It evaluates the constitutional limitations imposed under Article 19(2), the relationship between online expression and the dignity protected under Article 21, and the role of judicial review in determining whether digital restrictions are lawful, necessary, and proportionate.[10] It further examines the growing influence of private platforms and

[3] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.
[4] Nicolas P. Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, 4 SOC. MEDIA + SOC'Y 1 (2018).
[5] Shreya Singhal v. Union of India, (2015) 5 SCC 1.
[6] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[7] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.
[8] INDIA CONST. art. 19, cls. 1(a), 2.
[9] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[10] Modern Dental Coll. & Rsch. Ctr. v. State of Madhya Pradesh, (2016) 7 SCC 353.

algorithmic systems over the visibility of political speech. The paper argues that protecting online dissent requires more than simply extending Article 19(1)(a) to the internet. It requires a constitutional framework capable of addressing State restrictions, private platform power, algorithmic amplification, misinformation, and individual dignity simultaneously. Digital constitutionalism must therefore ensure that technological power does not become a substitute for constitutional accountability, while preserving the digital public sphere as a meaningful space for democratic participation and dissent.

2 DIGITAL CONSTITUTIONALISM AND ONLINE DISSENT

Digital constitutionalism examines how constitutional principles operate within an increasingly digital society. Unlike traditional constitutionalism, which largely focuses on the relationship between the State and the individual, the digital environment involves a wider distribution of power. Technology companies, social media platforms, algorithms, and other private actors increasingly influence how individuals communicate, access information, and participate in public life.[11] The constitutional question therefore extends beyond whether the State restricts a right to whether the digital environment itself enables or limits its meaningful exercise.

Freedom of expression occupies a central position within this framework. Digital platforms have become important spaces for political discussion, criticism of government policies, dissemination of information, and collective mobilisation. A single post can reach audiences far beyond geographical and institutional boundaries, allowing individuals and groups with limited access to traditional media to participate directly in public discourse. For student movements, this has made digital platforms particularly significant for organising protests, sharing experiences, and building public support.

However, digital spaces are not neutral or unrestricted public forums. The visibility of content can be influenced by algorithms, platform policies, moderation practices, and commercial interests. Content may be removed, restricted, or made less visible by private platforms even without direct State intervention. This creates a distinction between the formal existence of a right to expression and the practical ability to exercise it: an individual may be free to speak, while technological systems influence whether that speech is actually heard.

This tension becomes particularly important in relation to dissent. The ability to criticise public authorities, question policies, organise demonstrations, and express disagreement is fundamental to constitutional democracy. Digital platforms have transformed dissent by allowing movements to move rapidly between physical and online spaces. A student protest can begin with a physical gathering and acquire national visibility through videos, posts, hashtags, and online campaigns. At the same time, these platforms can facilitate misinformation, manipulation, harassment, and digital vigilantism. Governments may invoke public order, security, or other concerns to regulate online mobilisation, while private platforms apply community standards that may differ from constitutional standards governing State action. Online expression is therefore subject to a layered system of constitutional, statutory, governmental, and private regulation.

Digital constitutionalism must consequently address both State restrictions and private technological power. Its objective is not to create a digital space free from regulation, but to ensure that regulation

[11] Nicolas P. Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, 4 SOC. MEDIA + SOC'Y 1, 1–2 (2018).

remains consistent with constitutional values of liberty, dignity, accountability, and democratic participation. This is particularly relevant to student-led dissent, where the internet has become both a powerful instrument of mobilisation and a new site of constitutional conflict.

3 ARTICLE 19(1)(A) IN THE DIGITAL PUBLIC SPHERE

Article 19(1)(a) of the Constitution guarantees every citizen the freedom of speech and expression, a right central to democratic participation. Although the Constitution does not expressly refer to the internet, constitutional rights cannot be confined to the technologies existing at the time of its adoption. As methods of communication evolve, the protection of expression must extend to the mediums through which individuals actually participate in public discourse.[12]

The internet has significantly expanded the practical dimensions of this freedom. Social media allows individuals to express opinions, access information, criticise governmental action, and communicate with audiences beyond geographical and institutional boundaries. Unlike traditional media, digital platforms allow individuals to participate directly in public discourse without requiring access to newspapers, television, or other established institutions. This has made online spaces particularly important for political expression and collective mobilisation.[13]

The constitutional protection of online speech was significantly strengthened by *Shreya Singhal v. Union of India*,[14] where the Supreme Court struck down Section 66A of the Information Technology Act, 2000. The Court's distinction between discussion, advocacy, and incitement remains particularly relevant to digital dissent. Speech cannot be restricted merely because it is unpopular, offensive, or controversial;[15] constitutional restriction requires a connection with the grounds recognised under Article 19(2).[16]

This principle is especially important where digital platforms are used for political and student-led dissent. Criticism of government policy, calls for peaceful mobilisation, and expression of solidarity with a protest movement form an important part of democratic participation. The fact that such expression may generate controversy or reach a large audience does not, by itself, justify its restriction. At the same time, digital communication has altered the scale and speed at which speech operates. A single post can reach thousands within minutes, while online mobilisation can rapidly move between digital and physical spaces. This creates legitimate concerns relating to public order, misinformation, and unlawful conduct, but the scale of digital communication cannot itself become a justification for restricting constitutionally protected expression.[17]

The constitutional significance of online expression also intersects with other fundamental rights. In *K.S. Puttaswamy v. Union of India*, the Supreme Court recognised privacy as a fundamental right, linking it

---

[12] YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 213–16 (2006).

[13] *See generally* Sec'y, Ministry of Info. & Broad., Gov't of India v. Cricket Ass'n of Bengal, (1995) 2 SCC 161 (recognising the importance of freedom of speech and expression in the dissemination and receipt of information).

[14] Shreya Singhal v. Union of India, (2015) 5 SCC 1.

[15] *Id.* at 20–25. The Supreme Court distinguished between discussion, advocacy and incitement and held that only when discussion or advocacy reaches the level of incitement can it attract restriction under Article 19(2).

[16] INDIA CONST. art. 19, cl. 2.

[17] *Id.* art. 19(2). Restrictions upon freedom of speech and expression must fall within one or more of the constitutionally recognised grounds under Article 19(2).

with autonomy and dignity.[18] In *Anuradha Bhasin v. Union of India*,[19] the Court recognised the internet as an important medium through which constitutionally protected freedoms are exercised. Together, these decisions demonstrate that fundamental rights must remain meaningful within the digital environment.

The central issue, therefore, is not whether Article 19(1)(a) applies online, but how its protection should operate when digital speech is simultaneously shaped by State restrictions, private platforms, and technological systems. This question becomes particularly significant when online expression forms part of collective student dissent.[20]

4 Constitutional Restrictions on Freedom of Expression under Article 19(2)

The freedom guaranteed under Article 19(1)(a) is not absolute. Article 19(2) permits the State to impose reasonable restrictions on freedom of speech and expression, but only on the grounds expressly recognised by the Constitution. These include the sovereignty and integrity of India, security of the State, friendly relations with foreign States, public order, decency or morality, contempt of court, defamation, and incitement to an offence.[21] The existence of a recognised ground, however, does not automatically validate a restriction. It must also satisfy constitutional requirements of reasonableness, necessity, and proportionality.[22]

The digital environment makes this balance particularly significant. Online speech can reach large audiences within minutes and can move rapidly between platforms and physical spaces. However, the speed or reach of online communication cannot, by itself, justify restricting protected expression. The State must demonstrate a sufficient connection between the expression and the specific harm recognised under Article 19(2).[23]

4.1 Sovereignty and Integrity of India

This ground permits restrictions on expression that genuinely threatens the sovereignty and territorial or constitutional integrity of India. Political criticism, advocacy for governmental change, or discussion of regional issues cannot automatically be treated as threats to national integrity. In the digital sphere, where political content can reach audiences across borders, constitutional scrutiny remains necessary to distinguish legitimate political dissent from expression that genuinely threatens national integrity.[24]

4.2 Security of the State

Security of the State concerns serious threats to the functioning and survival of the State, rather than ordinary breaches of public order.[25] Mere criticism of government policy or politically unpopular expression cannot ordinarily be placed within this category. Where digital communication is used to facilitate serious violence or unlawful activity, restriction may be justified; however, the State must establish a

