International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Coding Accountability: Operationalizing Ethics-by-Design Frameworks in India's Algorithmic Public Administration

Lt. Kongala Sukumar

Assistant Professor of Public Administration
MALD Government Degree College,
GADWAL, Pin-509126, Telangana

Abstract:
The rapid diffusion of artificial intelligence and algorithmic decision-making tools across Indian public administration — from welfare targeting to predictive policing — has outpaced the development of institutional mechanisms capable of ensuring accountability. This paper examines the ethics-by-design paradigm, which advocates embedding ethical constraints into algorithmic systems at the design stage rather than retrofitting oversight after deployment, and investigates the extent to which such frameworks can be operationalized within India's administrative and regulatory context. Drawing on a review of NITI Aayog's National Strategy for Artificial Intelligence and its subsequent Responsible AI approach papers, sectoral case material on Aadhaar-linked welfare delivery, predictive policing pilots, and algorithmic credit scoring, and a comparative reading of the European Union's proposed risk-based Artificial Intelligence Act, the paper identifies structural gaps in transparency, auditability, and grievance redress that limit accountability in practice. It proposes a layered operational framework combining algorithmic impact assessments, mandatory audit trails, calibrated human oversight, and independent grievance mechanisms, adapted to India's federal administrative structure and capacity constraints. The analysis suggests that ethics-by-design principles, while conceptually robust, require substantial institutional retrofitting — including statutory backing, capacity building, and civil society participation — before they can meaningfully constrain algorithmic power in Indian governance. The paper concludes with a phased implementation roadmap.

Keywords: algorithmic accountability; ethics-by-design; artificial intelligence governance; public administration; India; algorithmic impact assessment

1. Introduction

Over the past decade, algorithmic systems have moved from the periphery to the core of Indian public administration. Biometric authentication now mediates access to subsidised food grains under the Public Distribution System (PDS); direct benefit transfers are routed and reconciled through automated matching of Aadhaar-linked bank accounts; several state police forces have piloted facial recognition and predictive-policing tools; and algorithmic credit-scoring models increasingly determine eligibility for micro-loans and government-backed financial products. These systems are typically introduced with an explicit welfare or efficiency justification — reducing leakage, speeding up service delivery, or improving the

allocation of scarce administrative attention — and are often celebrated as instruments of a more transparent and less discretionary state.

Yet the same opacity that algorithmic systems are meant to overcome in human bureaucracies frequently reappears in coded form. When a biometric authentication failure denies a household its monthly ration, or when a policing algorithm flags a neighbourhood for intensified surveillance, the citizen affected has few avenues to understand why the decision was made, contest it, or hold anyone accountable for its consequences. This is not a uniquely Indian problem; it is the central preoccupation of the international algorithmic accountability literature (Diakopoulos, 2016; Pasquale, 2015). What is distinctive about the Indian case is the scale at which these systems operate — Aadhaar alone covers over 1.3 billion residents — and the administrative context into which they are inserted: a federal structure with uneven technical capacity across states, a still-evolving data protection regime, and a bureaucracy accustomed to discretionary, paper-based accountability mechanisms that do not translate easily into code.

India's principal policy response to these concerns has been NITI Aayog's two-part Responsible AI approach, which followed the 2018 National Strategy for Artificial Intelligence and set out both the principles that should govern AI use (NITI Aayog, 2021a) and, subsequently, a roadmap for operationalizing them (NITI Aayog, 2021b). These documents mark an important discursive shift — they explicitly name accountability, transparency, and non-discrimination as governing values for AI in Indian public administration. However, principle statements are not, by themselves, operational mechanisms. The central question this paper addresses is therefore: how can the abstract commitments of India's responsible AI discourse be translated into concrete design, procedural, and institutional mechanisms capable of producing accountable algorithmic administration in practice?

The stakes of this question are not merely academic. India's Aadhaar system is the largest biometric identity infrastructure in the world, and its architecture has, since 2014, been progressively layered onto an expanding range of public services under the broader Jan Dhan–Aadhaar–Mobile (JAM) digital public infrastructure agenda — encompassing not only PDS authentication but also direct benefit transfer of subsidies, pension disbursal, and increasingly, private-sector know-your-customer verification. Each additional use case extends the population for whom algorithmic failure carries direct material consequence, while the accountability architecture governing that failure has, as this paper documents, lagged well behind the pace of functional expansion. It is this widening gap between functional scale and accountability infrastructure, rather than any single failure of technical design, that motivates treating "coding accountability" as an urgent and distinct object of policy analysis in its own right, separate from the broader and more general debate about AI ethics principles.

