

From Watchtowers to Welfare: The Ethical Trade-offs of AI Surveillance in India's Sustainable Governance Agenda

Lt. Dr. Kongala Sukumar

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

Abstract:

India's public administration is undergoing a rapid, technology-led transformation in which artificial intelligence (AI) is deployed simultaneously as an instrument of welfare delivery and as a mechanism of population surveillance. Biometric identification, facial recognition, predictive policing, and algorithmic targeting of beneficiaries have converged into what this paper terms a "watchtower-welfare" governance model, in which the state's capacity to see its citizens is inseparable from its capacity to serve them. Drawing on Bentham's panopticon and Foucault's theory of disciplinary power, this paper develops a conceptual framework for analysing the ethical trade-offs embedded in AI-enabled governance, and situates these trade-offs within India's sustainable governance agenda and its commitments under the United Nations 2030 Agenda for Sustainable Development. Using documentary analysis of policy instruments, case evidence on Aadhaar-linked welfare exclusion, facial recognition deployments, and smart city surveillance infrastructure, the paper identifies four recurring tensions: privacy versus service delivery, efficiency versus equity, security versus liberty, and centralisation versus accountability. The paper concludes that sustainable governance in the AI era requires embedding proportionality, purpose limitation, and participatory oversight into surveillance-linked welfare architecture, rather than treating ethics as an afterthought to technological deployment.

Keywords: Artificial intelligence, surveillance, sustainable governance, welfare state, Aadhaar, algorithmic ethics, India, public administration.

1. Introduction

The Indian state has, over the past decade, positioned artificial intelligence at the centre of its public administration reform agenda, promoting it as a tool for inclusive growth, efficient service delivery, and evidence-based policymaking (NITI Aayog, 2018). At the same time, the infrastructure that makes this vision possible—biometric identity databases, facial recognition systems, integrated command-and-control centres, and algorithmic targeting mechanisms—constitutes one of the largest surveillance apparatuses assembled by any democratic state. The coexistence of these two functions within a single technological architecture is not incidental. Systems built to verify identity for the purpose of distributing subsidised grain, cooking gas, or pension payments are the same systems that can, and do, track movement,

flag "suspicious" behaviour, and enable real-time monitoring of protest and dissent (Khera, 2019). This paper examines that coexistence directly, asking what ethical trade-offs are produced when a state's capacity for care and its capacity for control are built on the same digital foundation.

The metaphor in the title, watchtowers to welfare, is intended to capture this duality. The watchtower is Bentham's (1791) architectural device for the panopticon prison, later theorised by Foucault (1977) as a general model of disciplinary power: visibility, in Foucault's reading, is a trap, because the watched subject internalises discipline even in the absence of an observer. Welfare, in contrast, evokes the promise of a caring, redistributive state—India's constitutional commitment to social and economic justice, operationalised through schemes such as the Public Distribution System (PDS), the Mahatma Gandhi National Rural Employment Guarantee Act (MGNREGA), and direct benefit transfers. AI-enabled governance in India increasingly fuses these two logics: the delivery of welfare is made conditional on biometric visibility, and biometric visibility, once established for welfare purposes, becomes available for other governmental uses, including policing and security (Marda, 2018). This fusion is often described in policy discourse as efficient, transparent, and leakage-proof; this paper argues that it should also be understood as an ethically fraught reorganisation of the relationship between citizen and state.

The stakes of this inquiry are heightened by India's parallel commitments to sustainable development. As a signatory to the 2030 Agenda, India has framed digital public infrastructure as a vehicle for achieving Sustainable Development Goal (SDG) 16, which calls for peaceful, just, and inclusive institutions, as well as SDG 1, 2, 3, and 10, which concern poverty, hunger, health, and inequality (United Nations, 2015). Yet sustainable governance, properly understood, is not merely a matter of technical efficiency in reaching development targets; it requires that the institutions used to pursue those targets themselves be just, accountable, and rights-respecting. A welfare system that reduces hunger while entrenching mass surveillance and exclusion errors achieves one dimension of sustainability while undermining another. This paper therefore treats the ethics of AI surveillance not as a side issue to India's sustainable governance agenda but as constitutive of it.

