STATE RESPONSIBILITY FOR ARTIFICIAL INTELLIGENCE: ATTRIBUTION, DUE DILIGENCE, AND ACCOUNTABILITY UNDER INTERNATIONAL HUMAN RIGHTS LAW

Avni Sharma
LL.M. (Cyber Law)
Amity Law School
Amity University Noida , Uttar Pradesh

Abstract: *The rapid proliferation of artificial intelligence (AI) systems in governmental and quasi-governmental functions has exposed a fundamental lacuna in international legal doctrine: the existing framework of state responsibility, codified in the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), was constructed around paradigms of human agency that AI systems fundamentally challenge. This article examines three interlocking questions: first, how the attribution rules of ARSIWA apply when an AI system — developed by a private actor but deployed by a state — produces human rights violations; second, whether states incur responsibility through passive failure to regulate AI-driven harm caused by third parties within their jurisdiction; and third, what accountability mechanisms — spanning developer liability, operator liability, and institutional redress — are capable of bridging the gap between the autonomous character of AI decision-making and the anthropocentric assumptions embedded in existing law. Drawing on ICJ jurisprudence, comparative constitutional case law from India and the European Court of Justice, and emerging soft-law instruments including the EU Artificial Intelligence Act, the OECD AI Principles, and the UNESCO Recommendation on the Ethics of Artificial Intelligence, the article argues that state responsibility in the AI era requires a reconceptualised attribution standard, an expansive due diligence obligation, and a tiered liability architecture that tracks the lifecycle of AI deployment.*

Index Terms *— Artificial Intelligence; State Responsibility; ARSIWA; Attribution; Due Diligence; Algorithmic Accountability; Human Rights; EU AI Act; International Law; Cyber Law.*

I. INTRODUCTION

The integration of artificial intelligence into the architecture of state power has proceeded at a pace that international legal doctrine has struggled to match. Governments worldwide now rely on AI-driven systems for facial recognition and mass surveillance, predictive policing, credit and welfare eligibility determinations, autonomous military logistics, and criminal risk assessment. In each of these domains, AI systems exercise a form of functional authority that was previously exercised by identifiable human officials subject to established norms of accountability. Yet when such systems produce discriminatory outputs, infringe individual privacy, or cause cognisable harm, the legal mechanisms for assigning responsibility remain deeply uncertain [1].

The foundational text of international state responsibility — the Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), adopted by the International Law Commission (ILC) in 2001 — presupposes a world in which conduct can be traced to human actors exercising identifiable forms of agency. The attribution rules of ARSIWA operate through organically connected concepts: a state is responsible for

the acts of its organs, for the acts of entities exercising elements of governmental authority, and for the acts of private actors exercised under its direction or control. These concepts were formulated against a backdrop of traditional inter-state disputes — armed conflict, treaty breach, harm to foreign nationals — and their application to the distributed, adaptive, and often inscrutable decision-making of AI systems raises problems that have yet to be systematically resolved [2].

This article argues that three doctrinal challenges are particularly acute. The first is the attribution problem: when an AI system developed by a private technology company is procured and deployed by a state for governmental functions, the effective-control test established in Nicaragua v. United States and refined in Bosnia v. Serbia is an imperfect instrument for resolving responsibility, because the state's control over the system's outputs may be indirect, prospective, or technically opaque. The second is the due diligence problem: to what extent are states obliged, under both general international law and specific human rights treaty obligations, to regulate the AI activities of private actors operating within their territory or jurisdiction? The third is the accountability architecture problem: given that AI systems involve multiple actors across a lifecycle that spans design, training, deployment, and post-deployment adaptation, what liability frameworks are necessary to ensure that victims of AI-related harm have access to effective remedies? [3][4].

The article proceeds in four further parts. Part II examines the doctrinal foundations of state responsibility and analyses how the attribution and breach elements of ARSIWA operate in the AI context. Part III considers state obligations under international human rights law, including the duty to protect against third-party AI-related harm. Part IV surveys the liability frameworks that have emerged across domestic and regional jurisdictions, with particular attention to Indian constitutional jurisprudence and the European regulatory model. Part V offers conclusions and proposes a principled framework for reconceptualising state responsibility in the age of artificial intelligence.

