Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908
2026 Volume 7, Issue 1: 68–86

DOI: https://doi.org/10.48185/jaai.v7i1.1954

Infrastructure-Mediated Multilateralism: A Technical Architecture for International AI Governance

Kao-Cheng Huang[1*]
Chinese Association of Mere-Consciousness, Miaoli 35443, Taiwan

Received: 10.01.2025 • Accepted: 15.04.2026 • Published: 04.06.2026 • Final Version: 30.06.2026

Abstract: Global rules for AI are clearer than ever, and most countries agree on transparency, fairness, and accountability. That agreement, however, has not stopped ethics washing, concentrated power, or dangerous uses. This paper argues that the core issue is not insufficient agreement but the absence of infrastructure to turn principles into testable action. We propose the Multilateral AI Commons for Peace and Sustainability (MACPS): a five-part technical architecture that addresses three governance problems: coordination breakdowns across scattered rules, unequal AI capacity between the Global North and South, and weak trust that reduces compliance to performance. Drawing on Huayan Buddhism's concept of perfect interfusion (mutual constitution without loss of identity), we argue that shared infrastructure enables multilateral cooperation through material interdependence rather than shared belief alone. Because technical systems always carry values, we design MACPS to embed equity as a structural property. The architecture uses a shared semantic ontology for interoperability, pools compute on equitable terms, holds data in federated trusts, verifies performance at the mechanism level, and distributes machine-readable implementation playbooks. Shared infrastructure delivers governance outcomes that principles alone cannot secure, pointing the way to stronger cooperation even as geopolitics pulls countries apart.

Keywords: Multilateralism, Polycentric governance, Digital commons, Federated systems, Global South, Multilateral AI Commons for Peace and Sustainability (MACPS)

1. Introduction

Global AI governance does not lack principles; it lacks the infrastructure to make them work. Between 2016 and 2019, 84 AI ethics frameworks appeared, converging on transparency, fairness, non-harm, accountability, and privacy [1]. By 2025, however, ethics claims continue to proliferate while computing power remains concentrated in a few hands and dangerous AI is deployed with little accountability. This structural gap reflects a deeper failure: governance frameworks do not translate into action.

The consequences are visible. Five companies (Google, Microsoft, Meta, Amazon, and Baidu) run roughly 72% of the world's frontier AI compute, and three countries (the US, China, and the UK) supply more than 85% of the large-scale infrastructure used to train frontier models [2]. Current audits lack standardized procedures and independent verification [3]. From 2016 to 2023, the AI Incident Database recorded 1,342 cases of AI-related harm, yet fewer than 15% resulted in institutional accountability [4]. The problem is not a shortage of ideas; it is the absence of means to enact them.

[1*] Corresponding Author: k.huang@univ.oxon.org, ORCID: 0009-0008-6803-4473D

Current rules describe what should happen but not how. The EU AI Act [5] sorts AI by risk yet provides no compute for assessments. UNESCO's AI ethics recommendation promises fairness and oversight that remain unverifiable in practice [6]. The G7 Hiroshima Process code of conduct promotes responsible development but supplies no accountability framework [7]. Goals without mechanisms remain aspirations.

MACPS addresses three governance gaps directly. The coordination gap arises because rules split across jurisdictions force organizations to meet overlapping requirements without a way to share systems, favoring well-funded players. The capability gap persists because the Global North holds most of the compute, data, and technical expertise, excluding affected communities from meaningful participation [8, 9]. The credibility gap reflects an accountability failure: companies claim ethical goals while avoiding proof [10].

MACPS builds infrastructure, not norms. Semantic interoperability lets organizations meet every rule through a single mapping. Equity-based compute allocation directs real support to underfunded institutions. Verification of actual behavior, rather than self-reported compliance, generates trust. Shared material systems, not prior agreement, create durable ties.

The core argument is that states cooperate more reliably through shared technical systems that create real economic interdependence than through common beliefs or shared threats. Huayan Buddhism's notion of perfect interfusion captures a form of cooperation that standard Western frameworks do not fully express [11]: mutual shaping through shared infrastructure while each participant retains its identity. This differs from both harmonization and simple coordination.

Infrastructure carries rules and power dynamics [12–14]. Treating this as an opportunity for deliberate design (of values, methods, and accountability) opens important governance possibilities and clarifies how design choices shape outcomes.

This work shows how shared technical tools can narrow the gap between agreed principles and actual practice, contributing to polycentric [15] and global AI governance research [16–18]. The five-layer design includes technical specifications, a comparison with existing initiatives, testable claims supported by pre-specified statistical methods, and governance rules that seek effectiveness while protecting fair representation.

Section 2 reviews the literature. Section 3 establishes the theoretical foundations. Section 4 presents the system architecture. Section 5 offers comparative analysis. Section 6 describes pilot applications. Section 7 sets out governance and evaluation. Section 8 addresses limitations. Section 9 concludes.

2. Literature review

2.1. AI governance architectures and institutional design

Research on AI governance identifies a critical trilemma: centralized, distributed, or polycentric approaches [16]. Each carries trade-offs in trustworthiness, flexibility, and enforcement; none has stopped ethics washing or ensured fair access to compute [16]. Taeihagh [17] proposes a governing approach combining foresight, flexibility, and experimentation. Ulnicane et al. [18] argue that governing AI is harder because it remains a contested emerging technology whose framing is itself disputed.

Maas and Villalobos [19] review major institutions worldwide and report many ideas but few implementation plans. Roberts et al. [20] separate problems into rivalry and broken institutions, arguing that both require simultaneous attention.

Ho et al.'s [21] IAEA-for-AI model for inspection and monitoring, and Trager et al.'s [22] jurisdiction-based certification framework, represent important progress. Like MACPS, both emphasize implementation, but their focus on regulation over capability provision leaves Global South participation largely unaddressed.

2.2. Digital commons and platform governance

Ostrom's [15] work on commons governance provides the theoretical basis for institutional arrangements operating beyond market and state structures. Benkler [23] shows that networked information production sustains public goods without centralized coordination. Subsequent scholarship has applied digital commons principles to AI-related resources, including training datasets, computational infrastructure, and evaluation benchmarks. Couldry and Mejias [24] argue that contemporary data governance extracts value from marginalized communities through data colonialism. This perspective grounds MACPS's data trust architecture and its sovereignty-with-solidarity principle.

2.3. Critical infrastructure studies and the politics of design

Technical systems are shaped by politics, not neutrality. Winner [12] shows that everyday objects carry political choices; Star [13] demonstrates that infrastructure highlights some people while marginalizing others. Crawford [14] argues that AI systems carry enormous planet-wide costs the field has long underestimated. Benjamin [25] coined the "New Jim Code" to describe how computer systems silently encode and reproduce racial hierarchies. Zuboff [26] documents how corporate AI systems construct surveillance and extract value from users. Together, these perspectives inform MACPS's notion of constitutional infrastructure. The implication is clear: equity must be built into structure, because treating it as an unspoken corporate aspiration has proven insufficient. Linking ongoing support to concrete work and skill-building makes redistribution a route to self-reliance rather than dependence.

