

The Instrumental Turn in AI Governance: From Principles to Enforceable Policy Instruments

Sophia LI

Chinese Culture Society, Hong Kong SAR, China · lisophiahk@gmail.com · ORCID: 0009-0007-4866-7815

Abstract

AI governance confronts a persistent gap between normative consensus and enforceable policy. While over 160 ethics frameworks endorse transparency, fairness, and accountability, governance outcomes remain elusive. This article argues that AI governance is undergoing an **instrumental turn**: a paradigm shift from principle articulation to instrument operationalization. Drawing on Hood's (1983) NATO taxonomy, Howlett's (2011) instrument choice theory, and regulatory design literature, we develop a five-category instrument taxonomy—prohibition, licensing, disclosure, audit, and liability—and a **Principle-Instrument Matrix** mapping abstract principles onto enforceable mechanisms. Comparative analysis of the EU AI Act, OECD Principles, and UNESCO Recommendation reveals uneven instrumentalization progress across jurisdictions. An **Auditability Standard**—five criteria for principle-instrument translation readiness—explains why safety instruments have advanced furthest (5/5) while fairness instruments remain contested (1/5). The article identifies boundary conditions through four pathologies of **instrumental overfitting** and advances three testable propositions for future empirical research. Policymakers may adopt the Principle-Instrument Matrix as an analytical framework for translating AI ethics into actionable regulatory design.

Keywords policy instruments; AI ethics; instrumental turn; regulatory design; transparency; accountability; governance; Hood NATO taxonomy; auditability

1 Introduction

1.1 The Paradox of Ethical Consensus

We have achieved ethical consensus on artificial intelligence. Multiple systematic reviews confirm remarkable convergence across the hundreds of AI ethics frameworks produced over the past decade: transparency, fairness, accountability, privacy, and safety appear in nearly all major guidelines (Jobin et al., 2019; Haggendorff, 2020; Schiff et al., 2021). International organizations—the OECD, UNESCO, the Council of Europe—have endorsed these principles in formal instruments. Major technology companies have adopted internal ethics frameworks reflecting similar commitments.

Yet governance outcomes have not followed. Despite near-universal endorsement of transparency, algorithmic systems remain largely opaque to affected parties and regulators alike. Despite commitments to fairness, documented disparities persist across high-stakes algorithmic applications in criminal justice (Angwin et al., 2016), healthcare (Obermeyer et al., 2019), and employment (Raghavan et al., 2020). Despite accountability pledges, identifying responsible parties when algorithms fail remains stubbornly difficult.

This persistent gap between principle and practice has generated extensive commentary. Critics attribute the gap to corporate capture of the ethics agenda (Bietti, 2020; Wagner, 2018), the indeterminacy of ethical principles (Mittelstadt, 2019), the structural limitations of voluntary frameworks (Rességuier & Rodrigues, 2020), or the absence of meaningful enforcement mechanisms (Yeung, 2019). These critiques contain important insights, but they share a common limitation: they diagnose what is wrong with current approaches without specifying what alternative institutional arrangements would look like in operational terms.

This article proposes such an alternative. We argue that AI governance is undergoing an “**instrumental turn**”—a paradigm shift from principle articulation to instrument operationalization. The instrumental turn does not abandon ethical principles but recognizes that principles are not self-executing. They require translation into specific policy instruments: rules, standards, procedures, and enforcement mechanisms that channel behavior and enable accountability.

1.2 The Instrumental Turn Defined

Definition: The Instrumental Turn. The instrumental turn in AI governance is the paradigm shift from normative principle articulation to policy instrument operationalization, emphasizing the design, selection, and institutional embedding of enforceable mechanisms for achieving governance objectives.

Table 1: Core Concepts: Definitions and Diagnostic Tests

Concept	Definition	Positive Test	Negative Test
Instrumental Turn	Paradigm shift from principles to enforceable policy mechanisms	Focus on instrument design, selection, and enforcement infrastructure	Continued emphasis on principle articulation without operational translation
Policy Instrument	A mechanism through which governments pursue policy objectives	Specifies obligations, verification procedures, and enforcement consequences	Aspirational statement without operational specification
Principle-Instrument Gap	Disconnect between ethical commitments and governance outcomes	Principles endorsed but not operationalized in enforceable rules	Principles effectively translated into enforceable mechanisms
Auditability Standard	Minimum verifiable condition for principle-instrument translation	Principle can be verified by independent third parties against observable criteria	Principle stated without measurable or verifiable operationalization

The instrumental turn is conceptually informed by a longer trajectory in policy studies. Hood’s (1983) foundational taxonomy of governing instruments—nodality, authority, treasure, and organization (NATO)—demonstrated that governments possess a finite toolkit for achieving policy objectives. Howlett (2011) extended this analysis to instrument choice, showing that instrument selection is shaped by institutional context, policy capacity, and prior instrument commitments. In AI governance, these insights have been conspicuously absent: the overwhelming focus has been on what principles to adopt, not what instruments to deploy.