[18] Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1.
[19] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.
[20] INDIA CONST. art. 19(2).
[21] State of Madras v. V.G. Row, AIR 1952 SC 196; INDIA CONST. art. 19(2).
[22] Superintendent, Cent. Prison v. Ram Manohar Lohia, AIR 1960 SC 633.
[23] Kedar Nath Singh v. State of Bihar, AIR 1962 SC 955; *see also* Balwant Singh v. State of Punjab, (1995) 3 SCC 214.
[24] Romesh Thappar v. State of Madras, AIR 1950 SC 124.
[25] Kedar Nath Singh v. State of Bihar, AIR 1962 SC 955.

genuine connection between the expression and the threat rather than relying on speculative possibilities.[26]

4.3 Friendly Relations with Foreign States

The Constitution also permits restrictions in the interests of friendly relations with foreign States. The borderless nature of digital communication makes this ground increasingly relevant, as online statements may reach international audiences.[27] Nevertheless, the existence of an international audience cannot itself justify restricting political expression. The restriction must remain connected to the constitutionally protected interest.[28]

4.4 Public Order

Public order is particularly significant in the context of student protests and online mobilisation. A social media post supporting a peaceful protest or criticising governmental action does not automatically threaten public order merely because it attracts a large audience or provokes disagreement. There must be a proximate connection between the expression and the anticipated disturbance.[29]

The speed at which digital content spreads may increase the potential consequences of expression, but it cannot replace this constitutional requirement. Otherwise, the very characteristics that make digital platforms valuable for democratic participation could become a justification for suppressing dissent.[30]

4.5 Decency or Morality

The grounds of decency and morality permit restrictions on certain forms of legally objectionable expression. However, concepts of morality cannot become unrestricted tools for suppressing ideas merely because they offend prevailing social attitudes.[31] Restrictions must remain consistent with constitutional values and individual liberty. The Supreme Court's jurisprudence on obscenity has also recognised the importance of considering contemporary community standards rather than treating every form of offensive or unpopular expression as legally objectionable.[32]

4.6 Contempt of Court

Expression may also be restricted where it amounts to contempt of court. Digital platforms have significantly expanded public commentary on judicial decisions and proceedings, making this limitation increasingly relevant. At the same time, legitimate criticism of judgments and judicial reasoning remains an important aspect of public discourse. The distinction between criticism and legally punishable contempt must therefore be maintained.[33]

---

[26] INDIA CONST. art. 19(2).
[27] State of Madras v. V.G. Row, AIR 1952 SC 196.
[28] S. Rangarajan v. P. Jagjivan Ram, (1989) 2 SCC 574.
[29] Superintendent, Cent. Prison v. Ram Manohar Lohia, AIR 1960 SC 633.
[30] Shreya Singhal v. Union of India, (2015) 5 SCC 1.
[31] Navtej Singh Johar v. Union of India, (2018) 10 SCC 1.
[32] Aveek Sarkar v. State of West Bengal, (2014) 4 SCC 257.
[33] Contempt of Courts Act, 1971, §§ 5, 13.

4.7 Defamation

Defamation protects reputation against unlawful attacks, an issue that has become more significant in the digital environment because false statements can be disseminated almost instantaneously and remain accessible through reposts and screenshots. However, protection of reputation cannot become a blanket restriction on criticism. Political speech concerning public officials and public institutions must continue to receive meaningful constitutional protection.[34] Any restriction must therefore remain within the limits prescribed by Article 19(2).[35]

4.8 Incitement to an Offence

Incitement represents an important boundary between protected expression and unlawful conduct. Digital platforms can rapidly disseminate calls encouraging violence or other offences, making this ground particularly relevant to online communication. However, discussion, advocacy, and incitement must remain distinct. Expression cannot be restricted merely because it is provocative or unpopular; the nature of the expression and its connection with unlawful conduct must be considered.[36]

Thus, Article 19(2) does not create a general power to control undesirable online speech. It establishes specific constitutional boundaries within which restrictions may operate. In the context of student dissent, the central question is therefore whether the State can demonstrate that a particular expression falls within one of these grounds and whether the restriction imposed is reasonable, necessary, and proportionate.[37]

5 The Requirement of Reasonableness, Necessity and Proportionality

The enumeration of grounds under Article 19(2) does not provide the State with an unrestricted power to regulate speech. A restriction must also be reasonable. Constitutional review therefore requires an examination of the relationship between the legitimate objective and the measure adopted to achieve it.[38]

The principle of proportionality has become increasingly important in this assessment. A restriction should pursue a legitimate objective, be rationally connected to that objective, be necessary in the circumstances, and avoid imposing a disproportionate burden upon the fundamental right.[39] In the digital context, proportionality assumes particular significance because restrictions can affect not only the individual whose speech is targeted but also audiences, movements, and entire online communities.

For instance, removing an individual post may have a different constitutional impact from blocking an entire account, restricting a hashtag, shutting down internet access within a region, or preventing access to an entire platform. Although each measure may be justified in particular circumstances, the constitutional analysis must consider whether the extent of the restriction corresponds to the actual threat.[40]

The question, therefore, is not whether the State possesses the power to regulate online expression. It undoubtedly possesses constitutionally recognised powers to do so. The more difficult question is

---

[34] INDIA CONST. art. 19(1)(a), 19(2); Subramanian Swamy v. Union of India, (2016) 7 SCC 221.
[35] INDIA CONST. art. 19(2).
[36] Shreya Singhal v. Union of India, (2015) 5 SCC 1.
[37] Modern Dental Coll. & Rsch. Ctr. v. State of Madhya Pradesh, (2016) 7 SCC 353.
[38] State of Madras v. V.G. Row, AIR 1952 SC 196.
[39] Modern Dental Coll. & Rsch. Ctr. v. State of Madhya Pradesh, (2016) 7 SCC 353.
[40] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.

where the constitutional boundary lies between legitimate regulation and suppression of dissent. This boundary becomes particularly important when restrictions affect student movements, political mobilisation, and collective expression.

The Delhi student rallies provide a useful context for examining this tension. They demonstrate how digital platforms can operate simultaneously as instruments of mobilisation and as spaces subject to governmental and private regulation. The following section therefore examines the role of digital platforms in facilitating student-led dissent and considers whether restrictions imposed upon such mobilisation can withstand constitutional scrutiny under Article 19(2).[41]

6 DIGITAL VIRALITY, DIGNITY AND THE LIMITS OF ONLINE DISSENT

The constitutional tension becomes particularly complex when digital expression is directed at the dignity of an identifiable individual. During the Delhi student protests, alongside posts documenting injuries and expressing political disagreement, there were also provocative posts directed at public figures. One reported post carried an "RIP" message concerning a serving Union Minister, falsely suggesting his death. Even if presented as political rhetoric or satire, such an expression raises a question beyond Article 19(1)(a): does the protection of political expression extend to deliberately portraying a living person as dead?

This question must also be examined through Article 21. The Supreme Court has recognised dignity as an essential component of the right to life and personal liberty.[42] A public official cannot be insulated from criticism merely because the criticism is harsh or offensive. However, there is a distinction between criticising a Minister's policies or performance and publishing an "RIP" message for a person who is alive. The former concerns public conduct; the latter can become a deliberate form of humiliation directed at the individual. The constitutional challenge is therefore to distinguish criticism of the office-holder from dehumanisation of the individual.

The problem becomes more complicated because digital harm rarely remains confined to the original speaker. A post can be reposted, screenshotted, shared in groups, uploaded on another platform, and reproduced by hundreds or thousands of users. As dissemination increases, the original source becomes harder to identify while the audience becomes increasingly difficult to contain. Even when the original post is deleted, screenshots and subsequent copies may continue circulating.[43]

This creates a significant enforcement dilemma. Punishing every person who reproduced the content could result in an excessively broad restriction, particularly because individuals may differ in their intention, knowledge, and degree of participation. Conversely, focusing exclusively on the original creator may fail to address the harm caused by deliberate amplification. Digital communication therefore creates a gap between legal responsibility and technological reach. The scale of dissemination can also intensify the injury to dignity. A degrading statement communicated privately and the same statement circulated before thousands cannot necessarily produce equivalent consequences. Digital content can be repeatedly reproduced, making the harm extend beyond the moment of its original publication.

---

[41] *Id.*

[42] Francis Coralie Mullin v. Adm'r, Union Territory of Delhi, (1981) 1 SCC 608.

[43] HELEN NISSENBAUM, PRIVACY IN CONTEXT: TECHNOLOGY, POLICY, AND THE INTEGRITY OF SOCIAL LIFE 1–12 (2010).