II. ARSIWA AND THE ATTRIBUTION PROBLEM IN AI SYSTEMS

A. The Architecture of State Responsibility

The doctrine of state responsibility performs a foundational function in the international legal order: it establishes the conditions under which a state may be held internationally accountable for conduct inconsistent with its obligations and specifies the legal consequences that follow. ARSIWA codifies this doctrine in a comprehensive form that has achieved broad acceptance as an expression of customary international law. Under Article 2 of ARSIWA, an internationally wrongful act — the trigger for state responsibility — requires two conjunctive elements: the act must be attributable to the state, and it must constitute a breach of an international obligation incumbent upon that state [1].

The attribution rules of ARSIWA, set out in Articles 4 through 11, identify several bases on which conduct may be attributed to a state. Article 4 covers the acts of state organs — understood broadly to include any person or entity having organ status under internal law. Article 5 extends attribution to entities empowered to exercise elements of governmental authority. Article 8 — perhaps the most litigated provision in contemporary practice — provides that conduct is attributable to the state if carried out under the state's direction or control. The ICJ's elaboration of Article 8 in Nicaragua produced the demanding 'effective control' test, requiring that the state exercise actual direction and control over each specific operation that gives rise to the breach [2][3].

The standard was revisited in Bosnia v. Serbia, where the Court confirmed that the effective control test remained the applicable criterion for attributing acts of non-state actors to a state, while acknowledging that its application requires context-sensitive analysis. In the context of AI systems, these tests face a challenge that is qualitatively different from the scenarios for which they were designed [4].

B. Attribution and the Opacity of AI Agency

AI systems present at least three features that complicate the straightforward application of existing attribution standards. First, modern machine-learning systems — particularly deep-learning architectures — are characterised by epistemic opacity: the pathway from input to output is not transparent, even to their developers. A state that deploys an AI system for predictive policing may lack the technical capacity to explain, and therefore to direct, how the system generates risk scores for individual subjects. The question of whether 'effective control' can exist over a system whose decision logic is not fully accessible is one that ARSIWA's drafters never contemplated [5].

Second, AI systems are typically not state organs in the sense of Article 4, because they are developed and often maintained by private commercial entities. A facial recognition system procured by a law enforcement agency from a private vendor occupies an intermediate legal space: it exercises a form of governmental authority (Article 5) while remaining proprietary technology subject to minimal state control over its underlying algorithmic processes. The 'elements of governmental authority' formulation in Article 5 was designed to address entities like privatised utilities or contracted border services; its application to software systems that autonomously make consequential determinations is uncertain [5].

Third, the adaptive nature of AI systems means that conduct that was not 'directed' at the time of deployment may emerge through post-deployment learning. A state that procures and deploys an AI decision system and thereafter allows the system to operate without adequate oversight may find itself responsible for outcomes that no identifiable human within the state apparatus directed, intended, or even foresaw. On these facts, the effective control test is both under-inclusive and analytically ill-suited to the distributed character of AI-driven misconduct [6].

C. Towards a Functional Attribution Standard

A principled resolution of the attribution problem requires moving beyond the binary logic of control toward a functional standard sensitive to the specific role the state has played in bringing an AI system's harmful outputs into existence. Where a state procures an AI system for deployment in a domain of governmental authority — law enforcement, adjudication, benefit determination, surveillance — that system should, as a matter of international law, be treated as an extension of state power for purposes of attribution. The rationale is that the state has affirmatively chosen to vest in that system a function that would otherwise be subject to accountability norms applicable to state organs [5][6].

This functional approach finds indirect support in the ILC's own commentary, which emphasises that the attribution rules must be applied in light of their object and purpose — to ensure that states cannot evade responsibility by delegating governmental functions to nominally private entities. The same logic applies where the 'entity' is an AI system. A state that substitutes an AI system for a human officer in a function that carries legal consequences for individuals cannot coherently argue that the system's outputs are attributable to no one [2].

III. INTERNATIONAL HUMAN RIGHTS OBLIGATIONS AND THE DUTY TO REGULATE AI

A. The Tripartite Structure of Human Rights Obligations

International human rights law imposes on states a tripartite structure of obligations: to respect rights (the obligation to refrain from interference), to protect rights (the obligation to prevent and respond to interference by third parties), and to fulfil rights (the obligation to take positive measures to realise rights). It is the obligation to protect that is most relevant to AI-related harm caused by private actors. Under the International Covenant on Civil and Political Rights and its jurisprudence, states are required to ensure that private actors — including corporations operating AI systems — do not infringe the rights of individuals within the state's jurisdiction [7][8].