To answer this question, the paper draws on the ethics-by-design paradigm developed in the AI ethics literature, most systematically by Dignum (2019), which distinguishes between embedding ethical reasoning directly into system behaviour ("ethics by design"), evaluating the ethical implications of a system's use ("ethics in design"), and establishing external codes and standards that govern developers ("ethics for design"). The paper argues that operationalizing accountability in Indian algorithmic administration requires attention to all three registers simultaneously, organised as a layered framework spanning the value-specification, technical, procedural, and institutional levels of a system's lifecycle.

The remainder of the paper is organised as follows. Section 2 reviews the international and India-specific literature on algorithmic accountability and ethics-by-design. Section 3 develops the paper's theoretical framework. Section 4 outlines the paper's methodology. Section 5 examines three sectoral cases of algorithmic public administration in India. Section 6 proposes a layered mechanism framework for

operationalizing accountability. Section 7 situates the Indian case comparatively against the European Union's and OECD's regulatory approaches. Section 8 discusses structural barriers to implementation, Section 9 offers policy recommendations, and Section 10 concludes.

2. Literature Review

2.1 The Proliferation of AI Ethics Principles

A large body of scholarship has documented the global proliferation of AI ethics guidelines issued by governments, corporations, and multilateral bodies over the past several years. Jobin et al. (2019), in a widely cited mapping of 84 such documents, found substantial convergence around a small set of principles — transparency, justice and fairness, non-maleficence, responsibility, and privacy — but far less agreement on how these principles should be interpreted or implemented. Floridi et al. (2018), synthesising this convergence into the AI4People framework, added "explicability" as a distinct fifth principle, cutting across the more familiar bioethical quartet of beneficence, non-maleficence, autonomy, and justice. The European Commission's High-Level Expert Group on Artificial Intelligence (2019) similarly organised its Ethics Guidelines for Trustworthy AI around seven requirements, including human agency, technical robustness, transparency, and accountability. What unites this literature is an observation that has since become a common critique: principles alone do not constrain behaviour unless they are accompanied by mechanisms of implementation and enforcement (Jobin et al., 2019).

2.2 Algorithmic Accountability and Opacity

A parallel and more implementation-focused literature has examined how accountability can be built into algorithmic systems themselves. Diakopoulos (2016) frames algorithmic accountability as the practice of assigning responsibility for harms that arise from an algorithm's decisions and identifies reverse-engineering, auditing, and disclosure as the principal investigative tools available to those seeking to hold algorithmic systems to account. Pasquale (2015) documents how proprietary and regulatory secrecy combine to produce a "black box society" in which the logics governing credit, reputation, and search are largely inaccessible to those they affect. Ananny and Crawford (2018) push back against an uncritical faith in transparency as a solution, arguing that visibility into a system's code or data does not, on its own, produce accountability; what matters is whether affected parties have the practical capacity to observe, understand, interrogate, and intervene in an algorithmic system's decisions. Mittelstadt et al. (2016) similarly map the ethical debate around algorithms onto six recurring concerns — inconclusive, inscrutable, and misguided evidence; unfair outcomes; transformative effects; and traceability — several of which recur directly in the Indian cases examined below. At a more applied level, Selbst et al. (2019) caution that many proposed technical fixes for algorithmic fairness fail because they abstract away the social and institutional context in which a system is deployed, a caution with direct relevance for administrative systems embedded in India's highly heterogeneous local governance contexts.

2.3 Algorithmic Harm in Practice

A third strand of literature documents the lived consequences of algorithmic administration, particularly in welfare contexts. O'Neil (2016) coined the term "weapons of math destruction" to describe opaque, large-scale scoring systems that disproportionately harm the poor while evading meaningful challenge. Eubanks (2018), in a study of automated eligibility systems in the United States, shows how digitised welfare administration can reproduce and deepen existing patterns of surveillance and exclusion among low-income populations, a dynamic with clear parallels to India's experience with Aadhaar-linked welfare delivery.

2.4 The Indian Policy and Empirical Record

India's own policy engagement with AI governance begins with NITI Aayog's National Strategy for Artificial Intelligence (2018), which identified healthcare, agriculture, education, smart cities, and mobility as priority sectors and flagged privacy, security, and the absence of a formal data protection regime as key barriers to responsible adoption. This was followed by the two-part Responsible AI approach papers, the first setting out system-level and societal-level principles — including safety, equality, inclusivity, and accountability — for AI use in India (NITI Aayog, 2021a), and the second proposing regulatory, capacity-building, and "ethics by design" interventions to operationalize those principles across government and industry (NITI Aayog, 2021b). On the empirical side, a substantial body of field research has documented the accountability gaps that arise when biometric algorithmic authentication is layered onto welfare delivery. Drèze, Khalid, Khera, and Somanchi (2017), based on a household survey in rural Jharkhand, found exclusion errors as high as 20 percent in areas where biometric authentication was mandatory for every PDS transaction, with poor connectivity and biometric mismatch — particularly among elderly and manual-labour populations with worn fingerprints — as recurring causes. Khera (2019) extends this critique across multiple welfare schemes, arguing that the technical failures of Aadhaar-based authentication have in some documented instances contributed to the denial of essential entitlements, and that the institutional response of the Unique Identification Authority of India has tended toward denial rather than systematic measurement and correction of exclusion errors.