Methodologically, the paper adopts a qualitative, documentary approach, synthesising primary policy instruments issued by NITI Aayog and the Unique Identification Authority of India, judicial reasoning from the Supreme Court's privacy jurisprudence, and independent monitoring data compiled by civil society organisations, most notably the Internet Freedom Foundation's Project Panoptic tracker and published field research on Aadhaar-linked welfare exclusion. This approach is well suited to a research question that is fundamentally normative and institutional rather than statistical: the paper is not attempting to estimate the precise incidence of exclusion or misidentification across India, a task better suited to large-scale household survey research, but to characterise the structure of the trade-offs that AI-enabled governance produces, and to evaluate that structure against India's own stated sustainable governance commitments. Case material is drawn from welfare administration, policing, and urban governance because these three domains represent the points at which biometric and surveillance infrastructure most directly intersects with individual citizens' access to rights and entitlements.

The paper proceeds as follows. Section two develops a theoretical framework linking classical theories of disciplinary power to contemporary scholarship on surveillance capitalism and algorithmic governance. Section three maps the landscape of AI-enabled surveillance in Indian public administration, drawing on documented cases from welfare administration, policing, and urban governance. Section four examines the paradox of Aadhaar-linked welfare delivery, in which the pursuit of inclusion has produced new forms

of exclusion. Section five presents a structured matrix of ethical trade-offs, and section six maps these trade-offs against specific SDG targets. Section seven reviews the regulatory and institutional response to date, section eight offers recommendations for embedding ethical safeguards into India's AI governance architecture, and section nine concludes.

2. Theoretical Framework: From Panopticon to Datafied Governance

The analytical starting point for this paper is Jeremy Bentham's (1791) design for the panopticon, an institutional architecture in which a central watchtower allows a single observer to potentially see into every cell of a prison, while the prisoners themselves cannot verify when they are being watched. Michel Foucault (1977) extended this architectural principle into a general theory of modern power, arguing that the panopticon's genius lies not in constant surveillance itself but in the permanent possibility of surveillance, which induces self-discipline in the observed subject regardless of whether observation is actually occurring at any given moment. Foucault's insight travels well into the digital era: citizens enrolled in biometric identity systems, tracked by facial recognition cameras in public spaces, or scored by algorithmic welfare-eligibility systems cannot verify, at any given moment, whether or how they are being observed, matched, or flagged. The result is a diffuse, internalised form of governance that operates independently of any single instance of active monitoring.

Contemporary scholarship has extended this analysis to account for the specific characteristics of data-driven governance. Zuboff (2019) describes the extraction and monetisation of behavioural data as "surveillance capitalism," a logic in which human experience is rendered as data for the purpose of prediction and control; while Zuboff's focus is primarily on private technology firms, the underlying mechanism—the conversion of everyday activity into machine-readable, actionable data—applies equally to state surveillance conducted through welfare and security infrastructure. O'Neil (2016) contributes a complementary concern with opacity and scale, arguing that algorithmic systems used in high-stakes public decisions frequently function as "weapons of math destruction" precisely because their scoring logic is invisible to the people it affects, resistant to appeal, and capable of affecting large populations uniformly and simultaneously. Applied to Indian welfare administration, this means that a single algorithmic misclassification—for instance, a failed biometric authentication treated as evidence of an invalid claim—can be replicated across millions of transactions before it is identified or corrected (Dreze, Khalid, Khera, & Somanchi, 2017).

Together, Foucault's disciplinary framework and Zuboff's account of surveillance capitalism suggest that the significance of AI-enabled governance lies less in any single deployment than in its cumulative, infrastructural character: once biometric enrolment, facial recognition, and integrated command centres become the default substrate of public administration, the state's relationship to its citizens is reorganised at a structural level, in ways that are difficult to reverse even where individual applications are later curtailed or discontinued. This infrastructural permanence is a further reason to treat the ethical evaluation of AI governance as a matter of design and institutional choice from the outset, rather than as a corrective to be applied after deployment. A third theoretical strand relevant to this paper concerns the ethics of AI as a domain of applied ethics in its own right. Floridi et al. (2018) propose a framework organised around beneficence, non-maleficence, autonomy, justice, and explicability, arguing that AI systems deployed in the public interest must be evaluated not only by their aggregate outcomes but by their distributive effects and their respect for individual agency. This framework is useful because it resists a purely utilitarian

assessment of AI-enabled governance—an assessment that would ask only whether a system reduces leakage or improves targeting accuracy on average—and instead demands attention to who bears the costs of error, opacity, and non-consensual visibility. This paper draws on all three strands—Foucauldian disciplinary theory, surveillance capitalism, and AI ethics frameworks—to construct the trade-off matrix presented in section five.