The three core concepts of this article stand in a hierarchical relationship. The **principle-instrument gap** is the empirical problem: principles are endorsed but not operationalized. The **instrumental turn** is the theoretical response: a paradigm shift that redirects analytical attention from principle articulation to instrument design. The **Auditability Standard** is the technical condition for successful translation: it specifies the minimum criteria a principle must satisfy before effective instruments can be designed around it. Together, these three concepts form a diagnostic chain—from problem identification through theoretical reorientation to operational specification—that structures the article’s argument.

1.3 Scope and Limitations

Three clarifications delimit the scope of this article. First, the instrumental turn addresses the *design and selection* of policy instruments, not the normative question of which ethical principles should govern AI. Second, this article focuses on public policy instruments rather than private governance mechanisms. Third, the article addresses instruments applicable to AI systems generally rather than instruments specific to particular domains such as healthcare, criminal justice, or financial services.

The article makes three distinct contributions. First, it systematically identifies the structural reasons why principle-based AI governance fails, distinguishing five failure modes with distinct remedial implications. Second, it develops a Principle-Instrument Matrix that maps abstract principles onto concrete policy mechanisms. Third, it proposes an Auditability Standard—minimum verifiable conditions for translating principles into enforceable governance—that serves as a diagnostic tool for evaluating regulatory proposals.

The remainder of this article proceeds as follows. Section 2 examines the proliferation of AI ethics principles and the critique of principle-based governance. Section 3 develops the theoretical framework of the instrumental turn. Section 4 presents the Principle-Instrument Matrix. Section 5 provides a comparative analysis of EU, OECD, and UNESCO instrumentalization strategies. Section 6 analyzes institutional prerequisites for effective instrumentalization. Section 7 concludes.

2 Literature Review: The Principle Proliferation Problem

2.1 The Global Landscape of AI Ethics

The proliferation of AI ethics frameworks since 2015 is without precedent in the history of technology governance. [Jobin et al. \(2019\)](#) catalogued 84 documents and identified eleven clusters of recurrent principles; subsequent analyses suggest the total number now exceeds 160 ([Algorithm Watch, 2020](#)). [Hagendorff \(2020\)](#) evaluated these guidelines against philosophical criteria and found that while they exhibit convergence on high-level values, they systematically neglect institutional questions of implementation and enforcement. [Schiff et al. \(2021\)](#) analyzed documents from public, private, and NGO sectors and confirmed that principle articulation dominates while operationalization remains marginal.

This principle proliferation has been characterized as an exercise in “ethics washing”—the deployment of ethical language to forestall binding regulation ([Wagner, 2018](#); [Bietti, 2020](#)). While ethics washing is a genuine phenomenon, the critique risks oversimplification. Many principle frameworks were produced in good faith by organizations genuinely seeking governance guidance. The problem is structural rather than motivational: principles, by their nature, do not specify the instruments through which they will be achieved.

2.2 Why Principles Fail

[Mittelstadt \(2019\)](#) provides the most systematic critique of principle-based AI governance, arguing that principles alone cannot guarantee ethical AI because they lack the institutional supports that give medical ethics its practical force: professional licensure, shared professional identity, established enforcement mechanisms, and clear lines of accountability. AI development lacks each of these preconditions.

The critique has been extended in several directions. [Rességuier and Rodrigues \(2020\)](#) argued that AI ethics must recover its “teeth”—binding obligations with enforcement consequences. [Green \(2021\)](#) demonstrated that corporate AI ethics initiatives systematically avoid the structural reforms that would constrain profitable algorithmic systems. [Whittaker \(2021\)](#) showed that internal AI ethics teams within technology companies lack the organizational power to override business decisions, rendering their ethical oversight performative.

2.3 The Missing Instrument Dimension

What unites these critiques is a recognition—sometimes explicit, more often implicit—that the central problem of AI governance is not normative but instrumental. The ethics community has produced compelling accounts of what values should govern AI. What it has not produced is a systematic account of *how* these values should be institutionalized.

This is precisely the gap that policy instrument theory addresses. As [Linder and Peters \(1989\)](#) observed, the choice of policy instruments is among the most consequential decisions in governance design, yet it receives far less scholarly attention than the articulation of policy goals. In AI governance, this asymmetry is extreme: principle articulation has absorbed the overwhelming majority of scholarly and policy attention, while instrument design has been treated as a second-order implementation question.

The instrumental turn rejects this hierarchy. Instrument design is not an afterthought to principle articulation—it is the central challenge of governance.

3 Theoretical Framework: The Instrumental Turn

3.1 From Hood to Howlett: The Policy Instrument Tradition

The analysis of policy instruments has a substantial intellectual pedigree. [Hood’s \(1983\)](#) *The Tools of Government* identified four fundamental resources that governments deploy to achieve policy objectives, designated by the NATO acronym:

Table 2: Hood’s NATO Instrument Taxonomy Applied to AI Governance

NATO Category	Government Resource	Mechanism	AI Governance Application
Nodality	Information centrality	Government as information hub: collection, analysis, dissemination	Data registries; algorithmic transparency portals; public dashboards
Authority	Legal power	Binding rules: prohibitions, permissions, obligations	AI-specific regulation; conformity assessment; licensing
Treasure	Financial resources	Incentives and penalties: grants, taxes, subsidies, fines	R&D funding for trustworthy AI; regulatory sandboxes; penalties for non-compliance
Organization	Direct capacity	Government provision of services or functions	Public AI oversight bodies; government AI testing facilities; certification agencies

Hood’s taxonomy was not designed for AI governance, but its analytical power extends directly to this domain. Every AI governance arrangement deploys some combination of these four resources. The EU AI Act, for example, relies primarily on authority (binding classification rules), supplemented by organization (national competent authorities) and nodality (the European AI database). The US approach relies more heavily on treasure (R&D incentives) and authority wielded through existing sectoral regulators.

