AI Accountability Framework 2026

By

Dr. Pavan Duggal

Architect, Global AI Accountability

President, Global Artificial Intelligence Accountability Law and Governance Institute

Framework and Doctrines

PREAMBLE

Artificial intelligence now shapes decisions about health, work, education, finance, security, and public life. When these systems malfunction, discriminate, or are abused, the consequences fall on people who often have little visibility into how decisions are made and few effective remedies when things go wrong.

This Framework treats AI accountability as a foundational requirement of the Intelligence Age. Its purpose is to ensure that power exercised through algorithms is subject to law, traceable to human decision-makers, and answerable to those affected.

The global landscape is plagued by what we might call "principle fatigue", being a proliferation of well-intentioned but non-binding declarations that lack enforcement mechanisms and have failed to prevent algorithmic harms. This Framework breaks that cycle by establishing concrete, legally enforceable obligations with clear liability standards, robust auditing requirements, and meaningful remedies.

Four Core Objectives

1. **Enforceable Obligations:** To define AI accountability as a binding legal duty, not merely a moral aspiration.
2. **Adaptable Architecture:** To set out clear principles and doctrines that can be adopted across diverse legal systems and cultural contexts.

3. **Proportionate Regulation:** To create a risk-based structure for liability and redress that matches regulatory burdens to potential harms.
4. **Global South Leadership:** To articulate a vision of AI governance that resists digital colonialism and reflects diverse legal and cultural traditions, particularly those of developing economies.

This document states the core principles and doctrines of the Framework.

PART I: FOUNDATIONAL DEFINITION AND PRINCIPLES

1. The Accountability Imperative

AI systems increasingly mediate decisions that were previously taken by humans. They influence who receives credit, how patients are triaged, which candidates are shortlisted for jobs, how police resources are deployed, and what information citizens see. This raises a straightforward question: **who is accountable when AI systems cause harm?**

AI accountability is the enforceable obligation of identifiable natural and legal persons to:

- Ensure that AI systems for which they are responsible operate within applicable legal, ethical, and technical requirements.
- Explain and justify AI-mediated decisions to affected persons in terms they can understand.
- Provide timely and effective remedies where harm occurs.
- Maintain meaningful human oversight across the AI lifecycle, from design through decommissioning.

This definition rests on four pillars:

Prevention: Embedding compliant, safe, and fair operation through accountability-by-design principles.

Transparency: Making system operations and decisions understandable to those affected.

Remediation: Providing effective redress, appeal mechanisms, and appropriate compensation for harms.

Governance: Maintaining continuous, meaningful human control over AI systems.

AI accountability is distinct from related but less robust concepts:

- **AI Ethics** provides normative guidance but lacks binding legal force.
- **AI Governance** establishes institutional structures but may not assign clear individual responsibility.

- **AI Compliance** focuses on regulatory adherence but can become checkbox exercises without genuine answerability.
- **AI Responsibility** describes causal roles but may not create legally enforceable duties.
- **AI Transparency** enables accountability but is insufficient on its own.

The central challenge is that AI systems can learn, adapt, and decide with minimal human involvement in individual cases, creating ambiguity about where responsibility lies. Many advanced systems operate through processes difficult to interpret even for their creators. The AI value chain involves multiple actors including data providers, developers, deployers, infrastructure providers, thereby making it hard to assign clear accountability. Systems can make millions of decisions per second, affecting vast numbers of people simultaneously.

Traditional legal frameworks struggle with these realities. Current approaches rely too heavily on voluntary commitments routinely subordinated to competitive pressures, fragmented national laws that enable regulatory arbitrage, and reactive governance that responds to harms after they occur rather than preventing them through design requirements.

This Framework is designed to overcome these systemic failures through legally binding standards, globally harmonized wherever possible, with robust enforcement mechanisms and clear consequences for non-compliance.

2. Foundational Values

This Framework is grounded in multiple traditions: constitutionalism and human rights law, respect for persons as ends in themselves, consequentialist concern for real-world impacts, Eastern ethics of duty and harmony (including Confucian emphasis on hierarchical responsibility and Buddhist principles of minimizing suffering), and Indigenous insights on stewardship, intergenerational responsibility, and systems thinking.

Together, they support one central commitment: **AI must serve human dignity and cannot displace human responsibility.**

Human rights provide hard constraints on AI design and deployment. Systems that undermine autonomy, privacy, equality, due process, or freedom of expression are incompatible with this Framework, regardless of efficiency or profit. Where AI interferes with these rights, affected persons are entitled to explanation, human review, and effective remedy.

The Framework also draws on virtue ethics and institutional character. Sustainable AI governance depends not only on rules but on organizational cultures that value practical wisdom (phronesis), justice, transparency, and prudence, particularly the willingness to admit and correct error.

The defence that "the algorithm did it" is firmly rejected as both legally and ethically unacceptable. AI systems, regardless of sophistication, are instruments created and deployed by people. They do not possess moral agency, consciousness, or capacity for genuine ethical

reasoning. Accountability must always trace back to identifiable human actors and institutions.

3. The Eleven Foundational Principles

These principles distil the Framework's philosophical commitments into specific, actionable mandates. Each serves as a binding obligation with concrete implementation requirements.

Principle 1: Human Rights and Human-Centric Values

Mandate: AI systems shall be designed and deployed to augment human well-being, agency, and flourishing, never to replace or diminish human capacities, and shall operate within constitutional and international human rights frameworks.

Core obligations:

- Uphold the principle of "Do No Harm," preventing physical, psychological, economic, or legal injury to individuals or communities.
- Ensure all deployments respect privacy, freedom of expression, freedom of association, equality, and non-discrimination.
- Clearly inform users when they are interacting with an AI system rather than a human being.
- Prohibit AI use for purposes fundamentally incompatible with human dignity, such as mass manipulation or exploitation of vulnerable populations.

Principle 2: Transparency and Explainability

Mandate: People are entitled to know when AI is used in consequential decisions and to receive explanations they can understand.