Taken together, this literature identifies a clear gap that motivates the present paper: while India's AI policy discourse has matured considerably at the level of stated principle, and while field research has documented in granular detail how existing algorithmic systems fail citizens in practice, there has been comparatively little systematic work connecting the two — that is, specifying what concrete design, procedural, and institutional mechanisms would need to be in place for NITI Aayog's responsible AI principles to be operationalized within the specific sectoral systems where accountability failures have already been documented.

2.5 Administrative Law and Bureaucratic Discretion

A related but distinct strand of scholarship, situated within administrative law and public administration theory rather than computer science, has long examined how discretion is structured, constrained, and made accountable within bureaucracies. Classical accounts of street-level bureaucracy emphasise that front-line officials routinely exercise interpretive discretion in applying rules to individual cases, and that accountability in such systems has historically rested on layered mechanisms — supervisory review, appeals processes, judicial review, and legislative oversight — that operate on human decision-makers whose reasoning can, at least in principle, be interrogated through testimony, file notes, or hearings. When algorithmic systems replace or structure this discretion, as in Aadhaar-linked PDS authentication, they do not simply automate an existing process; they substitute a form of reasoning that is opaque to conventional administrative-law tools of interrogation for one that, however imperfectly, could previously be reconstructed through the ordinary apparatus of bureaucratic record-keeping. This substitution is significant for the present paper because it suggests that operationalizing accountability for algorithmic public administration in India is not solely a technical or AI-governance problem; it also requires re-establishing, in coded form, the functional equivalents of supervisory review, appeal, and reasoned explanation that older administrative-law accountability structures assumed would exist. The ethics-by-design framework adopted in Section 3 is, in this sense, best understood as an attempt to rebuild

administrative accountability infrastructure for a bureaucracy that increasingly delegates first-order decisions to code.

3. Theoretical Framework: Operationalizing Ethics-by-Design

This paper adopts Dignum's (2019) tripartite distinction between ethics by design, ethics in design, and ethics for design as its organising framework. Ethics by design refers to the direct embedding of ethical reasoning into a system's technical behaviour — for instance, building fairness constraints or explainability outputs into a scoring model itself. Ethics in design refers to the methods and processes through which development teams evaluate and negotiate the ethical implications of a system as it is built — impact assessments, participatory design exercises, and red-teaming exercises fall into this category. Ethics for design refers to the external codes, certification schemes, and standards that structure the incentives of the organisations and individuals producing algorithmic systems in the first place.

Read alongside Dignum's earlier "ART" principles — accountability, responsibility, and transparency — this typology suggests that "coding accountability" cannot be reduced to a single technical fix. Accountability, in Dignum's formulation, is the capacity of a system and its operators to explain and justify decisions to those affected by them; responsibility is the existence of a traceable chain linking a system's outputs back to identifiable human and institutional decision-makers; and transparency is the capacity to inspect the data, logic, and governance underpinning a system's behaviour. None of these properties can be achieved through code alone — they require that technical design choices be nested within procedural and institutional structures capable of enforcing and acting on what the code reveals.

For the purposes of analysing Indian algorithmic public administration, this paper operationalizes the ethics-by-design paradigm as a four-layer model, summarised in Table 2 below and applied throughout the remainder of the paper:

First, the value-specification layer, at which the objectives and constraints a system must satisfy — for example, a maximum tolerable false-rejection rate for a biometric authentication system serving welfare beneficiaries — are made explicit before development begins. Second, the technical layer, at which those values are translated into concrete system properties: fairness-aware model design, offline or fallback authentication pathways, and machine-readable audit logs. Third, the procedural layer, at which human institutions monitor, review, and intervene in a system's operation — algorithmic impact assessments prior to deployment, periodic third-party audits, and defined human-in-the-loop checkpoints for consequential decisions. Fourth, the institutional layer, at which independent bodies with statutory authority — whether a data protection authority, an ombudsperson, or a legislative committee — provide external oversight and a channel for grievance redress that does not depend on the goodwill of the implementing agency itself.

The central analytical claim of this paper is that accountability failures in Indian algorithmic administration to date have tended to occur precisely where this layered chain is broken — most often at the procedural and institutional layers, even where technical design has been reasonably sound.