[Howlett \(2011\)](#) extended Hood’s taxonomy into a comprehensive theory of instrument choice, arguing that instrument selection is shaped by three factors: the nature of the policy problem, the institutional context of governance, and prior instrument commitments that constrain future choices. This “instrument choice” perspective yields an important insight for AI governance: the reason different jurisdictions adopt different governance instruments is not primarily ideological but institutional. Instrument choice is path-dependent, shaped by existing regulatory infrastructure, legal traditions, and institutional capacities.

[Salamon \(2002\)](#) added a further dimension by analyzing “third-party governance”—the use of non-governmental actors to achieve public objectives. In AI governance, third-party instruments are pervasive:

auditing is performed by private firms, standards are set by industry consortia, and compliance monitoring relies heavily on developer self-assessment. The instrumental turn requires systematic attention to the strengths and limitations of third-party arrangements.

3.2 The Instrumental Turn as Paradigm Shift

The instrumental turn is not merely a call for “more implementation” of existing principles. It represents a paradigm shift in how AI governance is conceived:

Table 3: Principle Paradigm vs. Instrument Paradigm

Dimension	Principle Paradigm	Instrument Paradigm
Central Question	What values should govern AI?	What mechanisms achieve governance objectives?
Primary Output	Ethical frameworks, guidelines, declarations	Enforceable rules, standards, procedures, penalties
Success Criterion	Normative consensus across stakeholders	Behavioral change in regulated entities
Failure Indicator	Principle disagreement or incompleteness	Principles endorsed but not operationalized
Analytical Focus	Value clarification, deliberation, consensus-building	Instrument design, selection, combination, enforcement
Expertise Required	Moral philosophy, stakeholder engagement	Regulatory design, institutional analysis, enforcement
Accountability	Reputational; voluntary reporting	Legal; administrative; financial sanctions
Mechanism		

The paradigm shift does not render the principle paradigm obsolete. Rather, it recognizes that principle articulation is a *necessary but insufficient* condition for effective governance. Principles provide the normative foundation; instruments provide the institutional superstructure.

What, then, does the instrumental turn add beyond the established policy instrument tradition? Three features distinguish AI governance from the regulatory domains in which Hood, Howlett, and Salamon developed their frameworks. First, **multi-stage temporal configuration**: AI systems function as autonomous agents embedded in complex social environments (Rahwan et al., 2019), exhibiting lifecycle-specific risks—training bias, deployment drift, emergent behavior—that require instruments distributed across *ex ante*, concurrent, and *ex post* phases in ways more demanding than traditional industrial regulation. Second, **principle decomposition**: composite ethical principles like “transparency” must be disaggregated into governance-relevant variants before instrument selection can proceed. Third, **institutional capacity asymmetry**: the technical expertise required for AI oversight resides overwhelmingly in the private sector (Busuioc, 2021), creating regulatory dependencies that the traditional instrument literature does not adequately address.

The instrumental turn thus situates AI governance squarely within the **policy design and regulatory studies** tradition, rather than within ethics or science and technology studies (STS) alone. Its analytical vocabulary—instrument taxonomies, resource bundles, institutional preconditions, enforcement mechanisms—is drawn from the policy sciences (Linder & Peters, 1989; Howlett, 2011), and its evaluative criteria are institutional rather than philosophical.

3.3 Policy Instrument Taxonomy for AI Governance

Drawing on the instrument choice literature, we identify five primary instrument categories applicable to AI governance:

These five instruments differ not only in mechanism but in their relationship to the policy cycle. Prohibition and licensing operate *ex ante*—before system deployment. Disclosure operates *during* deployment. Audit may operate at any stage. Liability operates primarily *ex post*—after harm has materialized. Effective governance regimes combine instruments across the temporal spectrum.

A clarification of the relationship between Hood’s NATO resource primitives and the five-instrument taxonomy is warranted. The NATO categories are *resource primitives*—the raw materials from which gov-

Table 4: AI Governance Instrument Taxonomy

Instrument	Mechanism	Hood Cat.	Examples	Strengths	Limitations	Preconditions
Prohibition	Categorical ban on defined activities	Authority	EU AI Act Art. 5; biometric moratoriums	Maximum certainty; clear compliance standard	Rigidity; over/under-inclusion	Clear definitions; enforcement capacity
Licensing	Pre-market authorization	Authority + Org.	Conformity assessment; FDA-style approval	Gatekeeping; quality control	Barriers to entry; resource-intensive	Technical review capacity; admin infrastructure
Disclosure	Mandatory information provision	Nodality + Auth.	Model cards; impact assessments	Light-touch; market-enabling	Audience capacity; strategic compliance	Standardized formats; verification
Audit	Independent compliance verification	Org. + Auth.	Third-party bias audits; conformity verification	Independent verification; expertise	Capture risk; access barriers	Auditor accreditation; access mandates
Liability	Legal consequences for harmful outcomes	Auth. + Treasure	Product liability; strict liability; admin fines	Compensation; deterrence	Causation difficulties; resource asymmetry	Accessible forums; evidentiary frameworks

ernments construct governance mechanisms. The five instruments identified here are *governance mechanisms assembled from these primitives*, each drawing on a distinctive resource bundle. This distinction matters because instrument selection is constrained not only by policy objectives but by the resource bundles available to regulators.