Core obligations:

- Disclose, in clear language, when AI systems are used in decisions that affect legal rights, access to services, or significant opportunities.
- Maintain documentation (Cards and Data Sheets) describing system purpose, data sources, performance metrics, and key limitations.
- Provide affected individuals, on request, with a meaningful explanation of the main factors that influenced a decision and how they were weighted.
- Preserve technical records sufficient for independent expert review by regulators, auditors, and courts.

Principle 3: Accountability and Liability

Mandate: Clear accountability for AI system outcomes shall be assigned to identifiable human actors and institutions across the entire value chain.

Core obligations:

- Establish explicit chains of responsibility documenting which persons and entities bear accountability for each lifecycle stage.
- Prohibit "AI autonomy," "algorithmic complexity," or "machine learning opacity" as defences against legal liability.
- Maintain comprehensive, immutable audit trails documenting decisions, data inputs, and system states.
- Implement clear governance structures assigning ultimate decision-making authority to identified senior leaders.

Principle 4: Fairness and Non-Discrimination

Mandate: AI systems shall be designed, tested, and continuously monitored to prevent unfair bias, discriminatory outcomes, and disparate impacts against protected demographic groups.

Core obligations:

- Conduct rigorous pre-deployment bias testing across all legally protected categories including race, ethnicity, gender, age, disability, and religion.
- Utilize and publicly report on established fairness metrics including demographic parity, equalized odds, and calibration across groups.
- Assemble diverse development and testing teams to identify a wider range of potential biases.
- Perform regular independent fairness audits by qualified third parties.

Principle 5: Privacy and Data Protection

Mandate: Organizations shall protect individual privacy and ensure ethical data stewardship throughout the AI lifecycle, complying with all applicable data protection laws and respecting individual data rights.

Core obligations:

- Implement "Privacy by Design" principles, incorporating privacy-enhancing technologies including anonymization, differential privacy, and federated learning.
- Obtain explicit, informed, unbundled consent for personal data use in AI training and operation.
- Guarantee individuals the right to know what data was used, to access their data, and to have data deleted.
- Conduct Data Protection Impact Assessments (DPIAs) for systems processing personal or sensitive data before deployment.

Principle 6: Safety, Security, and Robustness

Mandate: AI systems shall be demonstrably safe, secure, and resilient by design, with risks continuously assessed and mitigated throughout their operational lifecycle.

Core obligations:

- Conduct comprehensive pre-deployment risk assessments using established frameworks such as NIST AI RMF or ISO 31000.
- Perform adversarial testing ("red teaming") to identify vulnerabilities to malicious attacks, data poisoning, and model inversion.
- Implement robust cybersecurity measures protecting training data, model parameters, and operational systems.
- Design fail-safe mechanisms ensuring graceful degradation or safe shutdown in critical failures, including emergency "kill switches" for high-risk applications.

Principle 7: Human Oversight and Control

Mandate: Human beings shall retain ultimate authority, meaningful control, and final decision-making power over AI systems, with ability to intervene, override, or shut down systems, particularly in high-stakes domains.

Core obligations:

- Implement tiered oversight models calibrated to risk: Human-in-the-Loop (HITL) for active review of significant decisions; Human-on-the-Loop (HOTL) for continuous monitoring with intervention capability; Human-in-Command for ultimate strategic authority.
- Mandate qualified human review for all high-stakes decisions affecting fundamental rights, safety, liberty, or economic opportunities.
- Provide comprehensive training for human overseers on system capabilities, limitations, common failure modes, and automation bias.
- Establish documented escalation pathways and incident response procedures.

Principle 8: Auditability and Traceability

Mandate: AI systems shall be designed to enable independent verification, validation, and forensic investigation of their functioning, decisions, and compliance.

Core obligations:

- Maintain comprehensive, automated logging of all significant system decisions, data inputs, model outputs, and operational events.
- Ensure creation of tamper-evident audit trails securely stored and accessible to authorized auditors and regulators.

- Implement strict version control for all models, algorithms, training datasets, and system configurations.
- Retain audit records for a minimum of seven years to support investigations, regulatory inquiries, and legal proceedings.

Principle 9: Contestability and Redress

Mandate: Individuals adversely affected by AI decisions must have access to effective mechanisms to challenge those decisions, seek human review, and obtain meaningful remedies including compensation.

Core obligations:

- Establish accessible, user-friendly complaint and appeal channels that do not require technical expertise or legal representation.
- Guarantee the right to appeal automated decisions to qualified human reviewers with appropriate expertise and decision-making authority.
- Adhere to defined response timelines: acknowledge complaints within 5 business days and provide substantive responses within 30 days.
- Implement fair compensation frameworks ensuring appropriate remedies for material harms, psychological injuries, and reputational damage.

Principle 10: Inclusivity, Diversity, and Sustainability

Mandate: AI development and deployment shall actively promote social inclusion, reflect human diversity, and contribute to environmental sustainability and long-term flourishing.

Core obligations:

- Ensure development teams are diverse across gender, race, ethnicity, age, disability, geographic origin, and disciplinary background.
- Conduct meaningful consultations with affected communities, particularly historically marginalized groups, during design, testing, and evaluation.
- Assess and mitigate environmental footprint including energy consumption, carbon emissions, electronic waste, and water usage.
- Design systems for accessibility across cultures, languages, literacy levels, and disabilities.

Principle 11: Governance and Adaptability

Mandate: AI governance frameworks must be adaptive and evolutionary, designed as living instruments that evolve with technological advancement and societal learning.

Core obligations:

- Establish formal internal AI governance bodies with genuine executive authority and direct reporting to boards of directors or senior leadership.
- Mandate periodic review and revision of all AI-related policies and procedures at intervals not exceeding 18 months.
- Participate actively in multi-stakeholder forums and international standard-setting bodies.
- Implement structured processes for incorporating lessons from incidents, audits, and emerging research into updated practices.