2.4. Global South participation and epistemic justice in AI governance

Global South voices are systematically excluded from AI governance. Png [8] shows that Global South participants face real limits in governance spaces shaped by Northern rules. Effoduh [9] demonstrates that supposedly universal concepts such as explainability carry hidden assumptions that marginalize non-Western ways of knowing. Jobin, Ienca, and Vayena [1] find that, across 84 AI ethics guidelines, strong agreement on key principles coexists with near-total absence of Global South authorship. Inclusion alone is insufficient. Meaningful participation requires real resources: compute, well-chosen data, and technical expertise. Tracking concrete indicators (the share of governance work authored by Global South scholars, linguistic diversity in training data, and the representation of non-Western knowledge traditions in evaluation benchmarks) puts epistemic justice into practice. MACPS addresses these gaps through an equity-based allocation mechanism and a federated architecture, so that redistribution is built into the system rather than left to discretion.

2.5. Privacy-preserving computation and federated systems

Federated systems enable cross-border private data analysis. The technical groundwork for federated, privacy-preserving computation is now well established. Dwork and Roth [27] provide the mathematical foundations of differential privacy for protecting individual records. Bonawitz et al. [28] demonstrate practical secure aggregation methods for federated machine learning, enabling joint analysis of private data across jurisdictions without a single trusted intermediary. MACPS fills a major gap by integrating these technical tools into a unified governance system. Earlier research treated privacy-preserving computing as a purely technical problem and data sharing as a purely political one. Technical work frames data sharing as a cryptographic puzzle; governance work treats it as a political negotiation over control and consent. MACPS bridges this divide by addressing the mutual dependence between technical design choices and institutional rules that both traditions have overlooked.

3. Theoretical foundations

3.1. The limits of normative governance

Governance scholarship on regime complexity [16, 19], technical safety and alignment research [29, 30], and regulatory harmonization [31] shares a common assumption: that the core problem is securing normative convergence among actors with divergent interests. The evidence challenges this view. Despite convergence across 84 ethics guideline sets, effective governance outcomes have not followed. Agreement on principles is necessary but insufficient; operationalization requires infrastructure that principles alone cannot provide.

Three architectures dominate international AI governance proposals: centralized, distributed, and polycentric [16]. Centralized approaches offer consistency and enforcement strength but face legitimacy challenges and limited adaptability. Distributed approaches preserve sovereignty and enable experimentation but suffer from coordination failures, regulatory arbitrage, and downward pressure on standards. Polycentric approaches, grounded in Ostrom's [15] commons work, balance flexibility and coordination, but they depend on infrastructure capable of managing overlapping authorities and resolving competing claims.

Recent scholarship distinguishes first-order cooperation problems (arising from interstate competition and enforcement uncertainty) from second-order problems concerning institutional inability to adapt to rapid technological change [20]. A first-order example is states' reluctance to share safety-relevant AI research for fear of strategic disadvantage. A second-order example is the inability of bodies such as the OECD to update classification frameworks at the pace of model deployment. The proliferation of AI ethics guidelines illustrates these dynamics: convergence on principles is substantial, yet ethics washing remains prevalent because no mechanism distinguishes substantive commitment from performative compliance. Transparency requires audit infrastructure; fairness depends on bias detection and remediation; accountability needs verification procedures and enforcement capacity. Without such infrastructure, principles remain what Smuha and Yeung [32] describe as widely accepted but lacking practical substance. Operationalization is the key missing element in current frameworks.

3.2. Digital commons as governance infrastructure

The digital commons framework offers a theoretically grounded alternative to market-based and state-centric governance of shared resources [23]. Standard economic theory predicts that common-pool resources face overexploitation without private ownership or state regulation: Hardin's tragedy of the commons. Ostrom's [15] empirical research shows instead that communities manage such resources effectively through institutional arrangements outside both models. Her case studies across fisheries, irrigation systems, forests, and pastures identify eight design principles associated with durable commons institutions.

These principles translate directly to digital AI governance infrastructure. Clearly defined boundaries separate authorized from unauthorized participants and are implemented through MACPS membership criteria and API authentication. Proportional equivalence between benefits and contributions is implemented through an equity-weighted allocation algorithm that combines reward for contribution with guaranteed baseline access. Collective choice arrangements enable affected stakeholders to participate in rule modification through a Steering Committee with structured representation. Monitoring is supported via mechanism-level benchmarks and an independent Ombuds Office. Graduated sanctions ensure proportionate responses, and low-cost dispute-resolution mechanisms provide accessible conflict settlement. Recognition by external authorities legitimizes self-governance through initial United Nations hosting, with a pathway toward independent international organization status. Each Ostrom principal maps to a specific technical or institutional mechanism in MACPS.

The federated architecture operationalizes these principles while addressing digital sovereignty concerns that have impeded prior international data governance efforts [24]. Participants contribute to and benefit from shared infrastructure without ceding control of their compute or data to a single central

authority. Each node retains local autonomy while accessing collective capabilities that exceed any individual institution's resources. Privacy-preserving computation enables this architecture. Differential privacy provides mathematically grounded guarantees [27] limiting the information inferable about individuals from aggregate outputs, and secure multiparty computation enables joint analysis across distributed datasets without requiring any party to disclose raw data [28].

3.3. Huayan philosophy and mutual constitution

Conceptual precision matters, which is why we draw on Huayan Buddhist philosophy. Western systems thinking typically represents interdependence as a causal chain, where A influences B and B influences C. This captures relevant dynamics but not the form of cooperation MACPS is designed to realize. Fazing's Huayan formulation makes a stronger claim: mutual constitution without reduction [11, 33]. The central metaphor, Indra's Net, is an infinite cosmic network in which each intersection contains a jewel reflecting every other jewel; each jewel preserves its distinct identity while being constituted through these reflections.

Three Huayan principles shape MACPS design. Perfect interfusion (the mutual reflection across jewels) motivates modular layer interdependence: each layer gains capability through its relation to the others while preserving its own function. Mutual non-obstruction preserves distinct national regulatory frameworks while enabling interoperability: the semantic ontology lets regulatory frameworks from, for example, France, Kenya, and Japan reflect one another without requiring homogenization. Mutual constitution captures how individual pilot implementations instantiate universal principles while those principles are realized through concrete deployments.

This framework yields design requirements that a purely systems-theoretic approach does not. A systems-theoretic optimization would concentrate compute where marginal productivity is highest. Mutual constitution instead implies that each participant's improvement contributes to the whole; under-resourced participants therefore require enhancement not as redistribution but as structural necessity. Excluding Global South institutions would render the system incomplete (each jewel would reflect only partial representations), so the design mandates an equity-weighted allocation mechanism that prioritizes under-resourced institutions. Similarly, the principle of non-obstruction motivates semantic mapping rather than harmonization. Huayan serves here as a rigorous conceptual vocabulary rather than a metaphysical claim; the mapping between philosophy and implementation is necessarily approximate but sufficiently precise to generate concrete design constraints.