This raises the question of what constitutes a reasonable restriction under Article 19(2). The State may legitimately address expression that causes legally recognised harm, but restrictions cannot be so broad that legitimate political criticism becomes punishable merely because it is offensive. Proportionality therefore requires consideration of the content, context, intention, target, reach, and likely consequences of the expression.[44] The example also demonstrates why Article 21 cannot become a blanket justification for suppressing dissent. Students must remain free to criticise governmental policies, organise protests, and express anger or disagreement. At the same time, constitutional freedom does not operate in isolation from the rights and dignity of others.[45] The Delhi student rallies therefore reveal a central tension within digital constitutionalism: the same technology that amplifies legitimate dissent can amplify cruelty and harm. Article 19(1)(a) protects the speaker, Article 19(2) defines constitutionally recognised limitations, and Article 21 protects the dignity of the individual affected. The challenge is to preserve all three without allowing either unrestricted digital harm or excessive State control to dominate the online public sphere.[46]

7 JUDICIAL RESPONSES TO DIGITAL CIVIL LIBERTIES IN INDIA

The Indian judiciary has increasingly adapted constitutional principles to the digital environment. The decisions in *Shreya Singhal v. Union of India*, *K.S. Puttaswamy v. Union of India*, and *Anuradha Bhasin v. Union of India* collectively demonstrate that technological change cannot diminish the protection of fundamental rights.[47]

In *Shreya Singhal*, the Supreme Court struck down Section 66A of the Information Technology Act, 2000, holding that vague and overbroad restrictions on online expression could not withstand constitutional scrutiny. The Court's distinction between discussion, advocacy, and incitement remains particularly important to digital dissent, reaffirming that speech cannot be restricted merely because it is unpopular, offensive, or inconvenient.[48]

Puttaswamy expanded the constitutional framework beyond expression by recognising privacy, autonomy, and dignity as fundamental rights. In a digital environment, surveillance and the collection of personal information may discourage individuals from participating in protests or expressing controversial opinions, creating a chilling effect even without a direct restriction on speech.[49]

In *Anuradha Bhasin*, the Court recognised that freedom of speech and expression can be exercised through the internet. Although it did not declare internet access itself to be a separate fundamental right, it recognised the internet as an important medium for exercising existing constitutional freedoms and emphasised the need for constitutional scrutiny of restrictions on access.[50]

Together, these judgments establish that fundamental rights follow the individual into the digital sphere. However, they also expose a limitation in the traditional constitutional framework. Article 19

[44] Om Kumar v. Union of India, (2001) 2 SCC 386.
[45] Kaushal Kishor v. State of Uttar Pradesh, (2023) 4 SCC 1.
[46] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 1–3 (2004).
[47] Shreya Singhal v. Union of India, (2015) 5 SCC 1; Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1; Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.
[48] GAUTAM BHATIA, OFFEND, SHOCK, OR DISTURB: FREE SPEECH UNDER THE INDIAN CONSTITUTION 1–15 (2016).
[49] NEIL M. RICHARDS, INTELLECTUAL PRIVACY: RETHINKING CIVIL LIBERTIES IN THE DIGITAL AGE 1–20 (2015).
[50] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.

primarily regulates State restrictions, while contemporary digital speech is increasingly controlled by private platforms through content moderation, account suspension, and algorithmic visibility.[51]

This creates the next constitutional challenge: when private platforms become central spaces for democratic participation, how should their power over political expression be reconciled with constitutional values of liberty, dignity, transparency, and accountability?[52]

8 PLATFORM GOVERNANCE, ALGORITHMIC AMPLIFICATION AND PRIVATE POWER

The constitutional analysis of digital dissent cannot end with State restrictions. Much of contemporary political discourse occurs on privately owned platforms that influence what speech is visible, how widely it circulates, and which narratives receive public attention. Through content moderation, recommendation systems, account restrictions, and algorithms, platforms exercise significant control over the practical conditions in which freedom of expression is exercised.[53]

This becomes particularly important during student protests. Social media can enable students to circulate information, document injuries, organise demonstrations, and reach audiences beyond geographical boundaries. However, algorithmic systems may also amplify content that generates strong emotional engagement, allowing inflammatory or misleading material to spread rapidly. The same technological architecture that makes legitimate dissent visible can therefore magnify misinformation and digital harm.[54]

Content moderation creates a further constitutional difficulty. Private platforms may remove material under their community standards even where the expression does not fall within the grounds recognised under Article 19(2). Conversely, legally permissible political expression may still be restricted by private platform rules. This demonstrates the distinction between State censorship and private moderation. Article 19(2) primarily governs State restrictions, while private platforms operate through their contractual relationships with users. Nevertheless, when these platforms become essential spaces for political participation, their decisions acquire broader democratic significance.[55]

The problem is particularly visible in the circulation of harmful content. As demonstrated by the "RIP" post concerning a living public official, content may be screenshotted, reposted, shared across groups, and reproduced on different platforms even after the original post is removed. Legal responsibility therefore becomes difficult to determine. The creator, deliberate amplifier, and person who unknowingly forwards content may not possess the same intention or degree of responsibility.[56]

Digital constitutionalism must consequently examine not only who has the formal power to restrict speech, but who has the practical power to determine whether speech is heard. This does not require treating private platforms as identical to the State. Rather, where platforms function as major spaces of

---

[51] Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C.D. L. REV. 1149, 1150–55 (2018).

[52] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 1–23 (2018).

[53] JOSÉ VAN DIJCK, THOMAS POELL & MARTIJN DE WAAL, THE PLATFORM SOCIETY: PUBLIC VALUES IN A CONNECTIVE WORLD 1–10 (2018).

[54] SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER 1–11 (2019).

[55] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1600–10 (2018).

[56] Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 65–75 (2009).

democratic participation, greater transparency in content moderation, accessible review mechanisms, and accountability in algorithmic decision-making become necessary.[57]

The future of online dissent therefore requires safeguards against both excessive State intervention and arbitrary private control, ensuring that the digital public sphere remains consistent with liberty, dignity, transparency, accountability, and democratic participation.[58]

9 Balancing Online Dissent with Public Order and Digital Regulation

The regulation of online dissent requires a balance between the protection of political expression and the State's responsibility to maintain public order and protect the rights of others. This tension is particularly visible during student protests, where social media may be used to organise peaceful demonstrations, document events, communicate demands, and mobilise public support. Such activities ordinarily fall within Article 19(1)(a), and criticism of government action cannot, by itself, be treated as a threat to public order.[59]

At the same time, digital platforms can facilitate calls for violence, threats, or deliberately false information capable of producing immediate disorder. Article 19(2) therefore permits reasonable restrictions where constitutionally recognised grounds are genuinely involved. The essential distinction is between dissent and disorder. A constitutional democracy must tolerate disagreement, controversy, and even uncomfortable political expression; otherwise, public order could become a justification for suppressing legitimate opposition.[60]

The principle of proportionality provides the necessary safeguard. A restriction must pursue a legitimate objective, have a rational connection with that objective, and interfere with the right no more than necessary. This is particularly important in the digital environment because restrictions can affect enormous amounts of lawful expression beyond the content that authorities seek to regulate.[61]

Internet shutdowns illustrate this problem. Restricting internet access affects not only political speech but also education, employment, emergency communication, journalism, and access to information. Similarly, removing a specific post, suspending an account, restricting particular content, and blocking an entire platform are fundamentally different measures. Constitutional regulation must therefore favour targeted intervention over collective restriction.[62]

Misinformation presents a further challenge. False information concerning arrests, injuries, police action, or government decisions during protests may generate fear and hostility. However, giving authorities unrestricted power to determine what constitutes "false" information could itself threaten freedom of expression, particularly where government accounts are contested. Restrictions should therefore have identifiable legal authority, provide adequate reasons, and remain subject to review.[63]

---

[57] Frank Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information 1–18 (2015).

[58] Julie E. Cohen, Between Truth and Power: The Legal Constructions of Informational Capitalism 1–15 (2019).

[59] Yochai Benkler, Robert Faris & Hal Roberts, Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics 1–20 (2018).

[60] David Kaye, Speech Police: The Global Struggle to Govern the Internet 1–18 (2019).

[61] Aharon Barak, Proportionality: Constitutional Rights and Their Limitations 131–45 (2012).

[62] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637.