3.1 The Limits of a Purely Technical Reading of Ethics-by-Design

It is important to be precise about what the ethics-by-design paradigm can and cannot deliver. A narrow, purely technical reading of the term — in which "coding accountability" is understood as writing fairness constraints into a model's objective function or adding an explainability module to its output — risks reproducing the very abstraction error that Selbst et al. (2019) warn against: treating a sociotechnical problem as though it could be resolved entirely within the boundaries of the technical artefact, independent of the institutional and social context into which it is deployed. A biometric authentication algorithm can,

in principle, be tuned to minimise false rejections; but whether a false rejection that nonetheless occurs results in a beneficiary going hungry for a month, as documented in the Jharkhand cases examined by Drèze et al. (2017), depends entirely on whether a procedural fallback exists and whether an institutional actor is responsible for ensuring that fallback functions. This paper therefore treats the technical layer of its four-layer framework as necessary but insufficient, and insists throughout on tracing each technical design choice through to its procedural and institutional consequences, rather than treating technical mitigation as the endpoint of the analysis.

4. Methodology

This paper adopts a qualitative, document-based policy analysis design combined with a structured comparative case method. Primary sources comprise India's principal AI governance documents — the National Strategy for Artificial Intelligence (NITI Aayog, 2018) and the two-part Responsible AI approach papers (NITI Aayog, 2021a, 2021b) — alongside the European Union's proposed Artificial Intelligence Act (European Commission, 2021) and the OECD's Recommendation on Artificial Intelligence (OECD, 2019), used as comparative regulatory benchmarks. Secondary sources comprise peer-reviewed and policy-oriented academic literature on algorithmic accountability and on the specific Indian sectoral systems examined.

Three sectoral cases were purposively selected to represent distinct administrative functions in which algorithmic decision-making has become consequential in India: welfare delivery (Aadhaar-linked authentication in the Public Distribution System), law enforcement (predictive-policing and facial recognition pilots undertaken by state police forces), and financial inclusion (algorithmic credit scoring associated with government-backed digital lending and account-aggregator initiatives). These cases were chosen because each has generated a documented public record — whether through field research, investigative reporting, or government disclosure — sufficient to support an accountability-focused analysis, and because together they span high-stakes decisions affecting subsistence, liberty, and economic opportunity. The analysis proceeds by mapping each case against the four-layer ethics-by-design framework developed in Section 3, identifying at which layer(s) accountability mechanisms are present, absent, or only nominally specified. This is followed by a comparative reading against the EU and OECD frameworks to identify transferable regulatory design elements. The paper does not undertake primary fieldwork or original data collection; its contribution is synthetic and conceptual, organising an existing but fragmented empirical and policy record into an operational framework for accountability.

4.1 Scope and Limitations

This paper's scope is deliberately bounded in two respects. First, it focuses on accountability mechanisms rather than on the equally important but analytically distinct question of algorithmic bias measurement, which has generated its own extensive technical literature and would warrant separate treatment. Second, because the underlying algorithmic systems examined — particularly the predictive-policing pilots discussed in Section 5.2 — are subject to limited public disclosure, the analysis relies on the available secondary and journalistic record rather than on direct technical inspection of proprietary systems; this is itself a methodological illustration of the opacity problem the paper investigates, and the paper treats the difficulty of obtaining primary technical detail as a finding in its own right rather than as a limitation to be apologised away. Where the public record is insufficiently detailed to support a specific empirical claim, this paper describes the accountability structure such a system would require in principle rather than asserting facts about a specific deployment that cannot be independently verified.

5. Algorithmic Public Administration in India: Sectoral Case Contexts

5.1 Welfare Delivery: Aadhaar-Linked Authentication in the PDS

The Aadhaar-enabled Public Distribution System requires beneficiaries to authenticate biometrically — typically via fingerprint — at the point of sale before receiving subsidised food grains. The stated objective is to eliminate "ghost" and duplicate beneficiaries and reduce diversion of rations. Field research in Jharkhand found exclusion errors as high as 20 percent in villages where biometric authentication was compulsory for every transaction, driven substantially by connectivity failures and by fingerprint mismatches among elderly beneficiaries and manual labourers (Drèze et al., 2017). Khera (2019) documents that the institutional response to such findings has generally been to dispute the scale of the problem rather than to establish a systematic mechanism for measuring exclusion error rates and correcting them. From an accountability standpoint, the system's technical layer (biometric matching) has received considerable design attention, but the procedural layer — mechanisms for a beneficiary to promptly contest and remedy an authentication failure — and the institutional layer — an independent body empowered to investigate exclusion patterns — remain comparatively underdeveloped.

5.2 Law Enforcement: Predictive Policing and Facial Recognition Pilots

Several Indian state police forces have piloted algorithmic tools intended to support crime prediction, suspect identification, or crowd monitoring, typically procured from private technology vendors under limited public disclosure. Because the underlying models, training data, and accuracy benchmarks of these systems are rarely published, independent scrutiny of their error rates or demographic performance disparities is difficult, echoing the opacity concerns raised generally in the accountability literature (Pasquale, 2015; Ananny & Crawford, 2018). Unlike the welfare case, where the affected population and the nature of the harm (denial of rations) is relatively well documented, predictive-policing deployments in India present a more acute accountability challenge precisely because the value-specification layer — what error rate, and what distribution of errors across communities, is considered tolerable — is rarely made public before deployment, let alone subjected to procedural review.