3.4. Infrastructure as political: Values, power, and intentional design

Infrastructure carries values. Winner [12] shows that Robert Moses's low bridges were designed to prevent buses from reaching Long Island beaches, limiting access for poor and Black New Yorkers. Star [13] shows that infrastructure renders some people visible while keeping others out of view: training language models on English texts, for example, embeds linguistic hierarchies directly relevant to MACPS's multilingual mediation pilot. Crawford [14] and Benjamin [25] demonstrate that AI systems encode existing inequalities into technological design. Benjamin's critique of the New Jim Code raises a hard question: if infrastructure always carries power, new infrastructure may simply introduce a different form of control.

Recognizing that infrastructure carries values means governance must be designed with intention. The important questions are which values the technology embeds, how, and who bears responsibility. Relying on moral or legal rules alone does not remove embedded values; it preserves those quietly built into systems controlled by powerful actors. Most AI governance today operates through private companies whose systems embed goals of profit maximization, user engagement, and surveillance-based value extraction [26]. The real question concerns corporate versus public control.

MACPS responds through constitutional infrastructure: technical systems that build fairness in from the start. Resource distribution is built into the system so that equitable allocation occurs automatically. Representation is required by design, not left as an aspiration. Voluntary equity commitments rarely scale, which helps explain the persistent gap between agreed principles and actual outcomes.

Embedding equity into infrastructure changes the political and economic dynamics: dropping equity obligations means leaving the system and losing access to shared resources.

The MACPS model implements this constitutional logic across five layers, each constraining a different form of power imbalance. Layer 1, the Semantic Ontology Layer, limits regulatory capture and jurisdiction shopping by rendering rules in machine-readable form for cross-jurisdictional comparison. Layer 2, the Compute Commons, limits resource concentration through equity-based sharing. Layer 3, Data Trusts, prevents extractive data collection. Layer 4, Evaluation Benchmarks, prevents ethics washing by replacing self-reported compliance with computational audits. Layer 5, Implementation Playbooks, converts governance rules into executable instructions, preventing informational asymmetry. Embedded values remain explicit, debatable, and accountable; a Steering Committee, Ombuds Office oversight, and transparent resource allocation sustain ongoing deliberation about which values the system should embody.

3.5. Commitment, defection, and the game theory of infrastructure-mediated cooperation

Rational actors need reasons to join, remain, and refrain from defection, even when short-term defection looks attractive. International relations theory identifies three principal defection modes: resource hoarding (a state keeps resources rather than sharing), compliance evasion (a state accepts rules but avoids accountability), and institutional capture (powerful actors remain inside and gradually reshape rules to serve their interests). These categories are not exhaustive but represent the principal strategic threats to multilateral cooperation identified in [20]. MACPS addresses each through structural design rather than persuasion.

Resource hoarding is deterred by a coordination dynamic: once enough mid-level institutions adopt MACPS, the cost of remaining outside rises. Institutions without the shared semantic ontology face escalating compliance burdens across jurisdictions, and those without access to the compute commons fall further behind in capability. The dynamic mirrors TCP/IP adoption: once the network reached sufficient scale, opting out meant forfeiting shared benefits.

Compliance evasion is deterred structurally. The audit layer (Section 4.4) replaces self-reports with automated structural verification. Mechanism metrics cannot be certified by assertion alone; organizations cannot satisfy Structural-Audit requirements without verified checks. This shifts compliance from signaling to substantive verification, making genuine compliance the rational choice for institutions seeking MACPS certification benefits. Explanation fidelity, representation stability, and value consistency metrics jointly prevent post-training manipulation.

Institutional capture is deterred by a constitutional design (Section 3.4) that distributes authority across overlapping checks. No single component can be co-opted sufficiently to compromise the whole system. Supermajority voting, mandatory rotation, civil society seats, an independent Ombuds, and open-source licensing create robust checks and balances. Successful capture would require coordination across several independent decision-making bodies, while resistance requires only adherence to existing rules: an asymmetry that favors institutional integrity.

The remaining commitment challenge concerns major powers (the United States, China, and to a lesser extent the European Union) whose participation significantly affects MACPS scale and effectiveness (addressed in detail in Section 8). Major powers face a hegemon's dilemma in infrastructure governance: unilateral provision is costly and may generate geopolitical resistance, while non-participation means losing influence over emerging standards. Historical precedents, including the United States' eventual integration into the ITU framework despite initial resistance and China's engagement with WTO dispute mechanisms despite early scepticism, show that effective infrastructure shifts major-power strategic calculus over time.

MACPS builds functional infrastructure first and does not presuppose immediate universal participation. Without US or Chinese involvement, MACPS would operate as a mid-power coalition with reduced but viable scale. Simulations indicate that participation by the EU, Japan, Australia, Brazil, South Africa, and Indonesia would provide sufficient compute pooling and regulatory coverage to sustain all five layers at lower aggregate capacity. Section 8 addresses this scenario in detail

4. System architecture

Building on the theoretical foundations above, we now specify the MACPS architecture. The system is organized into five layers, each implemented as a service exposing well-defined APIs. This modular structure reflects an engineering choice and the design principle of mutual non-obstruction: each layer preserves functional independence while gaining capability through integration. In operation, the layers interact dynamically: evaluation benchmarks (Layer 4) draw on the semantic ontology (Layer 1) to map results across jurisdictions, and data trusts (Layer 3) feed into compute allocation decisions (Layer 2) based on demonstrated need. Organizations can adopt the system in phases by implementing individual layers incrementally, with benefits expanding as participation broadens across the architecture. Figure 1 illustrates the complete system design.