In the context of AI, the obligation to protect crystallises around several core rights. The right to privacy under Article 17 of the ICCPR is directly implicated by AI-powered surveillance, data collection, and profiling. The right to equality and non-discrimination under Article 26 is engaged wherever AI-driven decision systems produce systematically biased outcomes along protected grounds. The right to an effective remedy under Article 2(3) requires that individuals adversely affected by AI systems have access to procedurally adequate means of challenge and reparation [8][9].

B. Due Diligence as a Regulatory Obligation

The principle of due diligence — the obligation of states to take reasonable measures to prevent harm caused by activities within their jurisdiction — provides the doctrinal vehicle through which the obligation to protect is translated into concrete regulatory duties. As an expression of general international law, the due diligence principle imposes on states an obligation of conduct rather than result: states are not guarantors against all AI-related harm, but they must demonstrate that they have established and enforced adequate regulatory frameworks [6].

What constitutes adequate regulation in the AI context is a question that different jurisdictions have approached in different ways. The European Union's Artificial Intelligence Act establishes a risk-based classification system under which AI applications in high-risk domains — including law enforcement, migration control, and judicial decision-making — are subject to stringent requirements of transparency, explainability, human oversight, and pre-market conformity assessment. These requirements operationalise due diligence in a legally binding form: a member state that fails to enforce compliance with these requirements in respect of AI systems deployed within its territory may incur responsibility for the resulting harm [10].

The OECD Principles on AI and the UNESCO Recommendation on the Ethics of AI, while lacking binding legal force, represent convergent international consensus on the substantive content of the due diligence obligation. Taken together, they require transparency and explainability, mechanisms of accountability and redress, human oversight in high-stakes decisions, and non-discrimination by design. A state that fails to incorporate these standards into its domestic regulatory framework may, depending on the specific treaty obligations it has assumed, be in breach of its obligation to protect under international human rights law [11][12].

C. Algorithmic Discrimination as a Human Rights Violation

The phenomenon of algorithmic discrimination — the systematic production of biased outcomes by AI systems trained on historically unrepresentative data — is increasingly recognised as a human rights problem of the first order. Research has demonstrated that AI systems deployed in credit scoring, recidivism prediction, employment screening, and targeted advertising tend to reproduce and amplify existing structural inequalities, producing outputs that disadvantage persons on grounds of race, gender, socioeconomic status, and other protected characteristics [13][14].

From the perspective of state responsibility, algorithmic discrimination may engage two distinct heads of liability. First, where a state directly deploys a discriminatory AI system in a governmental function, the system's discriminatory outputs are attributable to the state and constitute a breach of its non-discrimination obligations under applicable human rights law. Second, where a private actor deploys a discriminatory AI system and the state has failed to establish adequate regulatory or enforcement mechanisms to detect and remedy such discrimination, the state may be responsible for its failure to protect [13][15].

IV. LIABILITY ARCHITECTURE AND ACCOUNTABILITY MECHANISMS

A. The Multi-Stakeholder Problem

A central challenge of AI accountability is that responsibility for an AI system's outputs is distributed across a chain of actors — developers who design the underlying model, organisations that provide training data, operators who deploy the system, and end-users who interact with its outputs. Conventional tort law, premised on a dyadic model of harm-doer and victim, provides an imperfect template for this multi-stakeholder context. The opacity of machine-learning systems further complicates causation analysis: where a discriminatory output cannot be traced to a specific design decision by an identifiable actor, the conventional burden of proof may be impossible to discharge [5].

B. Developer, Operator, and State Liability

A workable liability architecture must track the lifecycle of AI deployment. Developer liability — analogous to product liability for defective goods — attaches where a harmful output is causally connected to a defect in the design, training data, or testing of the AI system. The analogy is imperfect because AI systems evolve post-deployment, and developers may legitimately disclaim responsibility for emergent behaviours that post-date their involvement. Nevertheless, the product liability analogy provides a useful starting point, particularly where the defect is identifiable ex ante through appropriate pre-deployment testing [10].