5.3 Financial Inclusion: Algorithmic Credit Scoring

India's Jan Dhan–Aadhaar–Mobile (JAM) infrastructure has enabled a rapid expansion of digital lending products that rely on algorithmic creditworthiness assessment, often incorporating alternative data such as mobile usage or digital transaction history. These systems raise accountability questions distinct from the welfare and policing cases: because credit decisions are frequently intermediated by private non-banking financial companies and fintech platforms operating under a lighter regulatory touch than public agencies, the institutional layer of oversight is fragmented across the Reserve Bank of India, sectoral regulators, and individual lenders' internal policies, with no single body currently empowered to conduct algorithmic audits across the sector.

Across all three sectors, a common pattern emerges: the value-specification and technical layers of the ethics-by-design framework introduced in Section 3 have, to varying degrees, received design attention — authentication algorithms are tuned for accuracy, predictive-policing models are trained on available crime data, and credit-scoring models are validated against historical repayment data. What is far more consistently absent is documentation, at the point of deployment, of the tolerances and trade-offs these systems were designed around — what error rate was considered acceptable, for whom, and what recourse exists when that tolerance is exceeded in an individual case. This absence is precisely what converts a technically competent system into an administratively unaccountable one, and it is this gap that Section 6 seeks to address.

Table 1- *Sectoral Cases of Algorithmic Public Administration in India*

Sector	Algorithmic Tool	Stated Objective	Principal Accountability Gap
Welfare delivery (PDS)	Aadhaar biometric authentication	Eliminate duplicate/ghost beneficiaries	Weak procedural fallback and grievance redress for authentication failures (Drèze et al., 2017; Khera, 2019)
Law enforcement	Predictive policing / facial recognition (state pilots)	Crime prediction; suspect and crowd identification	Undisclosed model logic, training data, and accuracy benchmarks (Pasquale, 2015)
Financial inclusion	Algorithmic credit scoring via JAM infrastructure	Expand access to formal credit for underserved borrowers	Fragmented regulatory oversight across private lenders (RBI/sectoral gap)

Note. Compiled by the author from NITI Aayog (2018, 2021a) and the sources cited in each row.

6. Operationalizing Accountability: A Layered Mechanism Framework

Building on the four-layer model introduced in Section 3, this section proposes a set of concrete mechanisms through which India's stated responsible AI principles could be operationalized in the three sectoral contexts examined above.

6.1 Algorithmic Impact Assessments

An algorithmic impact assessment (AIA) is a structured, pre-deployment evaluation of a system's likely effects, intended outcomes, and risk profile, conducted before a system is put into production and revisited at defined intervals thereafter. Canada's Directive on Automated Decision-Making (Treasury Board of Canada Secretariat, 2019) offers a template already tested in a large federal bureaucracy: it assigns systems to one of four impact tiers based on the stakes of the decisions they inform, with the depth of required review — up to and including independent peer review — scaling with tier. Applied to India, an AIA requirement would compel the value-specification layer of the ethics-by-design framework to be documented and made available for scrutiny before a welfare-authentication or predictive-policing system is deployed at scale, rather than being inferred retrospectively from field research after harm has already occurred.

6.2 Mandatory Audit Trails and Logging

At the technical layer, systems used in consequential administrative decisions should be required to generate machine-readable logs of key decision points — including authentication failures, override events, and the specific data inputs relied upon — sufficient to reconstruct after the fact why a particular decision was reached. Such logging directly enables the kind of reverse-engineering and disclosure-based accountability that Diakopoulos (2016) identifies as central to algorithmic accountability practice, and

would allow the exclusion-error patterns documented by Drèze et al. (2017) to be identified through routine administrative data rather than through resource-intensive independent field surveys.

6.3 Calibrated Human Oversight

Not all algorithmic decisions warrant the same intensity of human oversight. This paper distinguishes between human-in-the-loop arrangements, in which a human reviews and can override each individual algorithmic recommendation before it takes effect, and human-on-the-loop arrangements, in which a human monitors aggregate system performance and intervenes only on exception or at defined review intervals. For high-stakes, low-volume decisions — such as a predictive-policing flag leading to targeted surveillance of an individual — a human-in-the-loop requirement is appropriate. For high-volume, comparatively lower-stakes decisions — such as routine PDS authentication — a human-on-the-loop model combined with a guaranteed, fast, offline fallback pathway (for instance, the alternative authentication mechanisms already piloted in Madhya Pradesh) better balances administrative feasibility with the need to prevent individual harm.