A further implication of the Foucauldian framework, often underemphasised in policy discourse, concerns the relationship between visibility and eligibility. In the disciplinary institutions Foucault (1977) analysed—prisons, schools, hospitals—visibility was a precondition for normalisation and correction. In India's welfare-surveillance architecture, visibility has become a precondition for existence within the administrative record: a citizen who cannot be biometrically verified is not merely unseen but, for practical purposes, ineligible. This is a subtle but important departure from classical disciplinary power, in which the subject's existence was never in question, only their conduct. Algorithmic welfare systems compress the boundary between being governed and being recognised as governable, such that a technical failure of recognition becomes indistinguishable, in its effects, from a substantive denial of rights (Khera, 2019). This compression is the analytical thread that connects the theoretical discussion in this section to the empirical material examined in sections three and four.

3. The Landscape of AI-Enabled Surveillance in Indian Governance

India's AI-enabled governance architecture has developed cumulatively since the early 2010s, beginning with the Unique Identification Authority of India's (UIDAI) biometric enrolment programme, popularly known as Aadhaar, and expanding into facial recognition, predictive policing, and integrated city surveillance. The 2018 National Strategy for Artificial Intelligence, published by the government's policy think tank NITI Aayog, explicitly identified smart cities and infrastructure, alongside healthcare, agriculture, and education, as priority sectors for AI deployment, citing gains in efficiency and the potential to address societal needs at scale (NITI Aayog, 2018). This strategic framing has translated into a proliferation of surveillance-capable systems operating under the banner of governance modernisation, several of which are summarised in Table 1.

Facial recognition technology (FRT) offers the clearest illustration of the ambiguity between welfare-oriented and security-oriented deployment. The National Crime Records Bureau has pursued a nationwide Automated Facial Recognition System intended to assist criminal investigation, while state and city police forces have independently deployed FRT for crowd monitoring, missing-persons identification, and protest surveillance (Internet Freedom Foundation, 2020). Civil society monitoring by the Internet Freedom Foundation's Project Panoptic identified dozens of active or planned FRT projects operated by state and city police departments as of the early 2020s, most functioning without a dedicated legal framework governing their use, retention periods, or accuracy thresholds (Internet Freedom Foundation, 2020). In Hyderabad, a large-scale "Command and Control Centre," reported to link hundreds of thousands of surveillance cameras with facial recognition capability, was documented by Amnesty International, Article 19, and the Internet Freedom Foundation as enabling real-time tracking of individuals across the city (Amnesty International, Article 19, & Internet Freedom Foundation, 2021). Such systems are frequently justified with reference to public safety and traffic management, yet their surveillance capacity extends well beyond these stated purposes, illustrating what surveillance scholars term function creep: the gradual expansion of a system's use beyond its original justification.

Predictive and preventive policing tools represent a further layer of this architecture. Several state police departments have piloted or adopted software intended to identify individuals or localities considered at elevated risk of criminal activity, drawing on historical crime records, informant reports, and, increasingly, integration with facial recognition and vehicle-tracking systems. Because such systems are trained on historical policing data, they risk reproducing existing patterns of discriminatory enforcement against marginalised communities, embedding past biases into ostensibly neutral, forward-looking predictions (O'Neil, 2016). Smart city missions, meanwhile, have normalised dense networks of CCTV cameras, environmental sensors, and integrated command centres across dozens of Indian cities, creating a baseline surveillance infrastructure that is, in principle, available for repurposing by law enforcement and other government agencies with minimal additional investment.

A further pattern worth noting is the tendency for biometric verification, once normalised in one domain of public life, to expand into adjacent domains that were not originally understood as welfare or security functions at all. Civil society review of NITI Aayog's 2022 discussion paper on facial recognition technology, for instance, highlighted the extension of facial-recognition-based verification into aviation, through schemes such as DigiYatra, alongside its established use in law enforcement and welfare administration (Internet Freedom Foundation, 2022). This expansion illustrates that the boundary between "welfare-oriented" and "security-oriented" biometric infrastructure is not stable; systems introduced for one stated purpose routinely become the technical and institutional foundation for others, often without a fresh assessment of proportionality or necessity for the new use case. The cumulative effect, across welfare administration, policing, and now travel and access infrastructure, is a governance environment in which biometric verification is increasingly treated as a default administrative procedure rather than an intervention that itself requires justification.

Table 1 summarises the principal categories of AI-enabled surveillance infrastructure currently operating within Indian public administration, indicating their stated governance purpose, their surveillance capacity, and the ethical concerns most frequently raised in relation to each.