Operator liability — analogous to employer liability — attaches where an organisation that deploys an AI system fails to exercise adequate oversight, fails to implement appropriate safeguards, or misuses the system in a manner inconsistent with its design parameters. This form of liability is particularly relevant where a private employer uses an AI system to make hiring decisions that produce discriminatory outcomes: the employer, as operator, owes a duty of care to affected individuals that is not discharged merely by reliance on the algorithmic process [14].

State liability attaches where governmental actors are directly involved in the deployment of AI systems, particularly in law enforcement, migration, and social welfare administration. The state cannot insulate itself from responsibility by interposing a private contractor between itself and the affected individual; the

functional attribution principle discussed in Part II ensures that responsibility follows governmental function, not the formal identity of the actor who discharges it [16].

C. Accountability Mechanisms: Transparency, Redress, and Impact Assessment

Effective accountability requires institutional mechanisms that operate before, during, and after AI deployment. Ex ante mechanisms — including algorithmic impact assessments and mandatory conformity testing for high-risk applications — allow prospective identification and mitigation of discriminatory or harmful design features. The EU AI Act's requirement of pre-market assessment for high-risk AI systems reflects a regulatory philosophy that preventive accountability is more effective and less costly than post-hoc remediation [10].

Ongoing accountability requires transparency and explainability obligations that enable regulators, civil society, and affected individuals to scrutinise the decision logic of AI systems. The concept of 'meaningful human control' — the requirement that consequential AI-assisted decisions remain subject to human review and override — serves as a minimum accountability safeguard in domains such as criminal justice, where the stakes of automated error are especially severe [9][10].

Ex post accountability is constituted by redress mechanisms that allow individuals adversely affected by AI decisions to challenge those decisions and obtain appropriate remedies. The Indian Supreme Court's decisions in Shreya Singhal v. Union of India and Anuradha Bhasin v. Union of India established important constitutional parameters: statutory provisions authorising broad-based digital restrictions must satisfy the proportionality test, and digital access — including the integrity of digitally mediated decisions — is constitutionally protected [17][18].

At the European level, the Court of Justice's decision in Google Spain SL v. AEPD articulated the 'right to be forgotten' as an exercise of individual control over personal data processed by algorithmic systems. This right — now codified in Article 17 of the GDPR — imposes on operators a legal obligation to erase personal data upon request in circumstances where the data's continued processing is no longer justified [19][20].

The US Supreme Court's decision in Carpenter v. United States articulates a principle of constitutional significance for AI governance more broadly: the aggregation of digitally collected personal data by state actors constitutes a form of intrusion that the constitutional framework was designed to prevent, even where each individual datum is lawfully obtained. The aggregation principle applies with special force to AI-powered surveillance systems that synthesise multiple streams of personal data to construct comprehensive profiles of individuals [21].

The recognition of electronic evidence in Praful B. Desai v. State of Maharashtra, while primarily a procedural precedent, establishes that Indian courts are willing to engage with the evidentiary and procedural dimensions of digital technology, providing a doctrinal foundation for the adjudication of AI-related claims [22].

V. CONCLUSIONS AND RECOMMENDATIONS

This article has argued that the existing framework of state responsibility, while not wholly inapplicable to the AI context, requires principled adaptation at three points. The attribution rules of ARSIWA must be supplemented by a functional attribution standard that treats AI systems deployed in domains of governmental

authority as extensions of state power, regardless of whether the system was developed by a private entity. The due diligence principle requires states to establish, maintain, and enforce adequate regulatory frameworks for AI systems that operate within their jurisdiction. The accountability architecture must be redesigned to reflect the multi-stakeholder and multi-phase character of AI deployment, incorporating developer liability for defective design, operator liability for inadequate oversight, and state liability for the acts of AI systems exercising governmental authority [1][6][16].

Several specific recommendations follow from this analysis. First, the International Law Commission should consider developing supplementary guidance on the application of ARSIWA to AI-related conduct, explicitly addressing the functional attribution principle and the standard of 'direction or control' in the context of AI procurement and deployment.