[63] Claire Wardle & Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making* 5–15 (Council of Europe Report DGI(2017)09, 2017).

The same principle applies to the earlier "RIP" post. A deliberately degrading statement concerning a living individual may raise legitimate concerns regarding dignity and reputation, but treating every person who subsequently shared the post identically to its creator could be disproportionate. Responsibility should consider intention, knowledge, context, participation, and actual harm.[64]

Ultimately, reasonable restriction cannot mean maximum restriction. The existence of a legitimate governmental concern does not justify every regulatory response. Online dissent should remain protected unless expression crosses a constitutionally recognised boundary, and any intervention must remain lawful, necessary, proportionate, and accountable. Freedom of expression must remain the starting point, while restriction remains the carefully justified exception.

10 THE FUTURE OF ARTICLE 19 IN THE DIGITAL AGE

The future of Article 19(1)(a) depends on ensuring that constitutional protection remains meaningful as democratic participation increasingly moves online. This does not require creating a separate right to digital expression. Rather, existing constitutional guarantees must be applied to the technologies through which individuals now communicate, organise, criticise, and dissent.[65]

Political expression must remain at the centre of this framework. Online criticism of government policies, public officials, and institutions should receive strong constitutional protection, even where it is controversial or unpopular. At the same time, digital constitutionalism must recognise that online expression can produce significant harm to dignity, privacy, reputation, and public order. The appropriate response is therefore not unrestricted speech or excessive censorship, but context-sensitive and proportionate regulation.[66]

Future regulation should consider the context, intention, audience, reach, and likely consequences of expression rather than treating isolated words as determinative. The same principle should inform platform governance. As private platforms increasingly influence which political content is visible, greater transparency in content moderation, accessible mechanisms to challenge removals, and accountability in algorithmic amplification are necessary to protect meaningful participation.[67]

The experience of student movements demonstrates the importance of this balance. Digital platforms have enabled students to organise, document events, challenge official narratives, and reach wider audiences. Yet the same spaces expose them to surveillance, misinformation, harassment, and the rapid spread of harmful content. Regulation should address these risks without eliminating the democratic possibilities created by digital communication.[68]

The future framework must therefore move beyond a purely State-centric model towards broader digital accountability. Governments, courts, platforms, and users exercise different forms of power and must operate within appropriate legal and constitutional boundaries. Restrictions under Article 19(2) must remain limited to constitutionally recognised grounds and satisfy reasonableness and proportionality, while the dignity and rights of individuals affected by digital expression must also be protected.[69]

---

[64] RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 184–205 (1977).
[65] ANDREW MURRAY, INFORMATION TECHNOLOGY LAW: THE LAW AND SOCIETY (2016).
[66] WOJCIECH SADURSKI, FREEDOM OF SPEECH AND ITS LIMITS (1999).
[67] REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM (2012).
[68] ZEYNEP TUFEKCI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST (2017).
[69] 2 H. M. SEERVAI, CONSTITUTIONAL LAW OF INDIA (4th ed.).

Ultimately, digital constitutionalism is not about creating an internet free from regulation, nor about granting governments or platforms unlimited control over speech. It is about ensuring that technological power remains subject to constitutional values. The central challenge is to protect dissent strongly enough to sustain democracy while ensuring that regulation of speech does not itself become more damaging to democracy than the speech it seeks to control.[70]

11 CONCLUSION

Digital platforms have transformed the way individuals participate in democratic life, making online spaces an increasingly important arena for political expression, protest, mobilisation, and dissent. Article 19(1)(a) must therefore remain meaningful regardless of the medium through which expression occurs. The experience of student-led protests demonstrates both the democratic potential of digital platforms and the risks they create through misinformation, harassment, amplification, and attacks on individual dignity.[71]

However, protecting online dissent does not mean treating freedom of expression as absolute. Article 19(2) permits reasonable restrictions on constitutionally recognised grounds, but such restrictions must satisfy legality, reasonableness, necessity, and proportionality. At the same time, Article 21 requires the dignity and rights of individuals affected by digital expression to be considered. The "RIP" post concerning a living public official illustrates the difficult boundary between legitimate political criticism and targeted degradation, particularly when digital virality makes accountability difficult.[72]

The decisions in *Shreya Singhal*, *Puttaswamy*, and *Anuradha Bhasin* demonstrate that constitutional rights must adapt to technological change. Yet the challenge now extends beyond State power to private platforms and algorithms that influence what speech is visible and amplified.[73]

Ultimately, digital constitutionalism must protect dissent without permitting either unrestricted digital harm or excessive control. Article 19 must evolve not by becoming weaker, but by becoming responsive to the new forms of power shaping expression in India's digital democracy.[74]

* * *

[70] JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD (2006).
[71] ZEYNEP TUFEKCI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST (2017).
[72] AHARON BARAK, PROPORTIONALITY: CONSTITUTIONAL RIGHTS AND THEIR LIMITATIONS (2012).
[73] REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM (2012).
[74] JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD (2006).

Law in the Digital Decade: Rights, Regulation and Accountability, edited by Gyan Prakash Kesharwani and Ritu Verma (VidhiAagaz 2026) · Open access under CC BY-NC 4.0 · https://doi.org/10.63108/VAB.LDD.1.22

CHAPTER 22

Deepfakes, Defamation, and the Limits of Section 79: A Post-Visakha Analysis in Light of the IT Amendment Rules, 2026

RAIHA OWAIS[*] AND ANHA TAHIR[**]

The introduction of deepfake technology has long exposed gaps in India's intermediary liability system under Section 79 of the Information Technology Act, 2000. This paper revisits Google India Pvt. Ltd. v. Visakha Industries Ltd., which clarified intermediary liability for defamatory content through the "publication by inaction" doctrine, holding that an intermediary's failure to act on actual knowledge of unlawful content may defeat the statutory safe harbour. While Visaka addressed static, user-generated defamation, this paper argues that its reasoning does not sit well with synthetically generated deepfake content, in which the intermediary's own algorithmic infrastructure is implicated in fabrication rather than mere hosting.

The recent Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026, notified 10 February 2026 and enforced from 20 February, represents the legislature's first direct response to this gap, introducing a formal definition of "Synthetically Generated Information" (SGI), mandatory labelling and metadata-tracing obligations, and a 3-hour takedown window for deepfake content, with loss of safe harbour for non-compliant intermediaries. This paper examines whether the 2026 Amendment adequately closes the doctrinal fissure left open by Visaka, or whether its "actual knowledge" and verification standards remain too narrow to address content that is synthetically fabricated rather than merely spread.

Moreover, this paper considers the relationship between the 2026 Amendment, the pending Draft Second Amendment Rules of March 2026, and emerging criminal liability under the Bharatiya Nyaya Sanhita, 2023, for defamation and identity-based offences. It concludes by proposing a graduated due diligence standard tailored to SGI, arguing that labelling and takedown timelines alone cannot substitute for a coherent liability framework that distinguishes conduit from co-creator.

Keywords: Intermediary Liability, Deepfakes, Section 79 IT Act, Synthetically Generated Information (SGI), Safe Harbour

1 INTRODUCTION

The technology behind deepfakes is the output of two competing neural networks, a generator that creates fake content and a discriminator that tries to detect it, trained against each other until the generator's output becomes indistinguishable from an authentic recording.[1] This technique, first demonstrated in 2014 and popularised under the "deepfake" label from 2017 onward, has since migrated from research laboratories into consumer-facing applications capable of producing a photorealistic video or a cloned voice from a handful of source images or seconds of audio.[2]

[*] Faculty of Law, Aligarh Muslim University, Aligarh, Uttar Pradesh, India.
[**] Faculty of Law, Aligarh Muslim University, Aligarh, Uttar Pradesh, India.

[1] Kashish Gupta, *The Future of Deepfakes: Need for Regulation*, 5 NAT'L L.U. DELHI (2023).

[2] John Thayer, *Defamation or Impersonation? Working Towards a Legislative Remedy for Deepfake Election Misinformation*, 66 WM. & MARY L. REV. 255 (2024).

The damage caused by it is immense. In 2018, the investigative journalist Rana Ayyub discovered that a pornographic video bearing her face, generated from publicly available photographs, was circulating on WhatsApp and Twitter, reportedly viewed and shared multiple times within days.[3] Unlike a forged photograph pasted onto a bulletin board, in Ayyub's case this was defamatory and sexually humiliating, and an algorithm manufactured the statement made about her.