Table 1 Major AI Surveillance Applications in Indian Public Administration

System / Programme	Governing Agency	Stated Governance Purpose	Surveillance Capacity	Key Ethical Concern
Aadhaar biometric identification	UIDAI, Ministry of Electronics & IT	Welfare targeting, subsidy delivery, de-duplication	Centralised biometric database, linkable across schemes	Authentication-based exclusion of legitimate beneficiaries
Automated Facial Recognition System (AFRS)	National Crime Records Bureau	Criminal investigation, missing-persons identification	Nationwide face-matching against police databases	No dedicated legal framework; misidentification risk

System / Programme	Governing Agency	Stated Governance Purpose	Surveillance Capacity	Key Ethical Concern
City FRT / CCTV networks (e.g., Hyderabad Command & Control Centre)	State and city police	Traffic management, public safety, crowd monitoring	Real-time tracking via extensive camera networks	Function creep into protest and dissent surveillance
Predictive policing tools	State police departments	Crime-pattern forecasting, resource allocation	Risk-scoring of individuals/localities from historical data	Reproduction of discriminatory enforcement patterns
Smart city command centres	Urban local bodies / Smart Cities Mission	Infrastructure and traffic management, service delivery	Integrated sensor and camera networks across cities	Repurposing of civic infrastructure for security use

Note. Compiled by the author from NITI Aayog (2018), Internet Freedom Foundation (2020, 2021), and Amnesty International, Article 19, & Internet Freedom Foundation (2021).

4. Surveillance and the Welfare State: Aadhaar and the Paradox of Inclusion

Nowhere is the entanglement of welfare and surveillance more consequential than in the Aadhaar biometric identification programme, which has become a near-universal precondition for accessing subsidised food grain, cooking gas subsidies, pension disbursement, and a wide range of other welfare entitlements. Aadhaar was designed, and is defended by the government, as a tool of inclusion: a single, portable, biometric identity intended to eliminate duplicate and fraudulent beneficiaries and ensure that entitlements reach the intended recipient (NITI Aayog, 2018). In practice, the technology's reliance on fingerprint and iris matching has produced a persistent pattern of authentication failure among precisely the population groups the welfare system is meant to serve: manual labourers whose fingerprints are worn or damaged by physical work, elderly beneficiaries whose biometric features degrade with age, and rural residents in areas with unreliable internet connectivity required for real-time verification (Dreze et al., 2017).

Field research conducted in Jharkhand by Dreze et al. (2017) found exclusion errors as high as twenty percent in villages where biometric authentication was mandatory for every transaction under the Public Distribution System, with documented cases of beneficiaries, including elderly and disabled residents, being denied rations or dying in circumstances linked to authentication failure. Subsequent research has extended these findings to healthcare access, documenting cases in which biometric authentication failures delayed or denied maternal and neonatal care for women in informal employment (Khera & Somanchi, 2021). The Unique Identification Authority of India has historically disputed the scale of these exclusion errors and has not published disaggregated, age- or occupation-based failure-rate data that would allow independent verification (Dreze et al., 2017). This absence of transparent, auditable failure data is itself

an ethical concern distinct from the exclusion errors themselves: a system that cannot account for its own failure modes cannot be meaningfully held accountable for them.

The Aadhaar paradox illustrates a broader structural tension in AI-enabled welfare governance: the same biometric infrastructure that promises to make the state more responsive to individual citizens also makes individual citizens more visible, and more procedurally vulnerable, to the state. A beneficiary who fails biometric authentication is not merely inconvenienced; she is rendered administratively invisible, unable to prove the very identity the system was built to verify. This is a distinctly algorithmic form of harm, in which the technology's design logic—treating non-match as evidence of fraud rather than as evidence of system failure—produces exclusion as a systemic by-product rather than an isolated bug (Khera, 2019). The 2017 Supreme Court judgment in Justice K.S. Puttaswamy (Retd.) v. Union of India recognised privacy as a fundamental right under the Indian Constitution and subjected biometric identification schemes to a proportionality test, requiring that any infringement of privacy be necessary, proportionate, and subject to procedural safeguards; the persistence of unresolved exclusion errors nearly a decade after Aadhaar's rollout suggests that this proportionality standard remains imperfectly realised in practice.