6.4 Independent Grievance and Oversight Mechanisms

At the institutional layer, this paper argues that the single most consequential missing element in current Indian algorithmic administration is an independent body — whether a dedicated algorithmic accountability cell, an ombudsperson, or a function embedded within a future data protection authority — empowered to receive complaints, compel disclosure of system logs and impact assessments, and order corrective action, independent of the implementing agency itself. This addresses precisely the pattern Khera (2019) identifies, in which the body responsible for a system (the Unique Identification Authority of India, in the Aadhaar case) is also the primary arbiter of whether that system is failing.

6.5 Calibrated Explainability Requirements

A final design-level mechanism concerns explainability — the capacity of a system to produce, for a given decision, a reason that is comprehensible to the person affected by it. As Wachter et al. (2017) demonstrate in the European context, explainability obligations cannot be assumed to arise automatically from broader data protection or privacy legislation; they must be explicitly specified, and specified at a level of technical and administrative detail sufficient to be enforceable. This paper argues that explainability requirements for Indian algorithmic public administration should be calibrated rather than uniform: a beneficiary denied a PDS transaction needs a short, actionable explanation — for instance, that a fingerprint match failed and that an alternative authentication pathway is available at a specified location — rather than a technical account of the underlying matching algorithm, whereas a researcher, auditor, or the institutional oversight body proposed in Section 6.4 requires access to the full technical logic and performance statistics of the system. Building this two-tier explainability requirement — a citizen-facing layer and an auditor-facing layer — directly into system specifications at the value-specification stage is, in the framework developed here, itself an act of coding accountability, in the literal sense of the term.

Table 2- *A Layered Mechanism Framework for Operationalizing Accountability*

Layer	Mechanism	Lifecycle Stage	Current Status in India
Value specification	Algorithmic impact assessment	Pre-deployment	Not statutorily mandated; recommended in

Layer	Mechanism	Lifecycle Stage	Current Status in India
			principle (NITI Aayog, 2021b)
Technical	Mandatory audit logging; fairness-aware model design	Development / deployment	Ad hoc; varies by implementing agency and vendor
Procedural	Calibrated human-in-the-loop / on-the-loop review; offline fallback	Deployment / operation	Partial (e.g., select state PDS fallback pilots); not uniform
Institutional	Independent oversight body / algorithmic ombudsperson	Ongoing	Absent; grievance handling largely internal to implementing agency

Note. Author's synthesis, informed by Dignum (2019), Treasury Board of Canada Secretariat (2019), and NITI Aayog (2021a, 2021b).

7. Comparative Perspective: India and Global Regulatory Models

India's Responsible AI approach papers were published against the backdrop of two major international regulatory developments: the OECD's Recommendation on Artificial Intelligence (OECD, 2019), the first intergovernmental standard on AI adopted by OECD member and partner countries including India, and the European Commission's proposed Artificial Intelligence Act (European Commission, 2021), which introduces a risk-tiered regulatory model imposing binding obligations — including mandatory conformity assessments — on AI systems classified as "high-risk."

The EU's risk-tiered approach is instructive for the Indian context because it directly operationalizes the value-specification layer discussed in Section 3: rather than applying a uniform standard to all algorithmic systems, it calibrates the intensity of required oversight to the stakes of the decision at hand, placing systems used in areas such as law enforcement, employment, and essential public services in its highest-scrutiny category — a category that would plainly encompass the PDS authentication and predictive-policing cases examined in Section 5. The OECD Recommendation, by contrast, remains principle-based and non-binding, closer in character to NITI Aayog's current approach papers than to the EU's proposed regulation. Wachter et al. (2017) further caution, in an EU-focused analysis directly relevant to any Indian data protection legislation modelled on the GDPR, that a general "right to explanation" cannot be assumed to follow automatically from data protection law and must instead be explicitly and specifically legislated if it is to be enforceable.

Table 3- *Comparative Regulatory Approaches to Algorithmic Accountability*

Dimension	European Union (Proposed AI Act)	OECD Recommendation	India (NITI Aayog Approach)
Legal status	Binding regulation (proposed)	Non-binding recommendation	Non-binding policy guidance
Risk calibration	Explicit risk tiers with differentiated obligations	General principles, no formal tiering	General principles; sectoral tiering not yet formalised
Independent oversight	National supervisory authorities; conformity assessment	Left to member states	Not yet established
Audit/impact assessment	Mandatory for high-risk systems	Encouraged, not mandated	Recommended (NITI Aayog, 2021b), not mandated

Note. Compiled by the author from European Commission (2021), OECD (2019), and NITI Aayog (2021a, 2021b).