India's principal statutory answer to intermediary liability for unlawful third-party content, Section 79 of the Information Technology Act, 2000 ('the IT Act'), was not written with this problem in mind. It was drafted for a much older internet world of static user content: something a real, identifiable person actually wrote, with the platform just hosting or passing it along. The most authoritative judicial gloss on that architecture, the Supreme Court's 2019 decision in *Google India Private Ltd. v. M/S Visakha Industries*[4], addressed such a scenario. An allegedly defamatory article was posted by one identified individual on a Google Group, which the intermediary declined to remove despite notice. The Court's resulting "publication by inaction" doctrine, borrowed from English authorities on club noticeboards and internet service providers, held that an intermediary's continued, knowing inaction in the face of a removable defamatory posting could itself constitute "publication" for criminal defamation under the Indian Penal Code.

This paper asks whether that doctrine, forged for passive hosting of human-authored text, can meaningfully govern a technology in which the intermediary's own generative infrastructure participates in the making of the defamatory statement. It argues that it cannot do so without strain, because *Visakha* assumes three things that deepfakes break: a fixed, identifiable piece of content; one clear moment when the platform "knew"; and a simple choice between taking something down or leaving it up. But when the platform itself provides the tool (a face-swap filter, a voice cloner, an image generator), the real question isn't whether it failed to remove something after being notified. It's whether it helped make the thing in the first place.

2 SECTION 79 AND THE ARCHITECTURE OF INTERMEDIARY IMMUNITY

Section 79, before its 2009 substitution, provided the narrower shield as it exempted what the statute then called a network service provider from liability under the IT Act and its subordinate rules or regulations for third-party information it made available, but only if the provider could show both that it had no knowledge of the offending act and that it had exercised due diligence to prevent it.[5] As the Supreme Court later clarified in *Visakha*, this pre-2009 exemption was narrow in a specific sense, because it sat in Chapter XII of the IT Act, it shielded intermediaries only from liability arising under the Act itself, and did not touch offences under general criminal law such as defamation under Section 356 of the Bharatiya Nyaya Sanhita (previously Sections 499–501 of the Indian Penal Code).[6]

The 2009 amendment changed this significantly. The revised provision opens with a non-obstante clause overriding conflicting law, exempting intermediaries from liability for third-party content they host, language broad enough to cover general criminal law, unlike the earlier version. This immunity is conditional: under Section 79(3)(b), it lapses if the intermediary fails to act expeditiously after gaining actual

[3] Abigail George, *Defamation in the Time of Deepfakes*, 45 COLUM. J. GENDER & L. 45 (2024) (recounting Rana Ayyub, *I Was the Victim of a Deepfake Porn Plot Intended to Silence Me*, HUFFPOST (Nov. 21, 2018)).
[4] AIR 2020 SC 350.
[5] *Id.*
[6] *Id.*

knowledge or receiving government notification of unlawful content. Under the 2011 Intermediary Guidelines Rules, Rule 3(4) required takedown once the intermediary knew, or was told, that hosted material was defamatory (among other things). Both these provisions were challenged before the Supreme Court in *Shreya Singhal v. Union of India, and the Court, concerned about turning intermediaries into unilateral judges over the legality of the vast volume of content that passes through them, narrowed the meaning* of actual knowledge in Section 79(3)(b) and Rule 3(4) so that it arises only once a court order or an appropriate-government notification communicates it.[7] *Visakha* later summed up the reasoning as essentially practical: without that kind of external, authoritative trigger, large intermediaries would have no real way to sort the genuinely legitimate takedown demands from the flood of requests they receive every day.[8] Constitutionally, then, there is safe harbour immunity from general law, and not merely from liability under the IT Act, at the cost of a narrow, externally-gated knowledge standard.

The 2021 Intermediary Guidelines and Digital Media Ethics Code Rules layered further due-diligence obligations onto this structure, particularly for larger platforms designated as significant social media intermediaries: a grievance redressal mechanism, expeditious removal of flagged sexually explicit or impersonating content, and encouragement, though not an unconditional mandate, to deploy automated tools against certain categories of unlawful material, subject to periodic human review.[9] Even here, the underlying conception of the intermediary's role did not change: it remained that of a facilitator with no editorial control over content authored by others, a characterisation the Government of India itself pressed in *Visakha* when distinguishing intermediaries from newspaper editors under the Press and Registration of Books Act, 1867.[10]

This entire architecture rests on the statutory definition of an intermediary under Section 2(1)(w) of the IT Act as an entity that, on behalf of another person, receives, stores or transmits an electronic record, or otherwise provides a service in respect of it. As the Government of India's own AI Governance Guidelines Committee has since observed, a definition built around telecom carriers, search engines and cyber cafes sits uneasily with AI systems that produce output from user prompts, or even act autonomously, refining their results through continuous learning.[11] The Committee went further, noting that Section 79 immunity depends on the intermediary not initiating transmission, choosing the recipient, or altering the data in question, conditions that many content-generating AI systems may simply fail to satisfy.[12]

3 THE PUBLICATION-BY-INACTION DOCTRINE

The complainant company initiated criminal defamation proceedings against accused 1, the coordinator of "Ban Asbestos Group", for authoring and publishing allegedly defamatory articles titled "Visakha Asbestos Industries making gains" (on 31-07-2008) and "Poisoning the system" (on 20-11-2008), and accused 2 (Appellant), Google India Pvt. Ltd., for hosting the publications. On 09-12-2008, the complainant issued a notice for takedown; the appellant failed to do so. Google India resisted on several

[7] Shreya Singhal, (2015) 5 SCC 1.
[8] *Supra* note 4.
[9] Gupta, *supra* note 1.
[10] Gambhirsinh R. Dekare v. Falgunbhai Chimabhai Patel.
[11] GOV'T OF INDIA, INDIA AI GOVERNANCE GUIDELINES (Ministry of Elec. & Info. Tech., Nov. 5, 2025).
[12] *Id.*

grounds: that it was not itself the intermediary (Google LLC, its US parent, owned and operated Google Groups), that the notice did not specify the objectionable URL and that, as a passive host, it could not be attributed actual knowledge of content posted by millions of users without editorial oversight.[13]

Drawing from *Byrne v. Deane* (1937) and *Godfrey v. Demon Internet Ltd.* (2001), the Court held that the test for publication in an internet context is that a party which, having the power, the right and the ability to remove the defamatory matter, and being called upon to do so, defies and rejects the request may itself be treated as a publisher. Passive hosting alone does not attract liability; it is the deliberate inaction after notice that does. However, for the offence under Section 500 IPC (Section 356(2) BNS), publication alone is insufficient. The accused must also have acted with the requisite intention or knowledge that the content would harm the reputation of the person concerned. This meant that an intermediary who is alerted to defamatory material and possesses the technical means to take it down, but refuses to do so, may thereby become a publisher of it, provided the separate mental element required by Section 356, intention or knowledge of likely harm to reputation, is also independently made out.[14]

Two qualifications the Court itself attached to this holding deserve emphasis, because they mark the doctrine's outer edge. First, the Court expressly noted that the noticeboard analogy from Byrne is not straightforwardly transferable to the internet, observing that the scale and structure of online publication are materially different from a wall in a golf club.[15] Second, and more importantly for present purposes, the Court left open whether, absent a court order, an intermediary could lawfully insist on judicial adjudication before removing content even where it was, on the facts, the true intermediary. The eventual disposition was correspondingly narrow: the Court set aside the High Court's findings on the merits of Google India's refusal to act on notice, and remanded the matter to the trial Magistrate to determine, on evidence, whether the appellant was in fact the intermediary and whether publication had occurred.[16]

What *Visakha* decided, in other words, was a question about the legal consequence of inaction toward a fixed, human-authored object of publication whose only relevant lifecycle event before the court was a single act of hosting and a single, later act of refusal to de-list. The appellant's conduct at issue was entirely custodial: it did not write, edit, or algorithmically transform the Ban Asbestos India articles; its only alleged wrongdoing was declining to take them down. That is precisely the fact pattern for which the notice-and-knowledge architecture of Section 79, as read down in *Shreya Singhal*[17], was designed.

4 FROM CONDUIT TO CO-CREATOR: WHY VISAKHA DOES NOT TRAVEL WELL TO SYNTHETIC MEDIA

Deepfakes disturb the *Visakha* framework at each of the three junctures identified above: the determinacy of the object of publication, the meaningfulness of a single point of "actual knowledge," and the binary remove-or-retain choice on which the publication-by-inaction test depends.