The inclusion-exclusion paradox is not confined to biometric authentication for the Public Distribution System; it recurs across other flagship welfare programmes that have incorporated digital verification and algorithmic targeting. Direct benefit transfer mechanisms, which route subsidies and wage payments directly into bank accounts linked to Aadhaar, similarly presuppose an uninterrupted chain of accurate biometric linkage, functioning bank infrastructure, and reliable connectivity; a break at any point in this chain—an unlinked account, a name mismatch, a failed authentication—can delay or deny payment to a beneficiary who otherwise meets every substantive eligibility criterion (Khera, 2019). What is analytically significant about this pattern is that the harm is procedural rather than substantive: the beneficiary is not deemed ineligible on the merits, but is rendered unable to demonstrate eligibility because the verification infrastructure itself has failed. A governance model that treats such procedural failures as rare edge cases, rather than as a predictable and recurring feature of biometric administration at national scale, is likely to systematically underestimate the human cost of its own efficiency gains.

5. Ethical Trade-offs: Mapping the Tensions

The preceding sections suggest that AI-enabled governance in India cannot be adequately evaluated along a single dimension of efficiency or accuracy. Rather, it produces a recurring set of trade-offs between values that a sustainable governance agenda is meant to hold simultaneously: privacy, equity, security, liberty, transparency, and accountability. Table 2 organises these trade-offs into a structured matrix, drawing on the theoretical framework developed in section two and the empirical cases discussed in sections three and four. Four trade-offs are given particular attention below.

The first trade-off, privacy versus service delivery, arises because access to welfare is increasingly conditioned on the surrender of biometric data, collapsing the distinction between consenting to be governed and consenting to be watched. The second, efficiency versus equity, reflects the observation that systems optimised for aggregate accuracy or cost reduction frequently impose disproportionate costs on already marginalised groups, as seen in Aadhaar authentication failure rates among elderly and manual-labour populations. The third, security versus liberty, is most visible in facial recognition and predictive policing deployments, where infrastructure justified by reference to public safety is repurposed for protest surveillance and dissent monitoring, narrowing the space for lawful civic expression (Internet Freedom

Foundation, 2020). The fourth, centralisation versus accountability, concerns the institutional architecture of AI governance itself: centralised biometric databases and integrated command centres concentrate both administrative capacity and risk, yet oversight mechanisms, audit rights, and grievance redress procedures have not scaled proportionately with the surveillance infrastructure they are meant to check (Marda, 2018). Read through the AI4People framework proposed by Floridi et al. (2018), each of these trade-offs can be restated as a tension between two of the framework's five core principles. Privacy versus service delivery pits autonomy, understood as the citizen's control over her own identity and data, against beneficence, understood as the state's obligation to deliver welfare efficiently. Efficiency versus equity pits an aggregate conception of beneficence against justice, understood as the fair distribution of both the benefits and the burdens of a system across the population it serves. Security versus liberty pits non-maleficence, in the sense of preventing crime and harm, against autonomy, in the sense of citizens' freedom to assemble, dissent, and move without being tracked. Centralisation versus accountability, finally, concerns explicability, the principle that intelligent systems affecting the public should be understandable and contestable by those they affect; a centralised system whose logic, error rates, and appeal mechanisms are opaque to the citizens it governs fails this criterion regardless of its technical accuracy. Framing the trade-offs in this way clarifies that none of the four tensions identified here can be resolved simply by improving technical accuracy; each requires an institutional choice about which value takes precedence, and under what safeguards.

Table 2: Ethical Trade-off Matrix in AI-Enabled Governance

Trade-off	Governance Value Advanced	Governance Value Compromised	Illustrative Case	Population Most Affected
Privacy vs. service delivery	Efficient subsidy targeting	Informational privacy and consent	Mandatory Aadhaar authentication for PDS access	Rural, low-income beneficiaries
Efficiency vs. equity	Reduced administrative cost and leakage	Equal access to entitlements	Biometric authentication failure rates	Elderly, manual labourers, persons with disabilities
Security vs. liberty	Public safety, crime prevention	Freedom of assembly and dissent	FRT deployment at protests and public gatherings	Protesters, journalists, activists
Centralisation vs. accountability	Administrative coordination and data integration	Independent oversight and grievance redress	Centralised biometric and command-centre databases	General citizenry; those denied services

Note. Compiled by the author, drawing on Dreze et al. (2017), Internet Freedom Foundation (2020, 2021), Marda (2018), and O'Neil (2016).