3.4 Instrument Interactions and Combinations

Policy instruments do not operate in isolation. Effective governance regimes combine instruments in mutually reinforcing configurations. The instrument choice literature ([Howlett & Rayner, 2007](#)) identifies several interaction patterns:

Table 5: Instrument Interaction Patterns

Combination	Interaction Effect	Concrete Application
Disclosure + Liability	Disclosure preserves evidence for subsequent liability claims	Mandatory incident reporting creating evidentiary record for litigation
Licensing + Audit	Licensing establishes baseline; audit verifies ongoing compliance	Pre-market approval followed by periodic third-party verification
Prohibition + Liability	Prohibition defines violation; liability provides remedy and deterrence	Banned AI practice + civil action for victims + administrative penalty
Disclosure + Audit	Disclosure provides audit material; audit verifies disclosure accuracy	Impact assessments subjected to independent review
Licensing + Disclosure	Licensing generates registry; registry supports transparency	Authorized AI systems listed in public database with deployment data

These interactions are not automatic—they must be designed. A common governance failure is “instrument layering” without attention to interaction effects, producing regulatory complexity without corresponding governance improvement.

4 The Principle-Instrument Matrix

4.1 Matrix Design

The Principle-Instrument Matrix maps ethical principles onto policy instruments, identifying which instruments can operationalize which principles and under what conditions. Two design choices warrant explanation. First, the matrix decomposes composite principles into distinct variants. “Transparency,”

for example, is disaggregated into process, model, outcome, and reason transparency, because each variant requires different instrumental approaches. Second, the matrix distinguishes primary from secondary instruments.

Table 6: Transparency Variant Decomposition

Variant	Meaning	Target Audience	Instrumental Approach
Process	How a system was developed: data sourcing, model selection, validation procedures	Regulators, auditors	Documentation requirements; audit access mandates
Model	How a system generates outputs: algorithmic logic, feature weights, decision pathways	Technical experts, researchers	Open-source requirements; technical disclosure standards
Outcome	What decisions a system makes: deployment scope, decision rates, error distributions	Affected parties, civil society	Notice requirements; decision records; public dashboards
Reason	Why a system reached a specific decision: causal factors, feature contributions	Affected individuals; appeal bodies	Explanation requirements; interpretable models (Rudin, 2019); contestability procedures

The distinction among transparency variants is not merely taxonomic. Current governance frameworks often invoke “transparency” without specifying which variant is required. The EU AI Act requires “transparency” for certain systems but operationalizes this primarily through user notification—addressing outcome transparency while leaving process, model, and reason transparency largely unaddressed (Veale & Zuiderveen Borgesius, 2021).

4.2 The Complete Principle-Instrument Matrix

Table 7: Principle-Instrument Matrix

Principle	Prohibition	Licensing	Disclosure	Audit	Liability
Transparency (Process)	—	• Entry condition	• Documentation mandates	• Process auditing	◦ Evidence preservation
Transparency (Model)	—	◦ Technical review	• Open-source / specs	• Technical audit	◦ Discovery proceedings
Transparency (Outcome)	—	◦ Registry obligations	• Decision notification	◦ Outcome monitoring	• Compensation proceedings
Transparency (Reason)	—	◦ Explainability gates	• Explanation provision	• Explanation verification	• Contestability rights
Accountability	◦ Removes unaccountable systems	• Assigns responsible party	• Responsibility declaration	• Compliance verification	• Consequence imposition
Fairness (Group)	• Prohibit discriminatory uses	◦ Conditional approval	• Disparity disclosure	• Bias auditing	• Disparate impact claims
Fairness (Individual)	—	—	◦ Limited utility	◦ Case review	• Individual claims
Fairness (Procedural)	—	• Process requirements	• Process disclosure	• Process auditing	◦ Procedural challenge
Safety	• Prohibit dangerous applications	• Safety assessment	• Risk disclosure	• Safety auditing	• Harm compensation
Privacy	• Prohibit invasive surveillance	• Data processing authorization	• Data practice disclosure	• Data protection auditing	• Data breach liability

Legend: • = Primary instrument; ◦ = Secondary/supporting instrument; — = Not applicable

4.3 Reading the Matrix: Patterns and Implications

The Principle-Instrument Matrix reveals several structurally important patterns.

Pattern 1: Multi-Instrument Necessity. No single instrument adequately operationalizes any principle. Transparency requires disclosure, audit, and in many cases liability. Single-instrument approaches

systematically underperform. This finding aligns with [Gunningham and Grabosky’s \(1998\)](#) analysis of “smart regulation,” which demonstrates that instrument combinations outperform individual instruments.

Pattern 2: Principle Decomposition Imperative. Composite principles must be decomposed into variants before instrument selection proceeds. The instruments appropriate for process transparency differ fundamentally from those appropriate for reason transparency.