On the first point, a defamatory article is, once posted, a fixed textual artefact: the same words persist until edited or removed, and "the defamatory matter" referred to throughout the *Visakha* judgment is a stable referent. A deepfake generated through a generative adversarial network, by contrast, is the terminal

[13] *Supra* note 4.
[14] *Id.*
[15] *Id.*
[16] *Id.*
[17] (2015) 5 SCC 1.

output of an iterative training process in which a generator network continuously refines its forgeries against a discriminator network until the two become computationally indistinguishable.[18] Because the underlying model can regenerate near-identical but technically distinct outputs on demand, and because photorealistic synthetic media collapses, in a way ordinary text does not, the reader's ability to distinguish signifier from signified, a deepfake's capacity to cause reputational harm survives even the viewer's knowledge that it is fake.[19] The Byrne noticeboard test asks, in essence, whether the defendant's own conduct effectively kept defamatory matter in place; but where the matter in question can be trivially regenerated, subtly altered to defeat hash-based detection, and redistributed through parallel channels, the very idea of one fixed item whose ongoing presence a defendant can be said to have caused becomes strained.

On the second point, *Visakha's* entire notice-based architecture presumes an identifiable moment at which knowledge attaches, the 9 December 2008 letter in that case, after which the intermediary's continued inaction becomes legally significant. Rana Ayyub's experience illustrates that the video was not merely disseminated widely but reproduced and re-shared across WhatsApp and Twitter simultaneously and virally, with no single upload event or single platform from which a notice-and-takedown obligation could meaningfully flow.[20] The Supreme Court itself, in Shreya Singhal, narrowed the actual-knowledge standard, reasoning that requiring intermediaries to adjudicate legitimacy across the volume of the internet, with countless posts every minute even back in 2015, risked either overwhelming platforms or effectively handing them a private censorship role that the Court was unwilling to sanction. Generative tools multiply this volume problem by an additional order of magnitude; content moderation research cited in the comparative deepfake-regulation literature confirms that automated detection tools deployed against synthetic media are themselves imperfect and even good-faith compliance with a notice may not reliably identify or suppress all derivative copies.[21]

The third and most consequential disturbance concerns the remove-or-retain binary itself. In *Visakha*, the intermediary's only relevant choice, when notified, was whether to take down content that a third party had already, independently, authored. Where the offending content is instead produced using a generative feature that the intermediary itself built into its platform, a generator function, the intermediary's causally significant act occurs before any question of notice arises at all: it lies in having supplied the generative capability that made the fabrication possible. This is not a hosting decision at all; it more closely resembles the kind of conduct the AI Governance Guidelines Committee flagged as falling outside the immunity altogether, initiating a transmission, choosing a recipient, or altering the underlying data, conduct that Section 79 was never designed to cover, and which the Committee concluded may put many generative AI systems beyond the safe harbour independent of any subsequent takedown failure.[22] Put differently, *Visakha's* publication-by-inaction doctrine is built to answer whether an intermediary wrongly kept hosting something it should have taken down. Deepfakes produced through platform-native tools instead raise the logically prior question of whether the intermediary wrongly helped make the thing in the first place, a question closer to authorship or origination under the IT Act's own definitions than to custody, and one the publication-by-inaction test was never built to ask.

---

[18] Thayer, *supra* note 2, at 256.
[19] Mohit Kar & Shreya Sahoo, *Deepfakes and Its Iniquities: Regulating the Dark Side of AI*, 5 NAT'L L.U. ODISHA STUDENT L.J. 58 (2020).
[20] George, *supra* note 3.
[21] Andrew Ray, *Disinformation, Deepfakes and Democracies: The Need for Legislative Reform*, 44 UNSW L.J. 983 (2021).
[22] GOV'T OF INDIA, *supra* note 11.

This is not merely a theoretical distinction. Comparative commentary on deepfake-enabled election misinformation in the United States has identified an analogous structural gap in Section 230 of the Communications Decency Act, which likewise assumes a sharp separation between the platform and the third-party source of the content it hosts, a separation that erodes once platforms themselves supply generative tools that materially contribute to the content's creation.[23] Free-speech scholarship on synthetic media has similarly observed that existing doctrinal exceptions built for authored falsehoods (defamation, intentional infliction of emotional distress, and the right of publicity) were not designed with an algorithmic co-author in mind, and each must be expanded to reach a wrong whose proximate cause is a model rather than a person.[24] The right-of-publicity literature makes a parallel point from a different angle: that doctrine emerged historically to address the portable camera and mass media together, as a *combination* of a new way of capturing someone's likeness and a new way of disseminating it, and deepfakes present exactly that same two-part structure, except that the device doing the capturing is now owned and operated, in many cases, by the very platform accused of merely hosting the result.[25]

None of this suggests that *Visakha* was wrongly decided on its own facts, nor that the publication-by-inaction doctrine is without continuing value. For content that remains genuinely third-party-authored and merely hosted, such as a deepfake created with an independent, off-platform tool and later uploaded to a hosting service, *Visakha's* notice-based inquiry remains apt and indeed necessary, since it is the mechanism by which victims can compel removal without waiting for a court order in every case where the intermediary is truly custodial. The argument, rather, is that *Visakha* cannot be the *sole* doctrinal resource for deepfake-enabled defamation, because a meaningful share of real-world deepfake harm is produced using generative infrastructure that the hosting platform itself supplies, and for that category the publication-by-inaction inquiry is simply the wrong question.

5 THE IT AMENDMENT RULES, 2026: DOES THE LEGISLATURE CLOSE THE GAP?

MeitY notified the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules, 2026 by Gazette Notification G.S.R. 120(E) on 10 February 2026. These amendments came into force on 20 February 2026. The Amendment is India's first substantive statutory framework directed at AI-generated content as such, following an October 2025 draft that had circulated for stakeholder consultation, as several contemporaneous commentators have noted.[26]

The centrepiece of the 2026 Amendment is a new, technology-neutral definition of "Synthetically Generated Information" (SGI) which is audio, visual, or audio-visual content that has been created or algorithmically altered using computer resources in a manner that makes it appear authentic, such that a person would be misled into believing it to be genuine.[27] Two features should be flagged. First, it is not confined to AI-generated content in the narrow, model-based sense: it is deliberately broad enough to capture other forms of algorithmic alteration as well, a choice several commentators regard as inviting both over- and under-inclusion.[28] Further, the Rules impose their obligations on the *intermediary that hosts*

[23] Thayer, *supra* note 2, at 277.
[24] Jacob Bourgault, *Free Speech and Synthetic Lies: Deepfakes, Synthetic Media, and the First Amendment*, 3 J. INTELL. PROP. L.
[25] Michael P Goodyear, *Dignity and Deepfakes*, 57 ARIZ. ST. L.J. 931 (2025).
[26] AZB & Partners, *India's New Law Governing Synthetic Media*, https://www.azbpartners.com/bank/88502/ (last visited Aug. 28, 2026).
[27] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, r. 2(wa).
[28] ANM Global, *Regulating Synthetic Content: Key Implications of the IT Rules Amendment, 2026*, https://anmglobal.net/articles/regulating-synthetic-content-key-implications-of-the-it-rules-amendment-2026 (last visited Aug. 29, 2026).

SGI, largely without differentiating between an intermediary that merely stores synthetic content uploaded from elsewhere and one whose own in-platform tools generated it.

Substantively, the 2026 Amendment layers three obligations onto covered intermediaries. First, mandatory labelling: SGI must carry a label that is prominent and noticeable to an ordinary viewer or listener, a qualitative standard that replaced an earlier, more mechanical draft proposal requiring a watermark covering at least ten per cent of a visual work's surface area, or running through the first ten per cent of an audio clip's duration.[29] Second, permanent, non-strippable provenance metadata, broadly consonant with international content-authentication standards such as the Coalition for Content Provenance and Authenticity (C2PA), which the Government's own AI Governance Guidelines Committee had recommended examining as a mechanism for tracing whether content was AI-generated or modified.[30] Third, and most sharply, a compressed takedown window reported as three hours for flagged unlawful synthetic content generally, a substantial reduction from the up-to-thirty-six-hour windows the 2021 Rules allowed for government or court-ordered takedowns, with an even shorter window reported for non-consensual sexual imagery specifically.[31] Non-compliance carries a severe consequence: forfeiture of Section 79 safe harbour, with the non-compliant intermediary potentially treated, for liability purposes, as if it had itself created the unlawful content.[32]

The doctrinal issue this paper identifies survives the 2026 Amendment largely intact, for four related reasons. First, the safe-harbour forfeiture mechanism is still triggered by a failure to act within the takedown window following a complaint or governmental notification. In other words, it retains precisely the "*actual* knowledge" gate that Shreya Singhal read into Section 79(3)(b) and that *Visakha* applied to a purely custodial host.[33] The 2026 Amendment compresses the clock; it does not relocate the trigger. It therefore leaves undisturbed the prior question this paper has pressed, whether the intermediary's own generative infrastructure contributed to the fabrication, which remains analytically anterior to, and unaddressed by, any timeline for post-notice removal.