PART II: DOCTRINES FOR AI ACCOUNTABILITY

Traditional legal concepts struggle to address the unique challenges of AI: opacity, autonomy, distributed responsibility, and probabilistic reasoning. To close these accountability gaps, this Framework introduces eleven legal doctrines that form the jurisprudential core for courts, regulators, and legislators.

Doctrine 1: Non-Delegable Algorithmic Responsibility

Principle: Ultimate legal accountability for AI system outcomes cannot be delegated to algorithms, artificial intelligence systems, or any non-human entity. Natural persons and legal entities remain fully accountable for decisions mediated by AI systems they develop, deploy, or control.

Rationale: This doctrine forecloses the "algorithm did it" excuse that threatens to create a fundamental accountability gap. AI systems are sophisticated instruments, but they are instruments nonetheless. Legal accountability presupposes moral agency, intentionality, and capacity for ethical reasoning. Since AI systems possess none of these attributes, responsibility must reside with the persons and entities who create and use them.

Illustrative application: A financial institution deploys an AI underwriting system that systematically denies mortgage applications from qualified members of a protected demographic group. The institution remains fully accountable for these discriminatory outcomes and cannot evade liability by attributing decisions to the algorithm's independent determinations. It must demonstrate adequate bias testing, appropriate safeguards, and meaningful human oversight or face legal consequences.

Doctrine 2: Perpetual Accountability of AI Systems

Principle: Accountability for AI systems persists throughout their entire lifecycle and beyond, from initial development through active operation and even after decommissioning. Past deployment or system retirement does not extinguish liability for continuing impacts or long-term consequences.

Rationale: The impacts of AI systems, particularly biases embedded in training data or flaws in algorithmic design, can manifest long after deployment and may continue to affect

individuals even after a system is retired. Historical decisions made by a flawed system can have cascading effects on life trajectories and accumulated disadvantage. This doctrine ensures organizations cannot evade responsibility for past harms by quietly discontinuing a problematic system.

Illustrative application: An employment screening algorithm operated from 2022 to 2024 is discovered in 2026 to have systematically discriminated against candidates with certain disability-related employment gaps. Even though the company decommissioned the system in 2024, it remains liable for discriminatory hiring decisions made during its operational period. Affected individuals screened during this time have standing to seek remedies including retrospective review and appropriate compensation.

Doctrine 3: Explainability as Due Process

Principle: When AI systems make decisions materially affecting individual rights, liberties, opportunities, or significant interests, including employment, credit, education, healthcare, criminal justice, and government benefits, affected individuals possess a fundamental, enforceable right to receive a meaningful explanation in terms they can understand.

Rationale: Due process of law requires both notice of decisions affecting one's interests and meaningful opportunity to be heard or to challenge those decisions. An effective opportunity to challenge a decision is impossible without understanding its basis, the factors considered, and the reasoning applied. This doctrine elevates algorithmic transparency from a technical best practice to a constitutional requirement in high-stakes contexts.

Illustrative application: A government agency uses an AI system to deny a citizen welfare benefits based on predicted likelihood of employment. The agency must provide more than bare notice of denial. It must explain in plain language which factors about the applicant's circumstances were most influential, how those factors were weighted, and what alternative circumstances might have led to approval. Failure to provide meaningful explanation constitutes a due process violation, rendering the decision legally defective.

Doctrine 4: Algorithmic Fiduciary Duty

Principle: When an AI system is deployed in contexts characterized by significant power asymmetry, information imbalance, or dependency, particularly in advisory relationships such as healthcare, financial planning, legal services, or educational guidance, the deploying entity bears fiduciary-like duties of care, loyalty, and good faith toward affected individuals.

Rationale: Fiduciary duties arise in relationships where one party possesses superior knowledge, expertise, or power, and the other must rely on the fiduciary's judgment in a position of trust and vulnerability. AI deployment often creates or intensifies such asymmetries: deploying organizations control algorithmic design, training data, and operational parameters that are opaque to users, while users must rely on AI outputs for consequential decisions. This doctrine imposes a higher standard of care than ordinary negligence.

Illustrative application: A financial advisory platform deploys AI algorithms to recommend investment portfolios to retail clients. Investigation reveals the algorithms systematically favour proprietary investment products generating higher fees for the platform over comparable third-party products that would better serve client interests. This violates the platform's fiduciary duty. The platform cannot defend by showing recommended products are "adequate". It must demonstrate affirmatively that recommendations prioritize client welfare over platform profit.

Doctrine 5: AI Harm Presumption

Principle: For AI systems classified as high-risk, there exists a rebuttable presumption that an unvalidated, inadequately tested, or improperly documented system will cause harm. The burden rests on the developer or deployer to affirmatively demonstrate the system's safety, fairness, and regulatory compliance before deployment.

Rationale: AI systems can cause widespread, rapid, and potentially irreversible harm at scale. Traditional "wait for harm" reactive regulation is inadequate when a single flawed system can affect millions before problems are detected. This doctrine implements a precautionary principle for high-risk contexts, shifting the default from permissive deployment to cautious, evidence-based validation.

Illustrative application: A healthcare technology company plans to deploy an AI-powered diagnostic tool for cancer screening in hospitals nationwide. Regulatory authorities can presume this high-risk system will cause harm unless the company provides comprehensive evidence from independent clinical validation studies demonstrating diagnostic accuracy, performance consistency across diverse patient demographics, appropriate calibration of confidence levels, and safe integration into clinical workflows. The burden of proof rests entirely with the company.

Doctrine 6: Reverse Burden of Proof in AI Harm

Principle: When an individual suffers demonstrable harm in a context involving an AI system and can establish a credible prima facie connection between the harm and the system's operation, the burden of proof shifts to the AI developer or deployer to demonstrate either that the system did not cause the harm or that all reasonable preventive safeguards were properly implemented.