Pattern 3: Institutional Prerequisite Chains. Each instrument carries institutional prerequisites. Licensing requires administrative infrastructure; audit requires competent, independent auditors; liability requires accessible judicial or administrative forums. Instrument selection must begin with institutional capacity assessment ([Black, 2001](#); [Coglianese, 2017](#)).

Pattern 4: Temporal Distribution. Effective governance distributes instruments across the policy cycle: ex ante (prohibition, licensing), concurrent (disclosure, audit), and ex post (liability). Governance regimes that concentrate on a single temporal phase leave governance gaps for post-deployment harms.

Pattern 5: Instrument-Principle Feedback. The matrix is not a one-directional mapping. Instrument implementation generates feedback that refines principles. When bias auditing reveals that “group fairness” admits multiple incompatible statistical definitions ([Chouldechova, 2017](#)), the instrument experience compels principle refinement. This feedback loop is consistent with the experimentalist governance framework ([Sabel & Zeitlin, 2012](#)).

4.4 Application: Operationalizing Accountability

Consider accountability—a principle universally endorsed but persistently under-operationalized. The Principle-Instrument Matrix suggests a multi-instrument configuration:

Table 8: Accountability: Multi-Instrument Configuration

Instrument	Accountability Function	Implementation Mechanism
Licensing	Assigns responsible party ex ante	Registration requirements designating a legally accountable entity for each deployed system
Disclosure	Preserves accountability information	Mandatory documentation of design decisions, training data, deployment contexts, known limitations
Audit	Verifies accountability practices in operation	Third-party review of responsibility allocation, override procedures, incident response protocols
Liability	Imposes consequences ex post	Legal remedies—administrative penalties, civil damages—for failures in accountability practice

This four-instrument configuration addresses accountability more comprehensively than any single mechanism. Licensing creates an identifiable responsible party. Disclosure preserves information necessary for accountability determinations. Audit verifies that accountability mechanisms function in practice. Liability provides the enforcement “teeth” that voluntary accountability pledges demonstrably lack.

5 Comparative Analysis: EU, OECD, and UNESCO Instrumentalization

5.1 The European Union: Authority-Centered Instrumentalization

The EU AI Act ([European Parliament, 2024](#)) represents the most comprehensive attempt at AI governance instrumentalization to date. Its instrument profile emphasizes authority: binding prohibitions (Article 5), mandatory conformity assessment for high-risk systems (Chapter III), and administrative penalties for non-compliance (Article 99). This authority-centered approach reflects the EU’s broader regulatory tradition and its institutional capacity for standard-setting and enforcement ([Bradford, 2020](#)).

To illustrate how the Principle-Instrument Matrix applies in practice, consider the EU AI Act’s Article 5 prohibited-practices regime:

Table 9: EU AI Act Article 5: Principle-to-Instrument Trace

Principle Origin	Article 5 Prohibition	Instrument Type	Enforcement Pathway
Safety + Dignity	Social scoring by public authorities	Prohibition (categorical ban)	Maximum administrative fines
Autonomy + Privacy	Real-time remote biometric identification in public spaces (except narrow exceptions)	Prohibition + Licensing (prior authorization for exceptions)	Market access denial; referral under national law
Fairness + Non-manipulation	Subliminal techniques causing harm	Prohibition	Administrative penalties
Safety + Vulnerability protection	Exploitation of vulnerable groups	Prohibition	Administrative penalties

Through the lens of the Principle-Instrument Matrix, the EU approach exhibits notable strengths and gaps:

Table 10: EU AI Act: Principle-Instrument Gap Analysis

Principle	EU Instrument Coverage	Gap Analysis
Safety	• Licensing (conformity assessment); • Prohibition (Art. 5)	Strong ex ante coverage; ex post liability partially addressed by the proposed AI Liability Directive (Hacker, 2023)
Transparency	◦ Disclosure (limited user notification)	Process and model transparency under-addressed; reason transparency nascent
Accountability	• Licensing (responsible person designation)	Audit mechanisms still developing; AI Liability Directive would add ex post consequences
Fairness	◦ Implicit through non-discrimination provisions	No dedicated bias audit instrument; fairness conceptions unspecified

The EU's approach is strongest on safety and weakest on fairness—a pattern consistent with its choice to instrument primarily through authority rather than through audit or liability mechanisms.

5.2 The OECD: Nodality-Centered Instrumentalization

The OECD AI Principles ([OECD, 2019](#)) represent an alternative instrumentalization strategy centered on nodality. The OECD lacks binding authority; its instruments are persuasive rather than coercive.