7.1 Why Direct Transplantation Is Insufficient

The comparative reading offered above should not be mistaken for an argument that India can simply transplant the EU's risk-tiered model or the OECD's principle-based approach wholesale. The EU's proposed Artificial Intelligence Act presupposes a mature, well-resourced regulatory ecosystem — including existing sectoral regulators, established conformity-assessment infrastructure, and a General Data Protection Regulation already in force — that supplies much of the institutional scaffolding the Act's obligations rely upon. India, by contrast, is simultaneously building its data protection framework, its sectoral regulatory capacity, and its algorithmic governance apparatus, with limited sequencing between the three. The OECD Recommendation, for its part, is explicitly designed to accommodate this kind of institutional heterogeneity across its signatories, which is precisely why it remains non-binding and principle-based rather than prescriptive. The comparative value of Table 3, in this paper's reading, lies not in identifying a single model for direct adoption but in isolating specific design elements — risk tiering, mandatory conformity assessment for high-risk uses, and a named independent supervisory function — that could be sequenced into India's existing responsible AI approach papers even in advance of comprehensive data protection legislation, for instance through sector-specific executive guidance issued by MeitY or the Reserve Bank of India within their existing statutory authority.

8. Challenges and Structural Barriers

Several structural features of Indian governance complicate the operationalization of ethics-by-design principles beyond the technical and design considerations discussed above.

First, India's federal structure divides jurisdiction over the sectors most affected by algorithmic administration — policing and, to a significant degree, welfare implementation are state subjects — meaning that any centrally issued accountability standard, including those recommended in NITI Aayog's

approach papers, must be adopted and enforced unevenly across states with widely varying technical and institutional capacity.

Second, at the time of this paper's preparation, India's comprehensive data protection legislation remains under parliamentary consideration, leaving the institutional layer of the accountability framework proposed in Section 6 without a clear statutory anchor; absent a data protection authority or equivalent body with investigative and enforcement powers, the independent oversight function this paper identifies as critical has no obvious institutional home.

Third, technical audit capacity within government remains limited relative to the scale and complexity of the systems being deployed, and much of the underlying technical work is contracted to private vendors whose proprietary claims over model architecture and training data can themselves become a barrier to the kind of disclosure-based accountability that Diakopoulos (2016) and Pasquale (2015) identify as foundational.

Fourth, mechanisms of citizen-driven accountability that have historically been effective in the Indian context — most notably the Right to Information framework — were designed for a bureaucracy of documents and files, and their applicability to requests for algorithmic model logic, training data, or audit logs remains legally and practically untested at scale, leaving affected communities with fewer well-established tools than field researchers such as Drèze et al. (2017) have had to rely on in documenting harm after the fact.

Fifth, much of the empirical evidence on algorithmic harm in India cited throughout this paper — the Jharkhand exclusion-error findings, for instance — has been produced through resource-intensive, one-off household surveys conducted by independent researchers rather than through routine administrative monitoring. This is itself symptomatic of the accountability gap the paper is concerned with: in a well-instrumented system, the audit-logging mechanism proposed in Section 6.2 would make such findings a matter of ordinary administrative reporting rather than of periodic academic fieldwork, and the fact that field survey remains the primary evidentiary basis for understanding these systems' real-world performance is itself evidence of how far current practice sits from an operationalized ethics-by-design standard.

Sixth, procurement practices constitute an underappreciated barrier in their own right. Where algorithmic systems are acquired through standard public procurement processes designed for conventional IT infrastructure, contracts frequently do not require vendors to disclose model architecture, provide audit access, or commit to post-deployment performance reporting, since these were not contractual categories contemplated when procurement templates were drafted. Retrofitting accountability into already-deployed systems is consequently far more difficult, and more legally contested, than building such requirements into procurement contracts and tender specifications from the outset — a point with direct implications for the phased recommendations set out in Section 9.

9. Policy Recommendations

Drawing on the analysis above, this paper proposes a phased approach to operationalizing ethics-by-design accountability in Indian algorithmic public administration.

First, algorithmic impact assessments should be mandated, in the first instance, for the highest-stakes systems already identified in this paper — Aadhaar-linked welfare authentication and predictive-policing deployments — before being extended to a broader class of systems, following the tiered logic of the EU's proposed Artificial Intelligence Act (European Commission, 2021).

Second, mandatory audit-logging requirements should be attached to any system used to determine access to welfare entitlements, drawing on the technical logging already feasible within existing DBT and PDS digital infrastructure, so that exclusion-error patterns can be identified through routine data rather than ad hoc field survey.

Third, an Algorithmic Accountability Cell should be established, whether within NITI Aayog, the Ministry of Electronics and Information Technology, or the data protection authority contemplated by pending legislation, with the specific statutory mandate to receive grievances, compel disclosure of impact assessments and audit logs, and order remedial action independent of the implementing agency.

Fourth, capacity-building partnerships between government agencies and academic and civil society institutions should be prioritised to build the technical audit expertise identified as currently insufficient in Section 8, following the model of embedded technical fellowships already used in some Indian digital public infrastructure initiatives.