Rationale: The profound information asymmetry between individuals and operators of opaque AI systems creates an insurmountable barrier to traditional litigation. Plaintiffs typically lack access to algorithmic source code, training data, operational logs, or technical documentation necessary to prove causation. Deployers possess exclusive control over this evidence. This doctrine corrects that structural imbalance, ensuring accountability is not defeated by complexity or secrecy.

Illustrative application: An individual with strong qualifications and excellent credit history is denied a mortgage by an AI underwriting system. She demonstrates that her profile is comparable to or stronger than approved applicants and identifies her membership in a

protected demographic group. Having made this prima facie showing, the burden shifts to the lending institution to prove through transparent documentation that the denial was based on legitimate, non-discriminatory factors appropriately weighted in the algorithmic model, and that the institution conducted adequate bias testing. If the lender cannot meet this burden, liability follows.

Doctrine 7: Digital Sovereignty in AI Decision-Making

Principle: Nations and individuals possess fundamental rights to self-determination over consequential decisions affecting them. AI systems designed and developed in one jurisdiction must meaningfully respect the legal norms, cultural values, constitutional principles, and ethical standards of the jurisdictions where they are deployed and whose citizens they affect.

Rationale: This doctrine protects against "algorithmic colonialism", which represents the imposition of values, biases, and priorities of a few dominant technology-producing nations upon the rest of the world through globally deployed AI systems. When algorithms trained on data reflecting one society's norms make decisions affecting individuals in radically different cultural and legal contexts, they can undermine local self-governance and democratic sovereignty.

Illustrative application: A global social media platform deploys content moderation AI developed in Western jurisdictions, trained primarily on Western cultural norms regarding acceptable speech. In a Global South nation, the AI systematically removes political commentary and cultural expression that, while acceptable and constitutionally protected locally, triggers Western-centric definitions of "problematic content." The affected nation can require the platform either to substantially modify the AI system to respect local constitutional values, to implement local human review for content decisions, or face operational restrictions.

Doctrine 8: Collective Algorithmic Rights

Principle: AI systems can harm entire groups, communities, or demographic categories, not merely isolated individuals. Legal frameworks must recognize collective rights to be free from algorithmic discrimination, exploitation, and systemic disadvantage, providing legal standing for communities to challenge AI-caused harms and seek collective remedies.

Rationale: Traditional legal frameworks focus on individual injury and require individual plaintiffs to demonstrate particularized harm. However, AI systems often cause systemic, distributed harm through pattern discrimination. No single individual may suffer catastrophic injury, but the aggregate effect disadvantages entire communities. This doctrine enables effective legal challenges to such systemic harms by granting standing to community organizations, advocacy groups, and civil society actors.

Illustrative application: A community organization representing a predominantly minority, low-income neighbourhood files suit against a municipal government's AI-powered resource allocation system. They allege the system systematically under-allocates public services,

including infrastructure maintenance, educational resources, health services, to their neighbourhood while over-allocating to affluent areas. Under this doctrine, the organization has standing to bring this collective action on behalf of the entire affected community. The court can order systemic remedies including algorithm modification, enhanced community consultation, and compensatory resource allocation.

Doctrine 9: Algorithmic Precaution

Principle: Where deployment of an AI system poses plausible risks of serious, widespread, or irreversible harm to individuals, communities, democratic institutions, or societal structures, lack of complete scientific certainty about those risks shall not justify postponing cost-effective preventive measures.

Rationale: AI systems, particularly powerful foundation models and systems deployed at massive scale, can cause harms including erosion of democratic discourse through targeted disinformation, entrenchment of systemic discrimination, manipulation of vulnerable populations, cascading economic disruption, that are extremely difficult or impossible to reverse once they occur. The precautionary principle, well-established in environmental law and public health regulation, mandates preventive action in the face of plausible, high-impact risks even before definitive proof of harm exists.

Illustrative application: Regulators evaluate a powerful new generative AI model capable of producing highly convincing synthetic text, images, audio, and video. While comprehensive empirical evidence of mass harm does not yet exist, the model's capabilities create plausible risks of widespread election interference, fraud, and erosion of epistemic trust. Under this doctrine, regulators can impose mandatory safeguards including output watermarking, staged rollout with monitoring, usage restrictions for political advertising, identity verification for certain applications, before allowing broad deployment.

Doctrine 10: AI Accountability Inheritance

Principle: When an AI system, its underlying intellectual property, or its deploying organization is sold, merged, acquired, or otherwise transferred between entities, accountability obligations transfer with it. Successor entities inherit full legal and ethical accountability for the past impacts, ongoing operations, and future consequences of acquired AI systems.

Rationale: This doctrine prevents organizations from evading responsibility through strategic corporate restructuring, asset sales, spin-offs, or bankruptcy proceedings. Without accountability inheritance, companies could develop and deploy problematic AI systems, profit from their operation, and then transfer them to shell entities or dissolve themselves to escape liability when harms become apparent.

Illustrative application: Company A develops and deploys a facial recognition system subsequently discovered to have severe racial bias problems. Before lawsuits are filed, Company B acquires Company A's assets including the facial recognition system. Under this doctrine, Company B inherits full accountability for the system's discriminatory impacts,

including both past harms from Company A's deployment and ongoing harms from Company B's continued operation. Company B cannot claim it is a good-faith purchaser without notice; due diligence obligations include evaluating AI systems for bias, safety, and compliance issues before acquisition.

Doctrine 11: Algorithmic Due Diligence

Principle: Organizations deploying AI systems in contexts affecting legal rights, economic opportunities, physical safety, or individual welfare must exercise heightened due diligence proportionate to potential impacts. This affirmative obligation extends beyond contractual warranties to independent verification and validation of AI system performance, fairness, safety, and fitness for purpose.

Rationale: Deployers cannot reasonably rely solely on vendor assurances when deploying AI systems affecting individuals' fundamental interests. The complexity and opacity of modern AI systems, combined with the severity of potential harms and information asymmetry between vendors and deployers, create an affirmative obligation for deployers to conduct independent validation. Good faith reliance on vendor claims, sufficient in many commercial contexts, is inadequate when algorithmic decisions affect rights, opportunities, or safety.