Second, the Internet Freedom Foundation's detailed submissions on the preceding draft, several of which persisted into critique of the final Rules, warned that the SGI definition's reliance on whether content would appear authentic to a reasonable viewer is broad and subjective, sweeping in low-risk domestic uses of AI, filters, translation tools, alongside high-risk political or sexual deepfakes, without any risk-tiering to differentiate intermediary obligations accordingly.[34] A regime that regulates by appearance rather than by the intermediary's own causal proximity to the generative act cannot, by construction, distinguish a platform that merely hosts an externally-produced deepfake from one whose own tools produced it, the very distinction this paper argues is doctrinally essential.

Third, constitutional commentary on the notified Rules has raised concerns continuous with those that led the Supreme Court to read down Section 79(3)(b) in Shreya Singhal in the first place. One detailed critique argues that mandatory content-based labelling risks operating as compelled speech, and that provisions conditioning safe harbour on compliance with MeitY's clarifications, advisories and directions,

---

[29] *Id.*

[30] GOV'T OF INDIA, *supra* note 11.

[31] AZB & Partners, *supra* note 26.

[32] The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, r. 7; AZB & Partners, *supra* note 26.

[33] *Supra* note 4.

[34] Internet Freedom Foundation, *IT Intermediary Amendment Rules, 2026 Contradict Their Purpose*, https://internetfreedom.in/ (last visited Aug. 27, 2026).

carried forward and expanded in the subsequent Draft Second Amendment, may reproduce the vagueness and overbreadth for which Section 66A of the IT Act was struck down in Shreya Singhal.[35] To the extent these objections are well-founded, they counsel against simply widening the existing notice-and-takedown gate further (for instance, by shortening timelines even more or broadening the SGI definition still further), since doing so intensifies the very features of the pre-existing architecture that this paper argues are ill-suited to the generation-stage problem, rather than curing them.

Fourth, and practically, the compressed three-hour window is reported to strain even well-resourced intermediaries' capacity to verify both the *legality* of a takedown request and the authenticity of the underlying content (that is, whether it is synthetic at all, and whether provenance metadata has been stripped or spoofed) within the time allowed, a strain analysts have suggested will incentivise precautionary, over-inclusive removals with attendant costs to legitimate speech.[36] This is, in effect, the volume-and-verification problem the Supreme Court worried about in Shreya Singhal and *Visakha*, replicated in a harder register: intermediaries are now expected not merely to judge legality quickly, but to judge legality *and* technical authenticity quickly, using a legal test (actual knowledge on notice) that was calibrated for neither task.

The net assessment, then, is that the 2026 Amendment answers the question of how to make synthetic content identifiable and remove it quickly once flagged, but leaves substantially open the question this paper considers doctrinally prior: who is responsible, independent of any notice, when the platform's own infrastructure was used to fabricate the content in the first place. Because *Visakha's* publication-by-inaction test cannot answer that question, and because the 2026 Amendment's safe-harbour forfeiture mechanism operates through the same notice-triggered gate *Visakha* applied, the doctrinal fissure between conduit and co-creator persists after the 2026 Amendment, notwithstanding its genuine advances on labelling and provenance.

6 THE DRAFT SECOND AMENDMENT RULES, MARCH 2026, AND CRIMINAL LIABILITY UNDER THE BHARATIYA NYAYA SANHITA

On 30 March 2026, MeitY published the Draft Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Second Amendment Rules, 2026 for public consultation, issued under Sections 69A, 79 and 87 of the IT Act. Whereas the February 2026 Amendment focused on SGI as a content category, the Draft Second Amendment shifts the register toward institutional and user-level obligations. It proposes to extend publisher-like duties (grievance-officer contact details, code-of-ethics adherence, and fixed takedown responsiveness) to individual users who post "news and current affairs" content beyond prescribed thresholds, treating them, in effect, as publishers; it proposes a continuous, duration-long display requirement for SGI labels rather than a momentary disclosure; and it proposes an inter-ministerial "Digital Content Review Panel" combining MeitY and Ministry of Information and Broadcasting officials to adjudicate user–publisher disputes.[37]

---

[35] Rudraksh Lakra, *Guest Post: A Constitutional Critique of the Synthetically Generated Information (IT Rules Amendment), 2026*, INDIAN CONSTITUTIONAL LAW AND PHILOSOPHY (blog, Feb. 11, 2026).

[36] ANM Global, *supra* note 28.

[37] Government of India Ministry of Electronics and Information Technology, *Draft Amendments to the IT Rules, 2021 in Relation to Strengthening Intermediary Compliance with Clarifications, Advisories and Directions Issued by the Ministry and Digital Media Oversight*, https://www.meity.gov.in/documents/act-and-policies/draft-amendments-to-the-it-rules-2021-in-relation-to-strengthening-intermediary-compliance-with-clarifications-advisories-and-directions-issued-by-the-ministry-and-digital-media-oversight-AjM2YjMtQWa?pageTitle=

For the purposes of this paper's argument, the Draft Second Amendment is best read as complementary rather than corrective. Its innovations address a different structural gap (the blurring line between individual social media users and institutional publishers) rather than the conduit/co-creator distinction pressed in Parts 4 and 5. If anything, by directing regulatory and public attention toward user-level publisher obligations and toward compliance with MeitY's advisory directions, the Draft Second Amendment risks further deferring scrutiny of the intermediary's own generative-AI features, which remain regulated, if at all, through the same notice-triggered SGI takedown mechanism examined in Part 5.

Running beneath both sets of Rules is the substantive criminal law under the Bharatiya Nyaya Sanhita, 2023, which replaced the Indian Penal Code with effect from 1 July 2024. Section 356 BNS[38] replaces IPC Section 500 as the offence of criminal defamation, retaining the core "making or publishing" structure that the *Visakha* Court applied to Google India. Section 336 BNS addresses forgery for harming reputation, what commentators describe as enhanced clarity for digital forgery and impersonation, a provision of obvious relevance to a deepfake that fabricates a statement or image attributed to the victim. Sections 351 and 353 BNS[39] address criminal intimidation and statements intended to cause public mischief through digital platforms, Section 77[40] extends voyeurism-adjacent liability to non-consensual capture and circulation of intimate images, and Section 319[41] addresses cheating by personation, potentially relevant to voice-cloning fraud. The Digital Personal Data Protection Act, 2023 supplies a further, non-criminal layer, since the unauthorised processing of biometric data (a face or a voice) to construct a deepfake independently attracts penalties under that statute, reported as high as ₹250 crore for serious violations.[42]

Because these BNS offences, like their IPC predecessors in *Visakha*, sit outside Chapter XII of the IT Act, the *Visakha* Court's foundational holding that Section 79 does not immunise an intermediary from criminal liability under general law once "publication" is independently established continues to apply in principle after the 2026 Amendment. But this is precisely where the conduit/co-creator distinction re-emerges with sharpened stakes: attributing the "making" of a synthetic statement of fact under BNS Section 356 to a platform's generative model, rather than to an individual human actor, poses an actus reus and mens rea problem that neither the 2026 Amendment nor the Draft Second Amendment purports to resolve, because both instruments operate in the regulatory register of safe-harbour conditions rather than the criminal-attribution register of who "made" the statement. A coherent answer to that criminal-law question, this paper suggests in Part 7, requires the same conduit/co-creator sorting this paper has pressed throughout: an evidentiary sorting mechanism analogous to the rebuttable presumption of editorial responsibility the *Visakha* Court itself discussed in the context of newspaper editors under the Press and Registration of Books Act.[43]

Draft-amendments-to-the-IT-Rules%2C-2021-in-relation-to-strengthening-intermediary-compliance-with-clarifications%2C-advisories-and-directions-issued-by-the-Ministry-and-digital-media-oversight (last visited Sept. 21, 2026).