Fifth, any new algorithmic system used in a high-stakes administrative decision should be required to guarantee a functioning, accessible, non-algorithmic fallback pathway for affected citizens, addressing directly the exclusion-error dynamic documented by Drèze et al. (2017) and Khera (2019), and should be subject to a defined sunset or mandatory review clause rather than open-ended deployment.

Taken together, these five recommendations are intended to be sequenced rather than simultaneous. Algorithmic impact assessments and audit-logging requirements (recommendations one and two) can be introduced administratively, through executive circulars and revised procurement templates, without waiting for new primary legislation, and should therefore be prioritised as immediate, low-cost first steps. The Algorithmic Accountability Cell and its statutory grievance-redress powers (recommendation three) are more consequential but can plausibly be sequenced alongside the passage of comprehensive data protection legislation already under parliamentary consideration, rather than requiring an entirely separate legislative vehicle. Capacity-building partnerships and mandatory fallback and sunset provisions (recommendations four and five) are lower-cost and can proceed in parallel with both. This sequencing logic reflects the paper's broader theoretical claim: that the technical and value-specification layers of ethics-by-design are, in the Indian context, more tractable in the near term than the institutional layer, but that lasting accountability cannot be achieved without eventually closing the institutional gap as well.

10. Conclusion

India's engagement with the ethics of artificial intelligence has matured considerably at the level of stated principle, as reflected in NITI Aayog's National Strategy for Artificial Intelligence and its subsequent Responsible AI approach papers. What remains underdeveloped is the operational architecture required to translate those principles into the design, procedural, and institutional practice of specific administrative systems. This paper has argued that ethics-by-design, understood as a layered framework spanning value specification, technical design, procedural review, and institutional oversight, offers a useful lens through which to identify precisely where this translation currently breaks down — most consistently, this paper finds, at the procedural and institutional layers rather than at the level of technical design itself.

The sectoral cases examined — Aadhaar-linked welfare authentication, predictive policing, and algorithmic credit scoring — each illustrate a variant of the same underlying problem: systems introduced to reduce administrative discretion and error have, in the absence of adequate accountability infrastructure, produced new and often less visible forms of harm and unaccountability. Coding accountability into India's algorithmic public administration will require more than better algorithms; it will require algorithmic

impact assessments with real consequences, audit trails that are actually audited, human oversight calibrated to the stakes of each decision, and an independent institutional home for grievances that does not depend on the implementing agency policing itself. Until these procedural and institutional layers are built out and given statutory force, the accountability commitments articulated in India's responsible AI discourse will remain, for the citizens most affected by these systems, largely aspirational.

Future research could usefully extend this analysis in at least two directions. First, as India's data protection legislation is finalised and, potentially, sector-specific algorithmic accountability rules are issued, empirical work tracking whether the layered mechanisms proposed here — impact assessments, audit logging, calibrated oversight, and independent grievance redress — are in fact adopted, and with what effect on measured exclusion and error rates, would provide a direct test of this paper's central claims. Second, comparative work examining how other large federal democracies with significant administrative capacity constraints, rather than the comparatively well-resourced European Union, have approached algorithmic accountability in welfare and policing contexts would offer a more directly transferable set of institutional design lessons than the EU comparison this paper has been able to offer. What this paper has sought to establish, in the interim, is that the vocabulary of ethics-by-design is of real analytical value for Indian algorithmic public administration only once it is disaggregated into the specific, auditable mechanisms — and the specific, accountable institutions — required to give it operational force.

REFERENCES:

1. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. New Media & Society, 20(3), 973–989. https://doi.org/10.1177/1461444816676645
2. Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110
3. Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way. Springer.
4. Drèze, J., Khalid, N., Khera, R., & Somanchi, A. (2017). Aadhaar and food security in Jharkhand: Pain without gain? Economic and Political Weekly, 52(50), 50–59.
5. Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.
6. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM(2021) 206 final.
7. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
8. High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. European Commission.
9. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

10. Khera, R. (2019, April 6). Aadhaar failures: A tragedy of errors. Economic and Political Weekly, 54(14).
11. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21. https://doi.org/10.1177/2053951716679679
12. NITI Aayog. (2018). National strategy for artificial intelligence: #AIforAll. Government of India.
13. NITI Aayog. (2021a). Responsible AI: Approach document for India, Part 1—Principles for responsible AI. Government of India.
14. NITI Aayog. (2021b). Responsible AI: Approach document for India, Part 2—Operationalizing principles for responsible AI. Government of India.
15. O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.
16. OECD. (2019). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449).
17. Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.
18. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 59–68. https://doi.org/10.1145/3287560.3287598
19. Treasury Board of Canada Secretariat. (2019). Directive on automated decision-making. Government of Canada.
20. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law, 7(2), 76–99. https://doi.org/10.1093/idpl/ipx005