Illustrative application: A large employer purchases a third-party AI hiring system marketed as "bias-free" and "validated." The employer deploys the system based solely on vendor claims without conducting its own bias testing against local applicant demographics. The system produces discriminatory outcomes systematically disadvantaging protected groups. Under this doctrine, good faith reliance on vendor representations is not a complete defence. The employer had an affirmative duty to conduct independent bias audits before deployment and continuous fairness monitoring during operation.

PART III: RISK, HARM, AND LIABILITY

1. The AI Risk Classification System

The Framework establishes a four-tier risk classification system ensuring regulatory burdens are proportionate to potential harms.

Unacceptable Risk (Prohibited)

Examples: Social scoring systems by public authorities for general population control; real-time biometric surveillance in public spaces without judicial authorization; subliminal manipulation techniques causing harm; AI systems deliberately exploiting vulnerabilities of children or disabled persons; indiscriminate scraping of facial images for recognition databases.

Requirements: Outright prohibition. Development, deployment, or operation is illegal.

Penalties: Criminal sanctions including imprisonment up to 7 years for deliberate deployment; fines up to 6% of global annual turnover; immediate shutdown orders; asset seizure.

High-Risk AI Systems (Strict Regulation)

Examples: Medical diagnosis and treatment planning; credit scoring and lending decisions; employment screening and evaluation; critical infrastructure management (energy, water, transportation); law enforcement tools including predictive policing; biometric identification systems; educational assessment and tracking; immigration and asylum decision support.

Requirements:

- Mandatory pre-deployment conformity assessment and Algorithmic Impact Assessment (AIA)
- Independent third-party audit before deployment and annually thereafter
- Continuous performance, fairness, and safety monitoring
- Mandatory human oversight with qualified reviewers
- Strict liability for harms caused by system defects or failures
- Registration in public databases of high-risk AI systems

Penalties: Administrative fines up to 7% of global annual turnover; license revocation; mandatory corrective orders; public disclosure of violations; civil liability with reverse burden of proof.

Limited-Risk AI Systems (Transparency Requirements)

Examples: Chatbots and conversational AI; emotion recognition systems; deepfake generators and synthetic media tools; biometric categorization for non-law-enforcement purposes; content recommendation algorithms.

Requirements:

- Clear, conspicuous disclosure to users that they are interacting with AI
- Mandatory labeling of AI-generated synthetic content (watermarking or metadata)
- User notification when emotion recognition or biometric categorization is deployed
- Basic documentation of system capabilities and limitations

Penalties: Administrative fines up to 1% of turnover; corrective orders; public disclosure of violations.

Minimal-Risk AI Systems (Light Regulation)

Examples: AI-powered spam filters and content moderation; inventory management and supply chain optimization; AI-enabled video games and entertainment; recommendation engines for non-consequential content; personal AI assistants for scheduling.

Requirements: Voluntary adoption of industry best practices and codes of conduct; standard consumer protection laws apply.

Penalties: Market discipline; reputation harm; standard consumer protection remedies.

2. Categories of AI Harm

Direct Harms: Immediate, tangible damage with clear causal links including but not limited to physical injury from autonomous systems; economic loss from erroneous credit denials; liberty deprivation from wrongful arrests based on flawed identification; immediate psychological distress from harmful content.

Indirect Harms: Consequential damage flowing from AI decisions, including but not limited to diminished life prospects from biased educational tracking; career trajectory alteration from discriminatory hiring; health deterioration from delayed or incorrect diagnoses; accumulated economic disadvantage from systematically higher pricing.

Systemic Harms: Widespread societal impacts affecting entire populations including but not limited to entrenchment of historical discrimination; erosion of democratic processes through algorithmic manipulation; degradation of privacy norms; concentration of economic and political power; environmental damage from energy-intensive AI systems.

Latent Harms: Delayed or cumulative harms manifesting over time including but not limited to long-term health consequences of misdiagnoses; intergenerational perpetuation of bias through self-reinforcing systems; gradual erosion of critical thinking and human judgment; accumulated psychological impacts of constant surveillance.

The Framework distinguishes between harms to specific, identifiable individuals (actionable through individual litigation) and harms affecting entire communities or demographic groups (actionable through collective mechanisms under the Doctrine of Collective Algorithmic Rights).

3. The Liability Framework

Strict Liability (High-Risk AI)

Principle: Developers and deployers of high-risk AI systems are automatically liable for harm caused by system defects, inadequate testing, or operational failures, regardless of fault, negligence, or reasonable precautions taken.

Rationale: Strict liability is appropriate for inherently dangerous or particularly consequential activities. It places the burden of unavoidable risks on those who create those risks and profit from them.

Limited Defences:

- Sophisticated, unforeseeable third-party intervention breaking the causal chain
- Unauthorized modification or tampering by the deployer beyond developer control
- Proper use by end-user in compliance with all safety warnings and limitations

Damages Recoverable: Economic losses (lost wages, medical expenses, property damage); non-economic damages (pain and suffering, emotional distress); reputational harm and dignitary injuries; punitive damages for wilful misconduct or gross negligence.

Negligence-Based Liability (Limited- and Minimal-Risk AI)

Standard of Care: Organizations must exercise reasonable care proportionate to foreseeable risks when developing or deploying limited-risk and minimal-risk AI systems.

Modified Elements:

- **Duty:** Established by statute and common law principles of reasonable care; heightened for professional contexts.
- **Breach:** Determined by comparison to industry standards, best practices, and Framework requirements.
- **Causation:** Must demonstrate but-for causation and proximate cause; Doctrine of Reverse Burden of Proof applies once prima facie case is made.
- **Damages:** Traditional categories apply.

Product Liability Adaptations

The Framework explicitly treats AI systems as "products" subject to product liability law:

Design Defects: Flaws in algorithmic architecture, inadequate bias mitigation, insufficient safety mechanisms, or inappropriate choice of training approach that render the system unreasonably dangerous.