6. AI Surveillance and the Sustainable Development Goals

India's AI-enabled governance agenda is frequently justified in the language of sustainable development, and digital public infrastructure is presented as a mechanism for accelerating progress on multiple SDG targets simultaneously. This framing is not without merit: biometric identity systems can, in principle, reduce leakage in food and fuel subsidy programmes relevant to SDG 1 (No Poverty) and SDG 2 (Zero Hunger), while integrated health data systems can improve service delivery relevant to SDG 3 (Good Health and Well-being). However, an unexamined equation of digitisation with sustainability risks obscuring the ways in which surveillance-linked governance can simultaneously undermine other SDG commitments, particularly SDG 10 (Reduced Inequalities) and SDG 16 (Peace, Justice and Strong Institutions), which explicitly call for accountable, transparent, and non-discriminatory institutions (United Nations, 2015). Table 3 maps the surveillance mechanisms discussed in this paper against the SDG targets they are intended to serve and the SDG targets they place at risk.

The SDG 16 framing is particularly instructive because it makes explicit that sustainable governance is not reducible to service-delivery metrics. SDG target 16.10, for instance, calls for public access to information and protection of fundamental freedoms, which is difficult to reconcile with facial recognition systems deployed without dedicated legislative authorisation, published retention policies, or independent audit mechanisms (Internet Freedom Foundation, 2020). Similarly, SDG target 16.6, which calls for effective, accountable, and transparent institutions, is placed under strain when biometric welfare systems generate exclusion errors that the responsible authority does not systematically measure or disclose (Dreze et al., 2017). A sustainable governance agenda that pursues SDG 1 through 3 via digitisation while neglecting SDG 16's institutional accountability requirements is, on this reading, only partially sustainable: it may reduce measured poverty or hunger indicators while degrading the quality of the institutions through which those reductions are achieved.

This tension also has a distributive dimension that intersects directly with SDG 10. Because exclusion errors in biometric welfare systems concentrate among elderly, disabled, and manual-labour populations, and because facial recognition and predictive policing infrastructure concentrates surveillance risk among protesters and marginalised communities, the costs of AI-enabled governance are not distributed evenly across the population that is meant to benefit from it. A poverty-reduction strategy that relies on infrastructure disproportionately likely to fail the very groups it targets—elderly recipients of old-age pensions, manual labourers whose fingerprints are most prone to wear, women accessing maternal health services—risks reproducing, in digital form, the same patterns of marginalisation that sustainable development commitments are meant to redress. Measuring progress on SDG 1 through 3 solely through aggregate service-delivery statistics, without disaggregating outcomes by the demographic characteristics associated with authentication failure, therefore risks masking exactly the inequities that SDG 10 requires governments to identify and correct.

Table 3 Mapping Surveillance-Enabled Governance to Selected SDG Targets

SDG	Target(s) Advanced by Digitised Governance	Target(s) Placed at Risk
SDG 1 – No Poverty	1.3 – Nationally appropriate social protection systems	1.4 – Equal rights to economic resources and basic services undermined by exclusion errors
SDG 2 – Zero Hunger	2.1 – Access to safe, sufficient, and nutritious food via PDS	2.1 – Undermined where biometric failure denies ration access
SDG 3 – Good Health and Well-being	3.8 – Progress toward universal health coverage	3.1 / 3.2 – Maternal and child health access delayed by authentication failure
SDG 10 – Reduced Inequalities	10.2 – Promotion of social and economic inclusion	10.3 – Equal opportunity undermined by disproportionate exclusion of marginalised groups
SDG 16 – Peace, Justice and Strong Institutions	16.6 – Development of effective public institutions	16.6 / 16.10 – Accountability and public access to information undermined by opaque surveillance systems

Note. Compiled by the author, based on United Nations (2015) SDG target language and the case evidence reviewed in Sections 3 and 4.

7. Regulatory and Institutional Responses

India's regulatory response to AI-enabled surveillance has developed unevenly, with constitutional and judicial safeguards outpacing dedicated statutory regulation of specific technologies. The Supreme Court's 2017 recognition of privacy as a fundamental right in *Justice K.S. Puttaswamy (Retd.) v. Union of India* established a constitutional proportionality standard applicable to state surveillance and biometric identification, and has since been invoked in litigation and civil society advocacy concerning facial recognition and Aadhaar-linked exclusion. At the policy level, NITI Aayog has published a sequence of documents addressing responsible AI, beginning with the 2018 National Strategy for Artificial Intelligence, followed by two-part Responsible AI approach papers addressing system-level and societal-level principles and their operationalisation (NITI Aayog, 2021a, 2021b), and a subsequent discussion paper specifically addressing the responsible use of facial recognition technology. These documents articulate principles of safety, equality, inclusivity, non-discrimination, privacy, transparency, and accountability, but remain, by design, non-binding guidance rather than enforceable law (NITI Aayog, 2021a).