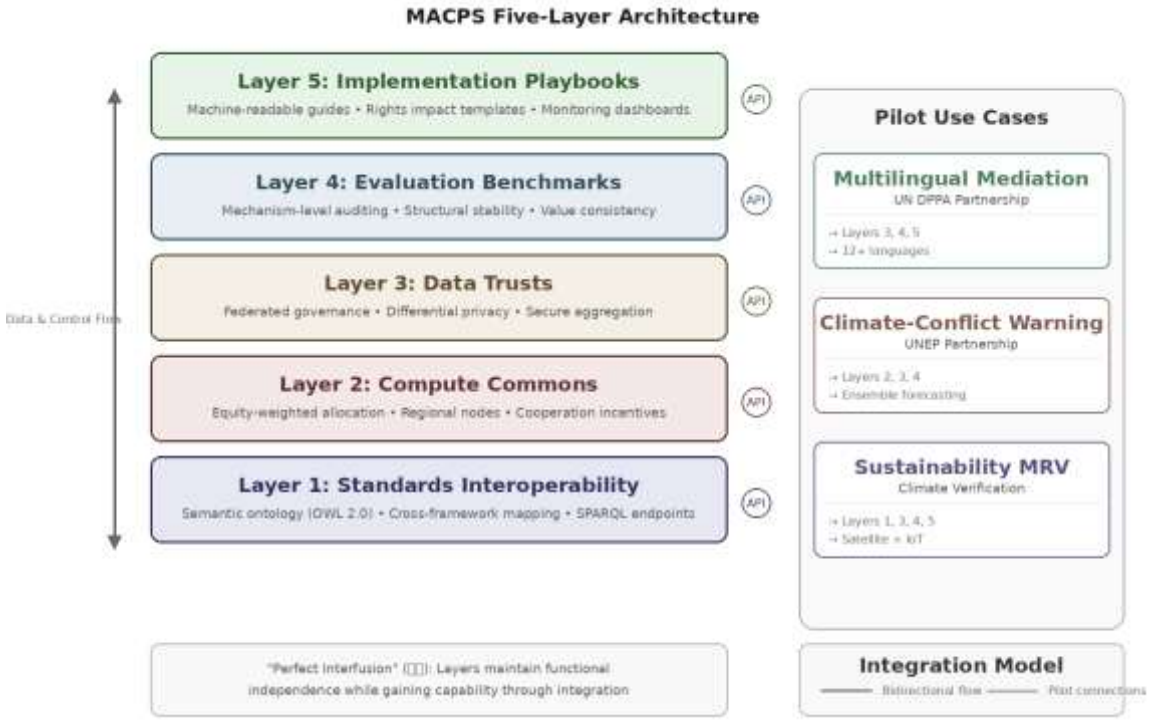

Figure 1. The five-layer architecture of MACPS highlights modular interdependence. Each layer operates as a service exposing well-defined APIs, while bidirectional arrows represent both data exchange and control flow between layers. Three pilot implementations utilize multiple layers, enabling “perfect interfusion” through an integration approach that strengthens rather than limits participant capabilities.

4.1. Standards interoperability layer

The foundational layer may suggest that regulatory fragmentation could be mitigated through semantic mapping rather than political harmonization. Moreover, a formal ontology is constructed using the Web Ontology Language (OWL 2.0), a structured language for representing knowledge systems that supports automated reasoning capabilities [34]. Furthermore, the significant evidence could indicate that this ontology encodes governance requirements across multiple jurisdictions, including control classes representing abstract obligations, such as "human oversight," framework instances corresponding to specific regulatory instruments (for example, UNESCO principles, EU AI Act provisions, and G7 practices), and equivalence relations implemented through owl: equivalent Class assertions that establish cross-framework mappings. In light of these key results, SPARQL Protocol and RDF Query Language (SPARQL) query endpoints might demonstrate that automated compliance verification could be supported through standardized query mechanisms.

Mapping process involves three sequential stages. However, the findings may suggest that expert annotation is first used to identify discrete regulatory requirements within source governance

frameworks. Additionally, embedding-based hierarchical clustering could demonstrate that semantically similar requirements appear to group effectively. Given that the evidence indicates formal verification is subsequently conducted using OWL reasoners, including HermiT and Pellet, the results might suggest that logical consistency within the ontology could be ensured through these important methods. Notwithstanding these results, initial implementation may suggest that coverage of approximately 80% of control elements across five major governance frameworks appears achievable. Nevertheless, the significant findings could indicate that this approach might demonstrate compliance across multiple jurisdictions through a single-point verification process, reducing redundant assessment burdens that could appear to disproportionately affect smaller organizations. Moreover, the evidence may suggest that institutions in the Global South might indicate insufficient resources to conduct separate compliance procedures across regulatory regimes [9].

4.2. Compute commons

The compute commons directly mitigate capability asymmetries through a federated pooling of computational resources. The technical specification includes distributed computing clusters deployed across at least three regional nodes (Africa, Asia-Pacific, and Latin America), with inter-node latency maintained below 100 ms. An initial allocation of 10,000 GPU-hours per month is provided in Year 1, increasing to 100,000 GPU-hours per month by Year 3. The total funding envelope is USD 50 million over three years, calculated using a subsidized rate of USD 0.15 per GPU-hour compared with prevailing commercial pricing of approximately USD 2.50 per GPU-hour.

The allocation mechanism operationalizes equity through the function: $Score(p) = 0.3 \cdot Impact(p) + 0.3 \cdot \text{Equity(p)} \cdot 1.5^{GS} + 0.2 \cdot Open(p) + 0.2 \cdot Coop(p)$, where $GS = 1$ for Global South institutions, applying a 1.5× equity multiplier. *Impact(p)* represents the projected humanitarian or societal benefit score (0–1), evaluated by a scientific advisory panel using a standardized rubric. *Equity(p)* quantifies the extent to which a proposed project benefits underserved populations (0–1). *Open(p)* captures the openness score, reflecting whether outputs are publicly licensed and reproducible (0–1). *Coop(p)* represents a cooperation bonus (0–1) assigned to multi-institutional or cross-regional projects.

This allocation design embeds distributive equity directly into the computational mechanism rather than relying on discretionary redistribution, which historical evidence suggests rarely scales reliably. The cooperation bonus further incentivizes cross-institutional collaboration, strengthening interdependence and increasing the relative cost of defection from the shared system. The 1.5 multiplier was calibrated to direct approximately 40% of total compute resources to Global South institutions, reflecting the Steering Committee's target for meaningful redistribution without rendering Global North participation unattractive.

The selected parameter weights (0.3, 0.3, 0.2, 0.2) and the 1.5× Global South multiplier are initial calibrations aligned with the Steering Committee's dual objectives of humanitarian impact and distributive fairness. Sensitivity analysis indicates that allocation outcomes remain stable under moderate parameter variation. Adjusting individual weights by ±0.1 while preserving normalization, and varying the equity multiplier between 1.3 and 2.0, produces rank-order changes in fewer than 15% of simulated proposals, primarily concentrated among borderline cases. The Steering Committee may revise these parameters through the Tier 2 decision procedure (Section 7), and annual recalibration based on observed outcomes is mandated by the governance framework.

To ensure rigorous and reproducible identification of “ underserved institutions, ” a composite Institutional Capacity Index (*ICI*) is defined. The index is given by:

$$ICI(i) = 0.35 \times GDP\_norm(i) + 0.25 \times Compute\_norm(i) + 0.20 \times HDI\_norm(i) + 0.20 \times Pub\_norm(i). \quad (1)$$

where *GDP_norm* denotes inverse-normalized GDP per capita derived from World Bank data; the inverse-normalized national AI compute capacity measured as publicly available GPU or TPU clusters per capita is represented by *Compute_norm*; *HDI_norm* corresponds to inverse-normalized human

development index values; and *Pub_norm* captures inverse-normalized AI publication output per capita based on bibliometric records from Scopus.