Manufacturing Defects: Errors in training execution, data quality failures, implementation bugs, or deployment mistakes that cause specific instances to deviate from intended design.

Failure to Warn: Inadequate disclosure of system limitations, failure to specify appropriate use contexts, insufficient documentation of known risks, or deceptive marketing claims about capabilities.

Implied Warranties: AI systems carry implied warranties of merchantability (fit for ordinary purposes) and fitness for particular purpose (when seller knows of buyer's specific needs).

PART IV: STAKEHOLDER RESPONSIBILITIES AND INSTITUTIONS

1. Differentiated Accountability Across the AI Value Chain

AI Developers & Model Creators

Design-Time Obligations:

- Conduct preliminary impact assessments evaluating potential harms before development commences
- Embed accountability-by-design features including comprehensive audit trails, explainability mechanisms, fairness constraints, and monitoring hooks
- Create detailed technical documentation describing capabilities, limitations, intended use contexts, and known failure modes
- Implement security-by-design principles protecting against adversarial attacks

Testing Requirements:

- Comprehensive functionality testing ensuring systems perform as intended across diverse conditions
- Rigorous fairness testing across demographic groups using multiple metrics
- Security testing including adversarial attacks, data poisoning, and model inversion
- Independent third-party verification and certification for high-risk systems

Liability Standard: Strict liability for design defects, inadequate bias mitigation, insufficient safety measures, failure to disclose known limitations, and deceptive performance claims.

Data Providers and Curators

Obligations:

- Ensure data quality, accuracy, completeness, and representativeness for intended purposes
- Maintain comprehensive provenance documentation (Data Sheets) tracking sources, collection methods, and known biases
- Obtain explicit, informed, unbundled consent for personal data use in AI training
- Assess training data for historical biases and implement appropriate mitigation strategies

Liability Standard: Strict liability if training data is demonstrably defective, unrepresentative, obtained through rights violations, or contains undisclosed biases that foreseeably cause algorithmic harms.

AI Deployers & Implementers

Obligations:

- Conduct context-specific Algorithmic Impact Assessments (AIAs) before deploying any high-risk system

- Provide clear notice to affected individuals when AI systems are used in consequential decisions
- Implement robust human oversight mechanisms calibrated to risk levels
- Conduct continuous performance monitoring in real-world operational environments
- Maintain incident response capabilities and establish clear escalation procedures
- Provide accessible complaint and appeal mechanisms with defined timelines

Liability Standard: Liable for inappropriate deployment contexts, inadequate human oversight, failure to monitor performance, inadequate response to known failures, and failure to provide effective redress. Good faith reliance on vendor claims is not a complete defence. It is imperative that deployers have independent due diligence obligations under Doctrine 11 entitled "Algorithmic Due Diligence".

State Actors & Government Bodies

Obligations:

- Meet the highest standards of constitutional compliance, transparency, and due process
- Conduct and publish mandatory Human Rights Impact Assessments for all rights-affecting systems
- Submit to independent audits with full public disclosure of findings
- Provide legislative oversight and ensure availability of judicial review
- Establish clear accountability chains to elected or appointed officials

Liability Standard: Subject to constitutional challenges, civil rights enforcement actions, and administrative law remedies. Qualified immunity generally does not apply to algorithmic discrimination or constitutional violations.

Cloud & Infrastructure Providers

Obligations:

- Provide infrastructure supporting accountability requirements including audit logging, data residency controls, and monitoring tools
- Implement robust security controls enabling deployer compliance with safety requirements
- Maintain service-level agreements ensuring availability and reliability
- Cooperate with lawful regulatory investigations and legal process

Liability Standard: Can be held liable as component manufacturers if infrastructure services are integral to defective AI systems and providers failed to meet reasonable care standards or contractual obligations.

Open-Source Contributors & Projects

Governance Requirements:

- Large open-source foundation models must adopt formal governance structures (foundations, steering committees)
- Maintainers bear responsibility for addressing known security vulnerabilities and safety issues within reasonable timeframes
- Projects must provide clear documentation of limitations and appropriate use

Safe Harbour Provisions:

- Individual contributors to volunteer-driven, non-commercial projects are generally shielded from liability
- Small-scale, community-based projects receive exemptions from formal compliance requirements
- Liability focuses on commercial entities deploying or building upon open-source models
- Good Samaritan protections for reasonable efforts to identify and remediate issues

2. Institutional Architecture

National AI Accountability Authority (NAIRA)

There is a need for countries to have in place their respective National AI Accountability Authorities (NAIRA)

Establishment: Created by national legislation with constitutionally protected mandate; independent from direct political control and industry capture; secure, multi-year funding ensuring operational independence.

Structure: Governed by multi-disciplinary commission including legal scholars, technologists, ethicists, and civil society representatives; appointed through transparent, merit-based process with public input; term limits and conflict-of-interest protections ensuring independence.

Powers:

- **Rulemaking Authority:** Develop detailed regulations implementing Framework principles
- **Investigative Powers:** Audit algorithms, compel document production, interview personnel

- **Enforcement Powers:** Issue fines, corrective orders, license revocations
- **Advisory Functions:** Provide guidance, publish best practices, facilitate stakeholder dialogue
- **International Cooperation:** Coordinate with foreign regulators on cross-border issues

Organizational AI Accountability Office

There is a need for organizations to have in pace their Organizational AI Accountability Office.