Civil society organisations, including the Internet Freedom Foundation, have played a significant monitoring and advocacy role in the absence of dedicated legislation, documenting the scale of facial recognition deployment through Project Panoptic and submitting formal responses to government

consultations, including a detailed critique of NITI Aayog's 2022 discussion paper on facial recognition, which argued that the paper inadequately addressed the harms of law enforcement use and function creep (Internet Freedom Foundation, 2022). This pattern—substantive ethical principles articulated at the policy level, constitutional safeguards established through litigation, and independent monitoring conducted by civil society, in the continued absence of a dedicated, binding legal framework specific to biometric surveillance and facial recognition—characterises the current state of India's AI governance architecture as this paper goes to press.

Data protection legislation, which would provide the most direct statutory route to enforceable safeguards, has followed a protracted and unsettled trajectory. An earlier Personal Data Protection Bill, introduced in 2019 to establish a comprehensive regulatory structure for personal data processing, was withdrawn amid extended parliamentary and stakeholder deliberation, and a revised Digital Personal Data Protection Bill was subsequently released for public consultation (Ministry of Electronics and Information Technology, 2022). The prolonged absence of enacted data protection legislation means that, at the time of writing, government surveillance and biometric welfare systems in India continue to operate largely under executive notification and administrative circular rather than a dedicated statutory regime specifying purpose limitation, retention limits, and independent oversight for state use of personal data. This regulatory lag is significant precisely because the pace of technological deployment—facial recognition, integrated command centres, and biometric welfare administration—has consistently outstripped the pace of the legislative and institutional safeguards meant to govern it.

8. Towards an Ethical Governance Framework: Recommendations

The trade-offs mapped in this paper are not, in themselves, an argument against the use of AI in Indian governance; digital identity and data-driven administration retain genuine potential to improve the reach and responsiveness of welfare delivery. They are, however, an argument against treating ethical safeguards as secondary to technological deployment. The following recommendations follow directly from the trade-off matrix and SDG mapping developed above.

First, purpose limitation should be legally mandated for biometric and surveillance infrastructure built for welfare administration. Data and infrastructure established to verify eligibility for a subsidy scheme should not be available by default for policing or security use without a separate, specific legal authorisation subject to judicial or independent oversight, directly addressing the function-creep pattern documented in facial recognition and command-centre deployments (Internet Freedom Foundation, 2021).

Second, authentication failure must be treated as a system defect rather than as evidence against the claimant. Welfare administration protocols should be redesigned so that biometric non-match triggers an accessible, time-bound alternative verification pathway rather than denial of entitlement, and authorities should be statutorily required to publish disaggregated failure-rate data by age, occupation, and geography to enable independent audit (Dreze et al., 2017).

Third, facial recognition and predictive policing systems should be subject to mandatory algorithmic impact assessments prior to deployment, evaluating accuracy across demographic subgroups, defining data retention limits, and establishing independent grievance redress mechanisms, consistent with the proportionality standard established in *Justice K.S. Puttaswamy (Retd.) v. Union of India* and with the AI ethics framework proposed by Floridi et al. (2018).

Fourth, India's forthcoming and evolving data protection framework should be extended to explicitly cover government surveillance infrastructure, not only private-sector data processing, closing the regulatory gap identified by civil society monitoring of facial recognition deployment (Internet Freedom Foundation, 2022).

Fifth, sustainable governance reporting linked to India's SDG commitments should incorporate institutional-quality indicators under SDG 16 alongside service-delivery indicators under SDG 1 through 3, so that gains attributed to digitisation are assessed against their surveillance and exclusion costs rather than reported in isolation.

Sixth, oversight of AI-enabled governance should include a standing, independent, and adequately resourced body with the statutory power to audit biometric and surveillance systems, publish findings, and order remedial action, rather than relying solely on ad hoc judicial review or civil society monitoring conducted without formal investigative authority. NITI Aayog's Responsible AI documents already articulate accountability as a core principle (NITI Aayog, 2021a, 2021b); the outstanding task is to give that principle an institutional home with the authority to act on it.