Institutions with $ICI \geq 0.60$ are classified as “ underserved” and receive the Global South multiplier ($GS = 1$) within the allocation mechanism. This threshold was determined through iterative calibration against World Bank and UNDP country classifications. Alternative thresholds of 0.50 and 0.70 were evaluated; 0.50 was too inclusive, capturing institutions with substantial AI capacity, while 0.70 was too restrictive, excluding many institutions in Latin America and Southeast Asia. The threshold is calibrated such that approximately 65%–70% of institutions in Africa, Southeast Asia, Latin America, and Small Island Developing States meet this criterion, while most institutions in high-income organizations for economic co-operation and development (OECD) countries do not. Limited exceptions may be granted for institutions located in economically disadvantaged regions within high-income countries, subject to case-by-case review by the Steering Committee. The *ICI* is recalculated annually using publicly available datasets, ensuring transparency, reproducibility, and external verifiability of the classification process.

4.3. Data trusts

Three specialized data trusts share control through privacy-safe federated computing. Moreover, the significant findings may suggest that conflict and humanitarian trust record checks using k-anonymity (with k greater than 10) and ε-differential privacy (ε at most 1.0) demonstrate that every event or need entry appears protected from individual identification. Furthermore, the evidence could indicate that setting ε at 1.0 or less may support that privacy remains intact while the data stays useful, matching approaches commonly found in census releases and health research.

In light of these results, the Climate/Environmental Trust might demonstrate that combining satellite images, IoT sensor data, and local community reports through federated learning could establish that spatial coarsening keeps map details from dropping below 1 km. Peace-process messages show nodes share data without saving raw files, deleting everything after 30 days. However, the key findings may suggest that the approach assumes the data keeper follows the rules but could indicate that attackers might attempt to link data back to real people. Additionally, the results could demonstrate that privacy spending accumulates over time, and any query exceeding the budget appears to require clear approval from the relevant stakeholders.

Given that secure aggregation prevents server-side inspection of gradients, federated learning might indicate that teams could train models on private data without sharing raw files centrally.

Federated governance could demonstrate that a significant privacy-versus-transparency trade-off exists, and the evidence may suggest that this appears to represent the main limit for whether the data trust approach remains viable. Thus, the findings might indicate that three key tensions could demonstrate important constraints.

In light of the evidence, differential privacy versus auditability may suggest that when privacy is strong at low ε, the data could appear less precise, which might make it harder to verify whether equity promises are being met in compliance reports. Furthermore, the results may indicate that federated learning and explainability present a significant tension, as gradient aggregation could demonstrate that server-side inspection appears constrained, which also limits how well the model can be explained afterward. Notwithstanding these results, data sovereignty and cross-jurisdiction analysis may suggest that rules like k-anonymity could indicate that personal records appear protected, but the evidence might demonstrate that spotting cross-border patterns requiring fine detail becomes harder.

However, the significant findings may suggest that MACPS could address these conflicts through layered protocol design, and the evidence appears to indicate that every trust keeps two separate budgets to balance privacy and accountability. Moreover, the key results might demonstrate that the public analytics budget stays at $\varepsilon\_pub \leq 0.5$ each quarter, while the evidence could suggest that the operational budget stays at $\varepsilon\_op \leq 0.5$ for routine queries, keeping the yearly total $\varepsilon\_total \leq 1.0$. Given that these budgets remain separate, the significant findings may indicate that transparency reports could appear privacy-safe regardless of how systems operate.

Furthermore, the results could demonstrate that the MACPS evaluation layer uses secure two-party computation to compare behavior summaries across nodes, which might suggest that auditing behavior appears possible while training remains private.

4.4. Evaluation benchmarks

Most current AI tests focus on how inputs turn into outputs, useful, but not enough for governance needs. Even if two models behave the same, they can work very differently inside, differences that matter for safety, making them more reliable, and keeping them aligned. MACPS audits at the mechanism level, checking the actual computational design instead of depending only on what outputs you can see. The Structural-Audit benchmark checks two key things: whether the model's inner activations stay the same for meaning-matched inputs, and whether its answers match its stated goals across different situations, so you can spot when it's gaming the specification. We set the target limits with caution, stability must be at least 0.85 and consistency at least 0.90, and we'll adjust them after pilot deployments. You'll find the complete benchmark details in Table 1 after the references section.

Table 1. Evaluation benchmarks and target thresholds.

Benchmark	Metrics	Target
Mediation-Natural Language Understanding (NLU)	Argument extraction *F1*, cross-lingual gap	*F1* ≥ 0.75, gap ≤ 0.10
EarlyWarn-CC	Brier skill score, lead time, false alarm rate	*BSS* ≥ 0.25, lead time ≥21 days, *FAR* ≤ 0.30
MRV-LandUse	*IoU* classification, change detection recall	*IoU* ≥ 0.80, recall ≥ 0.90
Structural-Audit	Representation stability, value consistency	Stability ≥ 0.85, consistency ≥ 0.90

Note: F1 denotes the harmonic mean of precision and recall; *BSS* refers to the Brier skill score; *FAR* denotes false alarm rate; *IoU* represents Intersection over Union.

To clarify the distinction between mechanism-level auditing and conventional compliance checking, two illustrative cases may suggest that the differences are significant. A standard fairness compliance test could indicate whether a hiring system satisfies the four-fifths rule by verifying that the selection rate ratio across demographic groups meets an adverse impact threshold of ≥ 0.80. Moreover, the significant evidence may suggest that such input-output evaluation can still be satisfied by a model that encodes discriminatory internal representations while concealing them through post hoc calibration, enabling forms of specification gaming. Given that the findings demonstrate that concealment remains possible, MACPS mechanism-level auditing might indicate that examining the internal model structure directly could provide more reliable results. In contrast, MACPS mechanism-level auditing appears to evaluate whether protected-attribute information is linearly decodable from intermediate representations. Furthermore, the significant representation stability metric (≥ 0.85) may suggest that internal activations for semantically equivalent inputs differing only in protected attributes maintain a cosine similarity above 0.85, constraining internal representational divergence rather than only observable outputs. Therefore, the key findings could indicate that this approach might demonstrate important advantages over conventional methods.

In light of these results, a parallel distinction appears to apply to transparency evaluation. Notwithstanding the widespread use of documentation artifacts, the significant evidence may suggest that conventional compliance approaches that treat model cards as sufficient evidence of transparency could demonstrate important limitations. Thus, the key findings might indicate that the MACPS mechanism-level audit introduces a verifiable EFS, defined as the correlation between feature attributions generated by the system's explanation module and feature importance derived from causal intervention testing. However, the important results could suggest that an EFS threshold of ≥ 0.80 may demonstrate that explanations correspond to actual causal model behavior rather than post hoc rationalizations that are only superficially plausible. Additionally, the significant evidence might indicate that Figure 2 demonstrates the distinction between standard compliance checks and the MACPS mechanism-level audit for both fairness and transparency evaluation. Moreover, the key

results may suggest that these mechanism-level metrics could demonstrate important properties when computed automatically by the MACPS evaluation infrastructure and exposed through standardized APIs. In light of the evidence, the significant findings might indicate that continuous monitoring of model properties appears to provide important advantages rather than reliance on periodic, snapshot-based compliance assessments.