Internal Structure:

- Independent office within organizations developing or deploying high-impact AI
- Direct reporting line to board of directors or C-suite leadership
- Authority to audit, suspend, or escalate concerns about AI projects
- Protected from retaliation for raising accountability concerns

Cross-Functional Governance:

- Chief AI Accountability Officer (CAO) with executive authority
- Mandatory members: Chief Data Officer, Chief Risk Officer, Chief Compliance Officer, Chief Ethics Officer
- At least two external independent members from civil society, academia, or affected communities
- Regular reporting to board of directors on AI risks and incidents

Accountability Documentation:

- Formal RACI (Responsible, Accountable, Consulted, Informed) matrices for each AI project
- Clear documentation of who executes work, who has final decision authority, who provides expert input, and who receives updates

PART V: GLOBAL SOUTH LEADERSHIP AND THE DECOLONIAL IMPERATIVE

The Challenge of Digital Colonialism

Effective AI governance cannot follow a one-size-fits-all model dominated by Western perspectives. Current patterns risk perpetuating historical cycles of technological dependence: during colonialism, colonies provided raw materials fuelling metropolitan industrialization while remaining dependent on manufactured imports. Today, Global South nations provide

the raw data fueling AI innovation concentrated in the Global North, with value and power accruing disproportionately to Northern corporations.

This Framework articulates a different path, one authored by and designed for the Global South, reflecting diverse legal traditions, developmental priorities, and cultural values.

Four Pillars of Global South AI Governance

1. Authorship, Not Adoption

Global South nations must author their own accountability frameworks reflecting unique legal, cultural, and developmental contexts. This Framework is designed in and for developing economies, not merely adapted from Western models. Development priorities, institutional capacities, and cultural values differ significantly across regions. The "technology transfer" model that treats the Global South as passive recipient is rejected.

2. Co-Design, Not Compliance

Global South nations are empowered as co-designers of international standards, with equal voice in international standard-setting bodies and treaty negotiations. The Framework moves beyond passive compliance with Northern-designed rules toward leadership roles in defining accountability for AI affecting Southern populations.

3. Consensus, Not Hegemony

International harmonization must be achieved through genuine multi-stakeholder consensus, rejecting hegemonic imposition by technologically dominant nations. Democratic processes must ensure all affected voices shape global governance, with respect for sovereignty in implementing international norms.

4. Flexibility, Not Uniformity

Implementation mechanisms must be flexible enough to accommodate varying institutional capacities, with tiered approaches recognizing different developmental stages. Technical assistance and capacity building are core obligations, with long-term commitment to building sustainable local expertise.

Countering Data Exploitation

Mechanisms of Exploitation:

- **"Free Services" Model:** Northern technology companies harvest vast data from Global South users, providing services "free" while extracting value from data. AI systems trained on Global South data serve primarily Northern interests, with profits and capabilities concentrating in technology-producing nations.
- **Value Capture Asymmetries:** Global South provides essential training data while Northern corporations capture overwhelming majority of economic value, with limited technology transfer or local capacity building, perpetuating technological dependency.

- **Algorithmic Dependency:** Global South nations dependent on foreign AI systems for critical functions, lacking local alternatives or competitive capabilities, vulnerable to service withdrawal or terms changes.

Framework Counter-Measures:

Mandatory Technology Transfer: Foreign AI companies operating in Global South must engage in meaningful technology transfer through joint ventures with local partners building domestic capabilities, training programs developing local AI expertise, and open-source contributions enabling local innovation.

Fair Compensation for Data Use: Legal requirements for fair compensation, royalties, or revenue sharing; Community Data Trusts enabling collective bargaining; transparent accounting of data value in AI systems; prohibition of exploitative terms of service extracting data for nominal consideration.

Community Data Governance: Establishment of community-controlled Data Trusts; collective ownership and governance of aggregated local data; fair compensation negotiated by representative bodies; democratic decision-making about data uses; culturally appropriate consent mechanisms respecting local languages, literacy levels, and cultural norms.

Differentiated Implementation

The Framework recognizes varying capacity levels and provides tiered implementation:

Tier 1 (Advanced Capacity Nations): Full framework implementation capability; resources to assist other Global South nations; leadership in regional standard-setting. Examples: India, Brazil, South Africa, Indonesia.

Tier 2 (Developing Capacity Nations): Phased implementation over 24-36 months; targeted international assistance for specific gaps; gradual scaling of compliance requirements; focus on building institutional infrastructure.

Tier 3 (Building Capacity Nations): Simplified framework focusing on most critical protections; extended implementation timelines (48-60 months); significant international technical and financial support; emphasis on fundamental rights protections and capacity development.

Sovereign AI Capability Development

Education and Research Infrastructure: Prioritize AI and data science education in national curricula; fund research infrastructure including computing resources and datasets; create competitive domestic opportunities retaining local talent.

South-South Cooperation: Collaborative research initiatives among Global South institutions; shared computing infrastructure reducing individual nation costs; common regional regulatory frameworks enabling harmonization; joint development of Global South-specific AI applications.

Open Source and Technology Sovereignty: Prioritize open-source AI solutions enhancing transparency; avoid vendor lock-in to proprietary Northern platforms; build local expertise through open-source contribution; create Global South-led open-source AI projects.

PART VI: ENFORCEMENT AND INTERNATIONAL COOPERATION

1. Enforcement Mechanisms

Regulatory Enforcement

Administrative Penalties:

- Fines up to 7% of global annual turnover for severe violations
- Escalating penalties for repeated violations
- License revocations for persistent non-compliance
- Public disclosure of violations and enforcement actions

Corrective Orders:

- Mandatory system modifications or decommissioning
- Enhanced monitoring and reporting requirements
- Appointment of independent compliance monitors
- Restrictions on development of new high-risk systems

Criminal Enforcement:

- Criminal liability for wilful deployment of prohibited systems
- Prosecution for deliberate discrimination or rights violations
- Penalties including imprisonment for severe violations causing death or serious injury
- Corporate criminal liability for systematic violations

Judicial Remedies & Individual Redress

Rights of Affected Individuals:

- Right to explanation of consequential automated decisions
- Right to human review and appeal
- Right to compensation for damages caused by AI systems
- Right to participate in class actions for widespread harms

Burden-Shifting Mechanisms:

- Doctrine of Reverse Burden of Proof applies once prima facie case established

- Deployers must produce evidence of compliance and proper safeguards
- Failure to maintain required documentation creates adverse inference

Remedies Available:

- Compensatory damages for economic and non-economic harms
- Punitive damages for wilful misconduct or gross negligence
- Injunctive relief halting harmful practices
- Declaratory judgments establishing rights and responsibilities
- Attorneys' fees and costs for prevailing plaintiffs

2. International Cooperation

Interoperability with Existing Frameworks

EU AI Act Alignment: Risk classification systems aligned for mutual recognition; streamlined certification process for systems approved under both frameworks; coordinated enforcement for multinational corporations.