Second, states should incorporate algorithmic impact assessment requirements into their domestic legal frameworks, modelled on — but not necessarily limited by — the approach taken in the EU AI Act. Such assessments should be mandatory for high-risk AI applications in governmental functions and should be subject to independent review. India, which has not yet enacted comprehensive AI-specific legislation, would benefit from a risk-based framework that builds on its existing constitutional jurisprudence on proportionality and the protection of digital access established in Anuradha Bhasin [10][18].

Third, states should establish accessible redress mechanisms for individuals adversely affected by AI-driven decisions, including the right to explanation of automated decisions, the right to human review, and appropriate remedies — including compensation — where AI-related harm is established. The principle of effective remedy, recognised in both ICCPR Article 2(3) and ARSIWA arts. 34–37, demands no less [8][23].

Finally, the international community should move towards binding multilateral standards on AI governance that complement and reinforce the existing human rights framework. The voluntary character of the OECD Principles and the UNESCO Recommendation, while valuable as expressions of emerging consensus, is insufficient to discipline the deployment of AI systems by states and private actors that choose not to align their practices with those standards. A treaty-based framework would provide the enforcement architecture that the present governance landscape conspicuously lacks [11][12].

Artificial intelligence is not merely a technological development that requires incremental legal adjustment. It represents a structural transformation of the exercise of power — by states, corporations, and hybrid actors — that demands a correspondingly structural response from the international legal order. The vindication of human rights in the age of AI depends on the willingness of the legal community to engage with these challenges with the creativity and doctrinal rigour they demand.

ACKNOWLEDGMENT

The author is grateful to the faculty of the Department of Cyber Law, National Law University, New Delhi, for their guidance and support in the development of this research.

REFERENCES

International Instruments

[1] Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), International Law Commission, UN Doc. A/56/10 (2001).

[2] Articles on Responsibility of States for Internationally Wrongful Acts (ARSIWA), arts. 4–11 and ILC Commentary thereto.

[3] Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America), Merits, Judgment, ICJ Reports 1986, p. 14, paras. 109–115.

[4] Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v. Serbia and Montenegro), Judgment, ICJ Reports 2007, p. 43, paras. 385–407.

[6] ARSIWA, art. 16; Trail Smelter Arbitration (United States v. Canada) (1941) 3 RIAA 1905; Corfu Channel Case (United Kingdom v. Albania), Merits, ICJ Reports 1949, p. 4.

[7] Universal Declaration of Human Rights, GA Res. 217 (III) A, UN Doc. A/RES/217(III) (10 December 1948).

[8] International Covenant on Civil and Political Rights, opened for signature 16 December 1966, 999 UNTS 171 (entered into force 23 March 1976).

[9] UN Human Rights Council, 'The Right to Privacy in the Digital Age', UN Doc. A/HRC/39/29 (3 August 2018).

[11] OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (22 May 2019).

[12] UNESCO, Recommendation on the Ethics of Artificial Intelligence (23 November 2021), SHS/BIO/PI/2021/1.

[16] UN Guiding Principles on Business and Human Rights, UN Human Rights Council Res. 17/4 (16 June 2011).

[20] General Data Protection Regulation, Regulation (EU) 2016/679, [2016] OJ L 119/1.

[23] ARSIWA, arts. 34–37 (forms of reparation).

Cases

[17] Shreya Singhal v. Union of India, (2015) 5 SCC 1 (Supreme Court of India).

[18] Anuradha Bhasin v. Union of India, (2020) 3 SCC 637 (Supreme Court of India).

[19] Google Spain SL v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja González, Case C-131/12, ECLI:EU:C:2014:317 (CJEU, 13 May 2014).

[21] Carpenter v. United States, 138 S. Ct. 2206 (2018) (US Supreme Court).

[22] Praful B. Desai v. State of Maharashtra, (2003) 4 SCC 601 (Supreme Court of India).

Books and Articles

[5] Calo R, 'Artificial Intelligence Policy: A Primer and Roadmap' (2017) 51 UC Davis Law Review 399.

[10] European Commission, Proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act), COM(2021) 206 final.

[13] Barocas S and Selbst AD, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671.

[14] O'Neil C, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy (Crown Publishers 2016).

[15] Selbst AD, 'Disparate Impact in Big Data Policing' (2017) 52 Georgia Law Review 109.