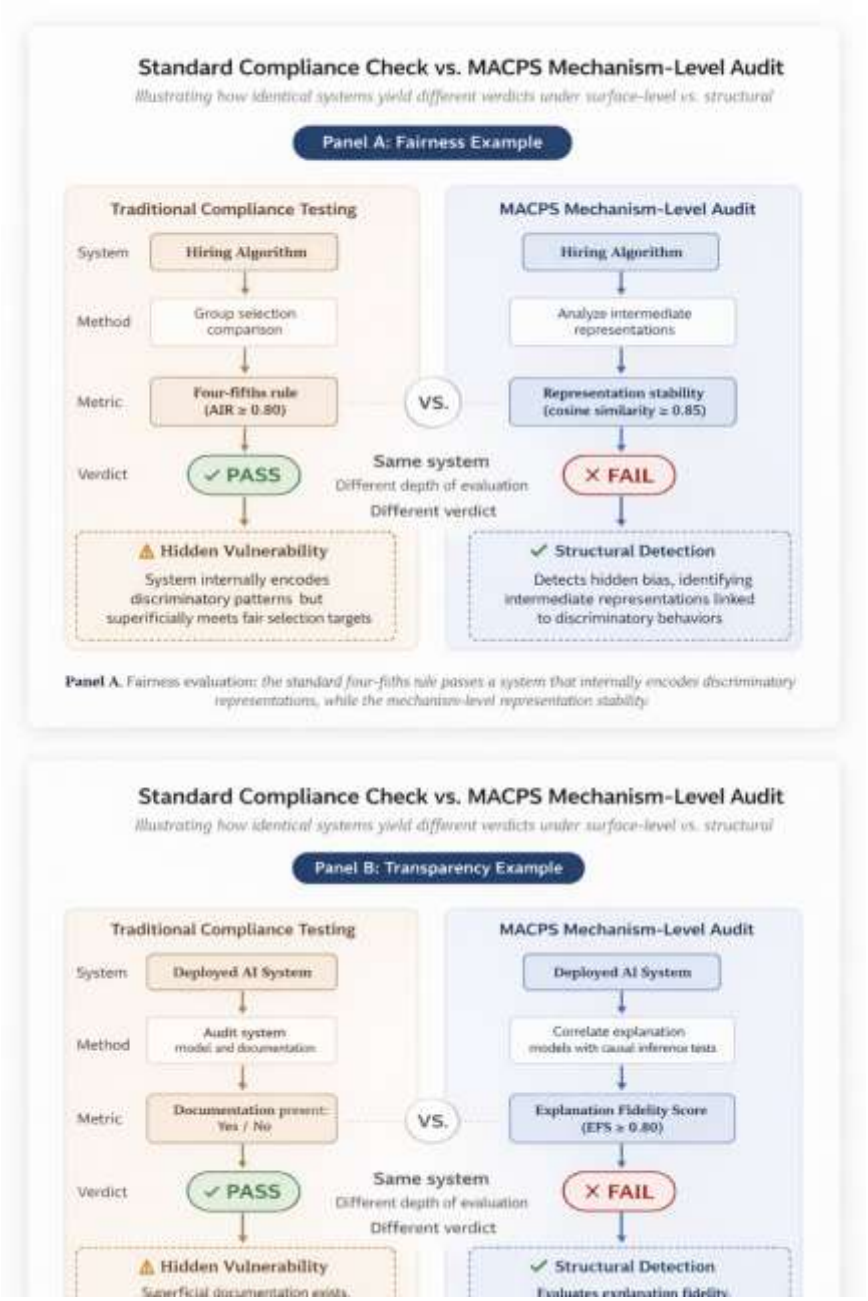

Figure 2. Comparison between a standard compliance check and the MACPS mechanism-level audit. Panel A (fairness): the conventional four-fifths rule approves a system that internally encodes discriminatory representations, whereas the mechanism-level representation stability metric (cosine similarity ≥ 0.85) identifies concealed bias within intermediate representations. Panel B (transparency): routine documentation verification approves systems possessing superficial documentation, whereas the MACPS EFS ≥ 0.80 assesses whether explanations reflect authentic model behavior by means of causal inference testing.

5. Comparative analysis

To clarify the distinctive contribution of MACPS, Table 2 situates it in relation to existing governance initiatives. The Global Partnership on AI (GPAI), the OECD.AI policy observatory, and the Partnership on AI (PAI) each perform important coordination roles, including stakeholder

convening, development of best-practice guidelines, and monitoring of policy developments [35, 36]. However, these initiatives do not provide an integrated technical infrastructure. This distinction is foundational: whereas existing frameworks primarily support convergence on what actions should be taken, MACPS supplies the operational capability required to implement those actions.

Table 2. Comparison of MACPS with alternative AI governance frameworks.

Dimension	MACPS	GPAI	OECD.AI	PAI
Compute access	Pooled (equity-based)	None	None	None
Data infrastructure	Federated trusts	None	None	Limited
Standards mapping	Semantic ontology	Working groups	Policy tracker	Best practices
Equity mechanism	Algorithmic (≥40%)	Aspirational	Limited	Aspirational
Accountability	Structural audit	Peer review	Metrics	Guidelines

Note: GPAI denotes Global Partnership on AI; PAI refers to Partnership on AI. The term "aspirational" indicates declared commitments that lack enforceable mechanisms.

The need for new infrastructure could indicate that three distinct structural differences exist between what current programs provide and what the capability gap actually requires. However, the significant findings may suggest that existing programs focus primarily on discussion and coordination rather than resource delivery, and attempting to add capability functions to consensus-driven groups could demonstrate serious governance conflicts. Moreover, the evidence may indicate that closing the capability gap requires real resources, computing power, well-made datasets, and technical know-how that current programs appear insufficiently equipped to provide. In light of these findings, the results might suggest that while GPAI establishes expert groups, it appears unable to distribute GPU resources in any meaningful way. Closing the credibility gap shows audits matter.

Skeptics might ask why GPAI, OECD.AI, and PAI have not built one shared technical system, even though they are widely seen as valuable. Furthermore, the significant evidence may suggest that this problem requires a clear structural explanation rather than vague institutional delay. Thus, the findings could indicate that three key factors explain the gap. Given that the original mission constrains institutional capacity, results might demonstrate that GPAI was established through a G7 statement to support responsible AI guidance rather than resource pooling. Mission limits infrastructure scope.