[38] Bharatiya Nyaya Sanhita, 2023, § 356.

[39] *Id.* §§ 351, 353.

[40] *Id.* § 77.

[41] *Id.* § 319.

[42] The Digital Personal Data Protection Act, 2023, The Schedule, item 1.

[43] *Supra* note 10.

7 TOWARDS A GRADUATED DUE DILIGENCE STANDARD FOR SYNTHETICALLY GENERATED INFORMATION

The preceding Parts converge on a single proposition: labelling obligations and compressed takedown timelines regulate the *visibility* of synthetic content and the *speed* of the response to a complaint, but neither allocates responsibility according to an intermediary's *causal proximity* to the act of synthesis. That question is precisely the one the Government's own AI Governance Guidelines Committee flagged as unresolved when it called for clarity on how liability should be apportioned "across the AI value chain", developer, deployer, and user, under an intermediary-liability statute never drafted with that chain in mind.[44] This paper proposes a three-tier due diligence standard responsive to that call, built around the intermediary's relationship to the generative act rather than solely around its response to notice.

Pure Conduit. Where an intermediary merely hosts SGI that was authored and generated entirely by a third-party using tools external to the platform (the closest analogue to the facts in Visakha itself), the existing publication-by-inaction, notice-based framework remains appropriate and should continue to govern. Safe harbour should remain available, subject to the labelling, provenance-preservation, and compressed takedown obligations the 2026 Amendment already imposes.[45] This is the category for which *Visakha's* doctrine was designed, and it continues to do useful work here.

Enabling Infrastructure. Where an intermediary provides in-platform generative tools that a user then operates to produce the SGI, the intermediary should bear a heightened *proactive* due diligence duty that attaches independently of any subsequent notice. Concretely, this would include mandatory embedding of provenance metadata at the point of generation (operationalising the C2PA-style standards the AI Governance Guidelines Committee has already endorsed for study),[46] technical safeguards against generating outputs that mimic identifiable real persons without their consent (for instance, matching against opt-out registries for public figures and complainants), and liability for failure to implement regardless of whether a takedown notice was ever served.

Co-creation and Active Curation. Where an intermediary's own algorithmic systems (recommendation engines that substantially amplify synthetic content, or those that synthesise third-party inputs into new outputs) themselves modify or synthesise content rather than merely transmitting it unaltered, the intermediary should be treated, for that specific output, as akin to a publisher or originator, and should not be entitled to Section 79 immunity for it at all. This tier operationalises the AI Governance Guidelines Committee's own observation that Section 79 immunity was conditioned on the intermediary not modifying transmitted data, and that many AI systems fall outside that condition as a matter of ordinary statutory construction, independent of any amendment.[47] It also extends, into the algorithmic context, the conduit/editor distinction the *Visakha* Court itself drew from the newspaper context, where an editor exercising selection and control over published matter attracts a presumption of responsibility that a mere distributor does not.[48]

Two safeguards would keep this graduated standard within the constitutional boundaries. First, judicial or designated-officer certification should remain the trigger for any *content-based* unlawfulness

[44] GOV'T OF INDIA, *supra* note 11.
[45] AZB & Partners, *supra* note 26.
[46] GOV'T OF INDIA, *supra* note 11.
[47] *Id.*
[48] *Supra* note 10.

determination: whether a given piece of synthetic content is in fact defamatory, or otherwise unlawful, should continue to require an external adjudicative gate, consistent with *Shreya Singhal's* read-down of "actual knowledge".[49] Second, and critically, the Tier Two obligations proposed above (provenance embedding and consent-matching safeguards) should be delinked entirely from that gate, because they are *ex ante design obligations* rather than *ex post content-adjudication obligations*. A duty to embed a watermark at the point of generation does not require anyone to judge whether the resulting content is defamatory; it only requires the platform to build its tools responsibly. Decoupling these two obligation-types in this manner should substantially answer the overbreadth and vagueness objections the Internet Freedom Foundation and constitutional commentators have raised against provisions that condition safe harbour on broad, content-based labelling duties or on compliance with open-ended ministerial advisories.[50]

Finally, this graduated standard would also assist the criminal-law attribution problem identified in Part 6. By these clear tiers of intermediary proximity to the generative act, it would supply courts adjudicating BNS Section 356 prosecutions with an evidentiary framework analogous to the rebuttable presumption of editorial responsibility already recognised for print editors, allowing a Tier Three intermediary's algorithmic co-creation to ground an inference of "making" under the BNS, while preserving Tier One intermediaries' position as mere custodians whose liability continues to depend on proof of knowing inaction, exactly as *Visakha* contemplated.

8 Conclusion

Google India Pvt. Ltd. v. Visakha Industries Ltd. filled a genuine gap in Indian intermediary-liability jurisprudence, establishing that an intermediary with both the power and the practical ability to remove defamatory material, which knowingly declines to do so, may be treated as a publisher through its inaction, whatever its ostensibly custodial role.[51] That doctrine remains sound for the case it was built for: a static, human-authored posting merely hosted by a passive conduit.

It does not, however, describe deepfakes produced using an intermediary's own generative infrastructure. There, the prior question is not whether the intermediary failed to remove something after notice, but whether its tools helped fabricate that thing in the first place, a question of origination that publication-by-inaction was never designed to ask.

The IT Amendment Rules, 2026 mark real progress: India's first statutory definition of Synthetically Generated Information, mandatory labelling and provenance-tracing, and a compressed takedown window.[52] Yet because safe-harbour forfeiture still hinges on the same actual-knowledge, notice-triggered gate Shreya Singhal read into Section 79(3)(b) and *Visakha* applied to a custodial host, the conduit/co-creator distinction remains largely unresolved, a gap independent critiques of the Rules bear out. The Draft Second Amendment Rules and the Bharatiya Nyaya Sanhita, 2023 add complementary but non-coextensive layers of publisher and criminal liability.[53]

[49] (2015) 5 SCC 1.
[50] Lakra, *supra* note 35.
[51] *Supra* note 4.
[52] AZB & Partners, *supra* note 26.
[53] Bharatiya Nyaya Sanhita, 2023, §§ 351, 353.

A graduated due diligence standard, tying obligations to an intermediary's causal proximity to the act of synthesis rather than only to its response to notice, offers a way forward, one faithful to *Visakha's* own insight that control over content, not merely custody of it, should determine liability.

* * *

About the Publisher

VidhiAagaz is an academic publisher based in Maihar, Madhya Pradesh, publishing since 2017. It is a member of Crossref (DOI prefix 10.63108), and its publications are indexed in prominent databases, including HeinOnline, Google Scholar and Manupatra. Every chapter of this book carries its own DOI and can be read free of charge at vidhiaagaz.com.

Publish your conference with us

Universities, departments and conference organisers can publish their conference papers with us as an edited book or as a special issue of our journals, the *International Journal of Law Management & Humanities* (IJLMH) and the *International Journal of Legal Science and Innovation* (IJLSI). We take a volume from manuscript to print: copy-editing, citation checks, typesetting, ISBN and DOI registration, open-access publication online, and printed copies.

Write to support@vidhiaagaz.com

vidhiaagaz.com/collaborate

Law in the Digital Decade

Rights, Regulation and Accountability

Algorithms now help decide who receives welfare benefits, how disputes are resolved, which speech stays online and how personal data is used. The law has had to answer in a decade, not a century.

This volume brings together twenty-two chapters from the CLRS International Conference on Law in the Digital Decade, held in September 2026. They ask how automated decisions by the State can be held to constitutional standards; how artificial intelligence should be regulated for fairness, accountability and liability; what AI means for courts and online dispute resolution; and how privacy, data protection and free speech are being reshaped online, with close attention to the Digital Personal Data Protection Act, 2023 and the Information Technology Rules.

The chapters combine doctrinal analysis with comparative material from the European Union, the United Kingdom and the United States. The book is written for students, researchers, practitioners and policy-makers working on law and technology in India.

About the Editors

Gyan Prakash Kesharwani is the Chairperson and Publishing Editor at VidhiAagaz, India.

Ritu Verma (B.A. LL.B., LL.M.) is an Editorial Executive at VidhiAagaz, India.

VidhiAagaz®
Inking Your Brain

www.vidhiaagaz.com

ISBN 978-81-947620-6-5