Table 11: OECD AI Principles: Instrument Specificity Assessment

OECD Principle	Recommended Approach	Instrument Specificity
Transparency	“Foster understanding”	Low—no mandatory disclosure, audit, or explanation requirements
Accountability	“Ensure accountability”	Low—no designated responsible party, enforcement mechanism, or sanction
Robustness	“Function appropriately”	Low—no safety assessment, testing protocol, or certification
Inclusive Growth	“Benefit people and the planet”	Minimal—aspirational framing without operational definition
Human-Centred Values	“Respect rule of law, human rights”	Low—references existing legal frameworks without AI-specific instruments

5.3 UNESCO: Treasure and Organization

The UNESCO Recommendation on the Ethics of Artificial Intelligence ([UNESCO, 2021](#)) adds distinctive instruments, particularly in the treasure and organization categories. UNESCO's emphasis on capacity building reflects a broader recognition that AI governance must address not only risks but also AI's potential contributions to sustainable development ([Vinuesa et al., 2020](#)):

Table 12: UNESCO AI Ethics Recommendation: Instrument Profile

UNESCO Instrument	Hood Category	Innovation	Limitation
Ethical Impact Assessment	Organization	Structured assessment methodology	Voluntary; no enforcement mechanism
Readiness Assessment	Org. + Nodality	National capacity evaluation	Diagnostic only; does not compel action
Multi-stakeholder governance	Organization	Inclusive deliberation	Slow; consensus-dependent; capture risk
Capacity building programs	Treasure + Org.	Institutional development in Global South	Resource-dependent; sustainability uncertain

5.4 Comparative Synthesis

Table 13: Comparative Instrumentalization: EU, OECD, and UNESCO

Dimension	EU	OECD	UNESCO
Primary Hood Category	Authority	Nodality	Organization + Treasure
Binding Force	Legally binding regulation	Soft law (recommendations)	Soft law (recommendation)
Instrument Specificity	High (conformity assessment, classification, penalties)	Low (general principles)	Medium (impact assessments, capacity building)
Enforcement Mechanism	Administrative penalties; market access denial	Peer review; reputational	None specified
Geographic Scope	EU member states (+ Brussels Effect)	38 OECD members	193 UNESCO members
Principle Coverage	Safety-centered	Broad but shallow	Ethics-centered; Global South oriented
Instrumental Turn Progress	Advanced	Nascent	Partial

The comparative analysis demonstrates that the instrumental turn is proceeding unevenly across jurisdictions and international organizations. The EU has advanced furthest in translating principles into enforceable instruments, but with notable coverage gaps.

6 Institutional Implications: Preconditions for Instrumentalization

6.1 The Auditability Standard

Effective instrumentalization requires that principles be translated into *auditable* criteria—criteria that can be assessed by independent parties against observable evidence. A terminological clarification is warranted: “auditability” here refers not to *audit as a policy instrument* but to a *translation condition*—the property that makes a principle amenable to instrumentalization through any mechanism.

Table 14: The Auditability Standard

Criterion	Description	Test Question	Failure Indicator
Definitional Clarity	Principle meaning is operationally unambiguous	Can independent parties agree on what the principle requires in specific cases?	Multiple conflicting interpretations persist among regulators
Observable Indicators	Compliance can be empirically assessed	Are there measurable indicators of principle satisfaction?	No measurement methodology exists or is agreed upon
Documentary Trail	Evidence of compliance can be preserved and inspected	Can compliance documentation be produced and retained for review?	Compliance is claimed but not documentable
Third-Party Verifiability	External parties can independently assess compliance	Can auditors with appropriate access verify claims?	Verification requires insider access that cannot be provided
Consequence Specification	Non-compliance carries defined, proportionate consequences	What happens when a violation is detected?	No enforcement mechanism; violation without sanction

Application: Auditing the Core Principles

Table 15: Auditability Scoring of Core AI Governance Principles

Principle	Def. Clarity	Observable	Doc. Trail	3rd-Party	Consequence	Score
Process Transp.	✓	✓	✓	✓	Varies	4.5/5
Reason Transp.	○	○	✓	○	Varies	3/5
Group Fairness	○	✓	✓	✓	Varies	3.5/5
Individual Fairness	×	×	○	×	Rarely	1/5
Accountability	○	○	✓	○	Varies	3/5
Safety	✓	✓	✓	✓	✓	5/5

Legend: ✓ = Satisfied; ○ = Partially satisfied; × = Not satisfied

Safety achieves the highest auditability score because it benefits from established engineering traditions: safety criteria can be defined, measured, documented, independently verified, and enforced. Individual fairness achieves the lowest score because there is no consensus on operational definition, no agreed measurement methodology, and limited third-party verification capacity. This divergence explains why safety-oriented instruments have advanced furthest while fairness instruments remain contested.

Two important caveats qualify the Auditability Standard. First, auditability is a *necessary but not sufficient* condition for good governance. Second, the standard carries a normative risk: by privileging measurable, verifiable principles, it may produce a “**streetlight effect**”—encouraging regulators to concentrate enforcement on easily audited dimensions while marginalizing dimensions that are ethically important but difficult to quantify. Low auditability should trigger investment in operationalization methods—not abandonment of the principle.