Additionally, the significant evidence may suggest that member countries retain their own AI plans, and the group's charter could indicate that it does not permit rules forcing specific technical standards or computing assignments. Notwithstanding these results, the findings might demonstrate that changing this arrangement requires states to renegotiate because their push for control over AI affects whether they can agree on shared resources. Therefore, the evidence may indicate that this reflects an expected match between institutional design and functional capacity rather than breakdown. In light of these key results, the study could suggest that political and economic factors limit what infrastructure investment can achieve across these governance structures. Politics constrains shared infrastructure.

Moreover, the significant findings may indicate that building shared technical systems, including computing clusters, federated data trusts, and interoperable systems, requires steady funding that voluntary contributions appear unable to reliably provide. Furthermore, the evidence could suggest that OECD.AI directs most resources toward research and policy work, while large-scale infrastructure requires long-term funding commitments similar to CERN's shared-cost approach. However, the results might demonstrate that this kind of coordinated fiscal planning does not yet appear established across AI governance groups. Notwithstanding these findings, the study may indicate that these institutions frequently converge on the lowest shared ground when major AI players collaborate with smaller countries. Consensus blocks strict requirements.

Given that significant evidence demonstrates that plans forcing strict audits or mandatory redistribution often face opposition from disadvantaged parties, the results could suggest that MACPS addresses this problem by beginning as a small group and building needed systems before expanding. Additionally, the findings may indicate that MACPS incorporates fairness into the design from the start rather than through subsequent negotiation, which could demonstrate a structurally distinct approach to governance. Thus, the significant evidence might suggest that this approach does not imply

that GPAI and OECD.AI failed to achieve their goals, since their institutional setup could indicate that infrastructure development falls outside their intended scope. Infrastructure work shows different governance logic.

6. Pilot use cases

Three pilot implementations may suggest that MACPS demonstrates significant capabilities across application domains characterized by high humanitarian relevance.

6.1. Multilingual mediation support

In collaboration with the UN Department of Political and Peacebuilding Affairs (DPPA), domain-adapted language models are being developed to support conflict mediation across 12 or more languages, including low-resource languages for which commercial translation systems are insufficient. The technical approach involves fine-tuning multilingual transformer architectures using federated learning, with four core functional capabilities:

i. Argument extraction: identification of claims together with their underlying reasoning structures.

ii. Summarization: preservation of semantic nuance across diverse cultural contexts.

iii. Integrative option identification: generation of potential creative resolution pathways.

iv. Escalatory rhetoric detection: identification of linguistic patterns associated with breakdowns in negotiation.

Safeguards are implemented to reflect the sensitivity of peace processes. Strict human-in-the-loop protocols ensure that the system provides advisory outputs only, while mediators retain full decision-making authority. Prohibition of long-term dialogue storage prevents the formation of exploitable communication archives. Mandatory review by cultural advisors is included to mitigate risks arising from Western-normed models misinterpreting culturally specific communication patterns.

i. A 25%–35% reduction in translation-related workload, based on DPPA field reports indicating that mediators currently spend approximately 30%–40% of session time on translation tasks during multilingual negotiations.
ii. A 40%–60% increase in the number of documented creative options explored, relative to baseline mediation logs reporting an average of 3.2 integrative proposals per negotiation round, according to DPPA internal analysis.

iii. An argument extraction performance of $F1 \geq 0.75$, with a cross-lingual transfer gap ≤ 0.10.

6.2. Climate-conflict early warning

A collaboration with the United Nations Environment Programme and humanitarian early-warning consortia supports integrated forecasting of climate-related conflict risk at district and provincial scales. Climate change functions as a “threat multiplier,” intensifying pressures related to resource competition, migration, and livelihoods; however, existing early-warning systems typically treat climate and conflict dynamics as separate analytical domains.

The technical framework relies on ensemble modelling that fuses multiple data sources, including downscaled Coupled Model Intercomparison Project Phase 6 (CMIP6) climate projections at 25 km resolution, socioeconomic indicators, conflict-event datasets derived from Armed Conflict Location and Event Data Project (ACLED) historical records, satellite imagery from Sentinel-2 capturing land-use change and vegetation stress, and structured community reports. Sentinel-2 and Sentinel-5P were selected for their open-access licensing, global coverage, and temporal resolution (5-day revisit for Sentinel-2, daily for Sentinel-5P), though cloud obstruction in tropical regions requires multi-temporal compositing.

To address dual-use concerns, several safeguards are implemented: analysis is restricted to regional-level outputs; system access is limited to authorized humanitarian organizations; use by security

services is explicitly prohibited; and quarterly algorithmic bias audits are mandated alongside remediation procedures.

Ethical risks associated with conflict prediction systems require sustained attention. Cederman and Weidmann [37] show that models trained on historical conflict data may inherit biases embedded in colonial-era borders and Cold War alliance structures, which can misdirect humanitarian resources away from emerging conflict patterns that diverge from historical precedents. MACPS addresses these risks through mandatory quarterly bias audits, spatial disaggregation designed to identify regional disparities in predictive performance, and community feedback mechanisms that enable affected populations to contest model outputs. Despite these safeguards, no technical mechanism fully resolves the structural tension between predictive accuracy and the risk that forecasts may become self-fulfilling through their influence on resource allocation and political decision-making.

Target outcomes are defined as the detection of 60%–80% of climate-related conflict events with a lead time of ≥ 21 days, compared to baseline systems that detect approximately 40% of events with a 14-day lead time according to United Nations Office for the Coordination of Humanitarian Affairs (OCHA) evaluations. Additional performance targets include a Brier skill score of ≥ 0.25 and a false-alarm rate of ≤ 0.30.

6.3. Sustainability verification

Even with AI-supported MRV, tracking and proof still fall short, which can weaken the strength of international climate deals. Even when countries make bold promises in their national climate plans, checking those promises is still expensive, not fully done, and often becomes political.

The technical framework consists of three integrated components: land-use classification via semantic segmentation applied to Sentinel-2 imagery with 10 m spatial resolution; methane detection using spectral analysis of Sentinel-5P observations; and supply-chain traceability supported by blockchain-anchored Internet of Things sensors. Governance safeguards are incorporated to reduce the risk of regulatory capture by evaluated parties. Institutional independence is maintained by separating technical assessment functions from political decision-making processes. An open methodological framework is included to enable independent external validation of results. Target outcomes include a 70%–85% reduction in verification costs compared with manual approaches, based on baseline MRV costs of USD 15–25 per hectare for ground-based verification reported in technical documentation of the United Nations Framework Convention on Climate Change. These estimates are derived from pilot simulations comparing automated satellite-based classification costs (USD 0.50 to 2.00 per hectare) against reported manual verification costs. Additional targets include a tenfold increase in geographic monitoring coverage, land-use classification performance with Intersection over Union (IoU) ≥ 0.80, and deforestation detection recall ≥ 0.90.

7. Governance, evaluation, and institutionalization

MACPS governance has to do two things well: make clear, consistent decisions for strong infrastructure, and include many voices for trust. We can't eliminate this tension entirely; we can only handle it through institutions [38].