9. Conclusion

This paper has argued that AI-enabled governance in India cannot be adequately understood through the language of efficiency alone, because the same biometric and surveillance infrastructure that enables welfare delivery also enables population monitoring, and the ethical costs of the latter frequently fall on the populations the former is meant to serve. Drawing on Foucault's theory of disciplinary visibility and contemporary scholarship on surveillance capitalism and algorithmic harm, the paper has shown that India's watchtower-welfare governance model produces recurring trade-offs between privacy and service delivery, efficiency and equity, security and liberty, and centralisation and accountability. These trade-offs are not evenly distributed: exclusion errors in biometric welfare authentication concentrate among elderly, disabled, and manual-labour populations, while facial recognition and predictive policing infrastructure concentrates surveillance risk among protesters, dissenters, and marginalised communities. A sustainable governance agenda, properly understood, requires more than digitised efficiency; it requires that the institutions delivering development outcomes be themselves accountable, proportionate, and rights-respecting, as SDG 16 explicitly demands. India possesses the constitutional foundation, in the Puttaswamy privacy judgment, and the policy vocabulary, in NITI Aayog's responsible AI documents, to build such a framework. What remains is the translation of these principles into binding, enforceable, and independently audited safeguards that treat the ethical trade-offs of AI surveillance as a central design constraint, rather than an externality to be managed after deployment. Only on that basis can India's watchtowers be reconciled with its welfare commitments.

Future research should build on the qualitative, documentary approach adopted here with systematic, disaggregated empirical measurement: nationally representative data on biometric authentication failure by age, occupation, and geography; independent audit of facial recognition accuracy across demographic subgroups in Indian deployments; and longitudinal tracking of welfare exclusion linked specifically to digital verification failure, as distinct from other causes of programme non-take-up. Such research would allow the trade-off matrix developed in this paper to be tested against systematic evidence rather than case material alone, and would provide the empirical foundation that any future statutory oversight body, of the kind recommended above, would need in order to function effectively. Until that evidence base is built

and acted upon, India's sustainable governance agenda will remain caught between the watchtower and the welfare office—each visible to the other, neither fully accountable to the citizen standing between them.

REFERENCES:

1. Amnesty International, Article 19, & Internet Freedom Foundation. (2021). Use of facial recognition technologies on a steep rise in India: The Hyderabad Command and Control Centre [Joint statement]. Internet Freedom Foundation.
2. Bentham, J. (1791). *Panopticon; or, the inspection-house*. T. Payne.
3. Dreze, J., Khalid, N., Khera, R., & Somanchi, A. (2017). Aadhaar and food security in Jharkhand: Pain without gain? *Economic and Political Weekly*, 52(50), 51–59.
4. Floridi, L., Cows, 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>
5. Foucault, M. (1977). *Discipline and punish: The birth of the prison*. Pantheon Books.
6. Internet Freedom Foundation. (2020). *Project Panoptic: Tracking the use of facial recognition technology in India*. Internet Freedom Foundation.
7. Internet Freedom Foundation. (2021). *The increasing use of facial recognition technology in India*. Internet Freedom Foundation.
8. Internet Freedom Foundation. (2022). *Response to NITI Aayog's draft discussion paper on facial recognition technology*. Internet Freedom Foundation.
9. Justice K.S. Puttaswamy (Retd.) v. Union of India, (2017) 10 SCC 1 (India).
10. Khera, R. (2019). *Dissent on Aadhaar: Big data meets big brother*. Orient BlackSwan.
11. Khera, R., & Somanchi, A. (2021). Aadhaar and healthcare access: Biometric authentication failure and maternal health outcomes. *Economic and Political Weekly*, 56(post-COVID special issue).
12. Ministry of Electronics and Information Technology. (2022). *Digital Personal Data Protection Bill, 2022 [Draft for public consultation]*. Government of India.
13. Marda, V. (2018). Artificial intelligence policy in India: A framework for engaging the limits of data-driven decision-making. *Philosophical Transactions of the Royal Society A*, 376(2133), Article 20180087. <https://doi.org/10.1098/rsta.2018.0087>
14. NITI Aayog. (2018). *National strategy for artificial intelligence: #AIforAll*. Government of India.
15. NITI Aayog. (2021a). *Responsible AI for all: Approach document for India, Part 1 — Principles for responsible AI*. Government of India.
16. NITI Aayog. (2021b). *Responsible AI for all: Approach document for India, Part 2 — Operationalizing principles for responsible AI*. Government of India.
17. O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.



18. United Nations. (2015). Transforming our world: The 2030 agenda for sustainable development. United Nations General Assembly.
19. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.