6.2 Failure Diagnosis Framework

Table 16: Institutional Failure Diagnosis

Failure Mode	Cause	Diagnostic Indicator	Real-World Example	Remedial Approach
Principle Proliferation	Normative activity substituting for institutional action	160+ frameworks, minimal behavioral change	Global AI ethics landscape (Jobin et al., 2019)	Shift resources to instrument design
Implementation Gap	Instruments designed but capacity insufficient	Formal requirements exist; compliance is cosmetic	GDPR Art. 22 provisions	Capacity building; institutional investment
Enforcement Deficit	Violations detected but consequences not imposed	Documented non-compliance without sanction	Pre-AI Act voluntary framework violations	Dedicated enforcement resources; penalty escalation
Definition Failure	Principle too vague for operational translation	Persistent interpretive disagreement	“Fairness” across statistical definitions (Chouldechova, 2017)	Decompose into verifiable sub-components
Gaming and Evasion	Strategic compliance satisfying letter but not spirit	Formal metrics satisfied; outcomes unchanged	Model cards that disclose without informing (Mitchell et al., 2019)	Outcome-oriented verification

6.3 The Capacity Challenge

A persistent obstacle to the instrumental turn is institutional capacity. Effective instrument deployment requires expertise, resources, and organizational infrastructure that many regulatory bodies currently lack.

Busuioc (2021) documented the “algorithmic capacity gap” faced by regulators: the technical expertise needed to oversee AI systems resides predominantly in the private sector.

The capacity challenge suggests that instrument selection must be calibrated to institutional realities. This does not counsel delay in instrumentalization. Rather, it counsels phased instrument deployment: beginning with instruments that match current institutional capacities while building capacity for more sophisticated instruments over time.

Table 17: Instrument Selection Matrix (Auditability $\times$ Enforcement Capacity)

	High Enforcement Capacity	Low Enforcement Capacity
High Auditability (score 4–5/5)	Authority-first pathway: Prohibition + Licensing + Audit. <i>Example:</i> Safety under the EU AI Act.	Disclosure-first pathway: Mandatory disclosure + Liability. <i>Example:</i> Safety in jurisdictions without dedicated AI agencies.
Low Auditability (score 1–3/5)	Sandbox-first pathway: Experimental governance + Capacity investment. <i>Example:</i> Fairness in the EU through sandbox programs.	Nodality-first pathway: Information provision + Norm-setting + International coordination. <i>Example:</i> Individual fairness under UNESCO/OECD soft governance.

This matrix transforms the Principle-Instrument Matrix from a static mapping into a **decision algorithm** for instrument selection.

6.4 Sandboxes and Experimental Governance

Regulatory sandboxes—controlled environments for testing AI governance instruments—represent a distinctive approach to the capacity challenge. The EU AI Act (Article 57) provides for AI regulatory sandboxes. Sandboxes operationalize the instrumental turn in miniature: they are meta-instruments—instruments for evaluating instruments. Their value lies in enabling iterative instrument design based on empirical evidence, consistent with “experimentalist governance” (Sabel & Zeitlin, 2012).

6.5 Testable Propositions

The instrumental turn generates empirically testable propositions:

Proposition 1 (Auditability-Adoption Nexus): Principles with higher auditability scores are more likely to be translated into binding legal obligations. *Operationalization:* the dependent variable can be coded as presence of “shall + verifiable requirement + sanction clause” in legislative text; the independent variable is auditability score as rated by regulatory analysts.

Proposition 2 (Instrument Combination Superiority): Governance regimes that combine disclosure, audit, and liability instruments achieve higher compliance rates than regimes relying on any single instrument. *Operationalization:* governance outcomes approximated through noncompliance rates and enforcement actions across matched jurisdictional pairs.

Proposition 3 (Capacity-Instrument Alignment): Jurisdictions with lower regulatory capacity disproportionately adopt nodality and treasure instruments rather than authority and organization instruments. *Operationalization:* regulatory capacity proxies include supervisory budget, technical staff per regulated entity, and presence of a dedicated AI oversight body.

6.6 Boundary Conditions: The Risk of Instrumental Overfitting

The instrumental turn carries its own pathologies. An exclusive focus on instrument design can produce what we term “**instrumental overfitting**”—a governance condition in which the instrument apparatus becomes so elaborate that it generates compliance costs, strategic gaming, and institutional rigidities that undermine the very principles the instruments were designed to serve. Four risk pathways deserve attention.

Risk 1: Compliance Theatre. When instrument compliance becomes an end in itself, organizations may invest in the *appearance* of compliance rather than its substance (Mitchell et al., 2019; Bietti, 2020).

Risk 2: Audit Capture. Third-party auditing depends on auditor independence, but the economics of the audit market may undermine this independence—a dynamic well-documented in financial regulation (Power, 1997).

Risk 3: Standard Laundering. If the standard-setting process is captured by industry incumbents (Büthe & Mattli, 2011), the resulting instruments may codify industry-preferred practices rather than genuinely protective governance requirements.

Risk 4: Innovation Chilling. Liability instruments, particularly strict liability regimes, may deter beneficial AI innovation if liability exposure is disproportionate (Hutt et al., 2014).

Instrumental overfitting does not invalidate the instrumental turn—it identifies its boundary conditions. These failure modes should be understood as a **checklist for instrument design review**.

7 Conclusion

AI ethics has achieved remarkable normative consensus. Yet governance outcomes have not followed—not because principles are wrong but because principles are not self-executing.

This article has argued that closing the principle–practice gap requires an “**instrumental turn**”—a paradigm shift from principle articulation to instrument operationalization. Drawing on Hood’s NATO taxonomy and Howlett’s instrument choice theory, we have developed a five-category instrument taxonomy and a **Principle–Instrument Matrix** that maps ethical principles onto enforceable mechanisms, revealing that effective operationalization requires multi-instrument configurations distributed across the policy cycle.