Integration with Global Principles: Transform OECD AI Principles from recommendations into binding obligations; operationalize UNESCO Recommendation on Ethics of AI; implement UN Guiding Principles on Business and Human Rights for AI; align with emerging Council of Europe AI Convention.

Cross-Border Enforcement Cooperation

Information Sharing Networks: Regular communication among national regulatory authorities; shared databases of AI incidents, violations, and best practices; coordinated investigations of multinational AI systems; early warning systems for emerging risks.

Mutual Recognition Agreements: Recognition of certifications and audits across participating jurisdictions; streamlined approval processes for compliant systems; coordinated enforcement actions against global actors; shared technical standards and testing methodologies.

Vision for Global AI Accountability Treaty

Treaty Objectives: Establish minimum global accountability standards; obligate signatory states to create competent regulatory authorities; create binding dispute resolution mechanisms; ensure global "race to the top" for responsible AI.

Key Provisions: Universal adoption of foundational principles; harmonized risk classification systems; cross-border liability and enforcement cooperation; technology transfer and capacity building for developing nations; regular treaty conferences to update standards.

PART VII: ADVANCED DOCTRINAL INNOVATIONS

Cognitive Sovereignty

As neurotechnology and AI converge, this Framework establishes protective rights around the final frontier of privacy, being the human mind itself.

The Five Neuro-Rights:

1. **Right to Mental Privacy:** Neural data including brain activity patterns, cognitive states, emotional responses, thought processes, is classified as the most sensitive category of personal data, requiring explicit, granular, purpose-limited, and freely revocable opt-in consent.
2. **Right to Cognitive Integrity:** Protection from AI-driven manipulation, alteration, or interference with cognitive processes, decision-making capacity, or mental states without fully informed, specific consent, including protection from subliminal manipulation and cognitive hijacking.
3. **Right to Mental Continuity:** Protection against AI systems or brain-computer interfaces that fragment, disrupt, or fundamentally alter an individual's sense of self, personal identity, or biographical continuity.
4. **Right to Augmentation Equity:** Fair, non-discriminatory access to cognitive enhancement technologies, preventing emergence of a "cognitive underclass" denied access to AI-mediated augmentation available to privileged populations.
5. **Right to Neurodiversity:** Recognition, respect, and accommodation of diverse cognitive styles, neurological variations, and information processing approaches rather than enforcing algorithmic conformity to neurotypical norms.

Algorithmic Sovereignty

Nations possess inherent capacity and constitutional authority to govern algorithmic systems within their jurisdictions, particularly when those systems affect their citizens, shape public discourse, or influence democratic processes.

Core Tenets:

Territorial Jurisdiction Over Effects: A nation may legitimately regulate any AI system whose outputs, decisions, or impacts materially affect persons, entities, or interests within its borders, regardless of where the system is developed or operated.

Mandatory Transparency for Democratic Accountability: The right to demand clear identification of system operators, beneficial owners, and decision-making authorities; accessible explanations of how systems function; auditable records of system decisions and impacts.

Democratic Control Over Algorithmic Power: Legislative scrutiny, parliamentary oversight, and judicial review for AI systems that shape public discourse, allocate public resources, enforce laws, or otherwise exercise quasi-governmental functions.

Protection of National Interests: The right to restrict, modify, or prohibit AI systems that demonstrably threaten national security, economic sovereignty, cultural integrity, public health, democratic institutions, or fundamental constitutional values as democratically determined.

CLOSING DECLARATION

We stand at a moment of choice. AI systems now mediate decisions that shape human lives, opportunities, and societies. The question is not whether we will have AI, but whether we will have accountability for AI.

This Framework provides a path forward, from voluntary ethics to enforceable law, from aspirational principles to concrete obligations, from reactive governance to preventive design, from Northern hegemony to Global South leadership.

The core commitments are clear:

Power exercised through algorithms must be subject to law and answerable to those affected. Human beings retain ultimate responsibility for AI-mediated decisions; the "algorithm did it" defence is rejected. Those harmed by AI systems are entitled to explanation, human review, and effective remedy. AI development and deployment must respect human rights, protect privacy, prevent discrimination, and maintain meaningful human control.

Implementation requires action from all stakeholders:

Governments must legislate with urgency, establish independent regulatory authorities with genuine enforcement power, and lead international cooperation toward binding global standards.

Industry must embrace accountability as foundational to sustainable innovation, implement robust governance structures, pursue third-party certification, invest in accountability technologies, and accept appropriate liability for AI-caused harms.

Civil society and academia must serve as vigilant watchdogs, conduct independent research, educate and empower communities, represent affected populations, and propose alternative approaches.

Individuals must demand accountability, support responsible organizations, participate in governance, and report harms when they occur.

The Framework is offered to the global community as a blueprint for ensuring AI serves humanity's highest values. It can be adopted into national legislation, adapted to local contexts, and refined through implementation experience. What cannot be compromised are

the core commitments: enforceability over voluntarism, prevention over reaction, human dignity over efficiency, and Global South leadership over digital colonialism.

For accountability that is not merely documented but actively demonstrated. For AI governance that is not imposed but democratically determined. For an Intelligence Age that serves human flourishing.

Document Information

Title: The AI Accountability Framework 2026: Framework and Doctrines
Version: 3.0
Date: January 2026
Status: Final
Authored by: Dr. Pavan Duggal

Citation: Duggal, P. (2026). The AI Accountability Framework 2026: Framework and Doctrines. [Version 3.0]