The **Auditability Standard** provides diagnostic criteria for translation readiness. It is a necessary but not sufficient condition: principles that satisfy its five criteria are operationally ready for instrumentalization, but effective governance still requires normative judgment about priority and proportionality. The standard’s divergent scores—safety at 5/5, individual fairness at 1/5—explain the uneven progress of AI governance instrumentalization and identify where further conceptual work is needed.

The instrumental turn does not diminish the importance of ethical principles. It recognizes that principles achieve governance impact only through institutional embedding—but also that instrumentalization itself must be governed, lest it produce the compliance theatre, audit capture, and standard laundering pathologies identified in §6.6. Policymakers may adopt the Principle–Instrument Matrix as an analytical framework for translating AI ethics principles into actionable regulatory design, guided by four immediate priorities:

1. **Conduct auditability assessments** of existing principle commitments to identify which are ready for instrumentalization and which require prior operationalization work.
2. **Select instrument pathways** using the Auditability × Capacity matrix (Table 17) to match governance ambition with institutional reality.
3. **Redesign instrument configurations** using the Principle–Instrument Matrix to ensure multi-instrument coverage across the policy cycle, calibrated to institutional capacity and SME proportionality.
4. **Test instrument combinations** in regulatory sandboxes, guided by the three testable propositions advanced in this article, to build empirical evidence on instrument effectiveness—while monitoring for instrumental overfitting.

Bridging the chasm between ethical aspiration and institutional reality is the defining governance challenge of the algorithmic age. The instrumental turn provides the analytical architecture for that bridge.

References

- Algorithm Watch. (2020). *AI Ethics Guidelines Global Inventory*. Algorithm Watch. <https://inventory.algorithmwatch.org/>
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. *ProPublica*, May 23.
- Baldwin, R., Cave, M., & Lodge, M. (2011). *Understanding regulation: Theory, strategy, and practice* (2nd ed.). Oxford University Press.
- Bietti, E. (2020). From ethics washing to ethics bashing: A view on tech ethics from within moral philosophy. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 210–219.
- Black, J. (2001). Decentring regulation: Understanding the role of regulation and self-regulation in a 'post-regulatory' world. *Current Legal Problems*, 54(1), 103–146.
- Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. Oxford University Press.
- Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review*, 81(5), 825–836. <https://doi.org/10.1111/puar.13293>
- Büthe, T., & Mattli, W. (2011). *The new global rulers: The privatization of regulation in the world economy*. Princeton University Press.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153–163. <https://doi.org/10.1089/big.2016.0047>
- Coglianesi, C. (2017). The limits of performance-based regulation. *University of Michigan Journal of Law Reform*, 50(3), 525–563.
- European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
- Green, B. (2021). The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing*, 2(3), 209–225.
- Gunningham, N., & Grabosky, P. (1998). *Smart regulation: Designing environmental policy*. Oxford University Press.
- Hacker, P. (2023). The European AI liability directives—Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, 51, 105871. <https://doi.org/10.1016/j.clsr.2023.105871>
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Hood, C. C. (1983). *The tools of government*. Macmillan.
- Howlett, M. (2011). *Designing public policies: Principles and instruments*. Routledge.
- Howlett, M., & Rayner, J. (2007). Design principles for policy mixes: Cohesion and coherence in 'new governance arrangements.' *Policy and Society*, 26(4), 1–18.
- Hutt, P. B., Merrill, R. A., & Grossman, L. A. (2014). *Food and drug law: Cases and materials* (4th ed.). Foundation Press.
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
- Linder, S. H., & Peters, B. G. (1989). Instruments of government: Perceptions and contexts. *Journal of Public Policy*, 9(1), 35–58.
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Ge-

- bru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.
- Power, M. (1997). *The audit society: Rituals of verification*. Oxford University Press.
- Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.
- Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Résséguier, A., & Rodrigues, R. (2020). AI ethics should not remain toothless! A call to bring back the teeth of ethics. *Big Data & Society*, 7(2). <https://doi.org/10.1177/2053951720942541>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Sabel, C. F., & Zeitlin, J. (2012). Experimentalist governance. In D. Levi-Faur (Ed.), *The Oxford handbook of governance* (pp. 169–184). Oxford University Press.
- Salamon, L. M. (Ed.). (2002). *The tools of government: A guide to the new governance*. Oxford University Press.
- Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2021). AI ethics in the public, private, and NGO sectors: A review of a global document collection. *IEEE Transactions on Technology and Society*, 2(1), 31–42. <https://doi.org/10.1109/TTS.2021.3052127>
- UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. UNESCO.
- Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., ... & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, 11, 233. <https://doi.org/10.1038/s41467-019-14108-y>
- Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act. *Computer Law Review International*, 22(4), 97–112.
- Wagner, B. (2018). Ethics as an escape from regulation: From “ethics-washing” to ethics-shopping? In E. Bayamlioglu et al. (Eds.), *Being profiled* (pp. 84–90). Amsterdam University Press.
- Whittaker, M. (2021). The steep cost of capture. *Interactions*, 28(6), 50–55.
- Yeung, K. (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework. *Council of Europe Study DGI(2019)05*.
Academic Editor Xin Huang, cpcl@cpcl.hk