The steering committee includes 15 members, each serving staggered three-year terms. Five of them belong to the UN system: UNU, UNESCO, DPPA, UNEP, and ITU. Five regional bodies take turns: the African Union, ASEAN, ECOWAS, the EU, and the OAS. Three members speak for civil society, while two represent the technical community. Decisions need a two-thirds vote. When minorities are blocked, it's only for major changes to equity commitments.

Voting comes in three levels. Tier 1 is for everyday work, needs more than half the votes (8 of 15), and allows 14 days for consultation. Tier 2 deals with real policy changes and needs a two-thirds vote (10 of 15), plus 30 days of review and public input. Tier 3 handles constitutional amendments; they need a four-fifths supermajority (12 of 15) plus a 90-day review and an independent impact check. Emergency rules let leaders decide by simple majority, pending later approval. For Tier 1, quorum is

11 out of 15; for Tiers 2 and 3, it's 13 out of 15. You can't use proxy votes, but pre-chosen alternates may vote.

Disputes get settled in three steps. For the first 30 days, the Ombuds Office helps run the discussions. Stage 2 lasts 60 days and sends unresolved disputes to a randomly chosen three-person panel that makes final factual and procedural decisions. After 90 days in Stage 3, value disagreements reach the Steering Committee, following Tier 2 voting rules. Every Stage 2 and Stage 3 decision is posted in full, dissent included. After finishing the set steps, members can leave by giving 180 days' notice, while still retaining the data they created.

Expect three types of value conflicts. We handle the conflict between privacy and transparency by using a layered privacy budget system, turning disagreements into clear technical limits. When safety access permissions clash, we handle it step by step: Tier 1 lets people enter with basic certification and fewer checks, while Tier 3 grants full credentials for critical situations. When countries disagree, they can join a shared data trust only if they choose. They earn benefits by contributing, while they keep full control over any data they don't share.

The federated setup avoids single points of failure and helps block outside pressure. No one node runs the core system. The semantic map is copied in different regions, computing is shared through agreed-upon consensus, and data is held by independent custodians using encryption-based access rules. Byzantine fault tolerance means any group smaller than one-third can't change the rules on its own. Every part runs under the Apache 2.0 open-source license, which means you can fork the infrastructure if needed. Together, these features make sure one actor can't decide another participant's fate.

A 20-person scientific panel offers technical advice, but it's not binding. The Ombuds Office looks into complaints and shares yearly transparency reports. MACPS starts by running under UNU, but it can become an independent international organization if the pilot works.

MACPS is distinguished from governance initiatives that articulate objectives without verification capacity. Pre-specified success criteria include achievement of ≥70% of technical performance targets, following established benchmarking practice in international institutional design [39], ≥80% stakeholder satisfaction, ≥40% Global South compute allocation verified through system logs, and zero critical safety incidents.

Four hypotheses are defined for empirical evaluation: Each hypothesis targets a distinct MACPS objective: cooperation enhancement (H1), equity in capability provision (H2), credibility of audit mechanisms (H3), and real-world humanitarian impact (H4).

i. H1: MACPS users exhibit higher cross-border collaboration, tested via quasi-experimental matched comparison and two-sample t-test ($n \geq 30$ per group, $\alpha = 0.05$).
ii. H2: Output quality increases disproportionately for Global South institutions relative to Global North institutions, tested using mixed-effects regression ($n \geq 50$ institutions).
iii. H3: Structural audit failures predict subsequent deployment issues, tested using chi-square analysis ($n \geq 20$ systems per group).
iv. H4: Early warning regions show reduced conflict escalation relative to synthetic control counterfactuals, tested using permutation inference ($n \geq 5$ treated regions).

Sample sizes are derived from a priori power analyses assuming medium effect sizes (Cohen’ s d = 0.50 for H1; $f^2 = 0.15$ for H2) at 80% statistical power, yielding minimum requirements of $n = 26$ per group for H1 and $n = 43$ institutions for H2. The chosen sample sizes of 30 and 50 provide conservative </w:t> <w: lastRenderedPageBreak/><w:t xml: space="preserve">margins. For H3, $n = 20$ per group ensures 80% power to detect an odds ratio of 3.0. For H4, synthetic control methods remain robust with at least five treated units, provided a donor pool of ≥20 untreated regions are available; MACPS regional coverage across Africa and Asia-Pacific satisfies this condition.

For the H1 test, we use propensity score matching based on institutional traits, budget, staffing, past international collaboration, academic field, and national income. We set the caliper to 0.2 standard deviations of the logit score to keep the two groups well balanced. For H2, the mixed-effects regression lets each institution and country have its own starting point (random intercepts). It then tests MACPS

participation, baseline output quality, and time as fixed effects, so we can compare within the same institution while accounting for prior differences. If fewer than half the targets are reached after 18 months, the plan calls for an independent review to share what happened and help redesign the program.

8. Limitations and risks

The study identifies that risks fall into six groups, each with mitigation steps to reduce harm.

However, the findings may suggest that data weaponization poses significant challenges, as states could withhold, alter, or restrict data to avoid unfavorable comparisons. Furthermore, the evidence could indicate that independent audits, multiple data sources, anomaly detection, and clear removal rules might reduce this risk substantially. Nevertheless, the results appear to show that these measures cannot eliminate risk entirely, since some trust remains necessary.

Regulatory pace could demonstrate that rules may evolve faster than the ontology can adapt, and the interoperability layer targets 80% coverage. Moreover, the findings may suggest that modular design and versioned ontologies could allow incremental expansion across key domains. In light of these results, even 70%–80% interoperability appears to prove useful, and open development might welcome outside contributions.

Additionally, the evidence could indicate that elite capture allows well-resourced organizations to dominate standard-setting and benchmark design. Thus, the study may suggest that mitigation includes rotating governance roles, increased transparency, formal voting, and meaningful exit options.

Given that major-power non-participation could demonstrate significant challenges, the findings may suggest that the US and China might decline to join and discourage allies from participating. However, the results could indicate that MACPS operates through a coalition of willing partners, including mid-sized countries, international bodies, and civil society. Furthermore, the evidence may suggest that open-source design might prevent lock-in, while regional hubs in Africa, Asia-Pacific, and Latin America could reduce dependence on major-power jurisdictions. Notwithstanding these challenges, historical precedents such as the ITU, CERN, and arms-control agreements could demonstrate that operational success may shift incentives over time.

Technical vulnerabilities could indicate that privacy-preserving computing remains demanding and may not withstand future attacks, and a compromised custodian could expose private data. Moreover, the findings may suggest that mitigation includes collaborative cryptographic research, calibrated privacy budgets, and prepared breach-response plans.

Therefore, the evidence could indicate that temporal contingency demonstrates that MACPS gains value by addressing regulatory fragmentation. In light of the key results, rapid global convergence might reduce the value of semantic mapping, though shared compute and evaluation benchmarks could remain useful.

The study may suggest that philosophical limits remain important, as the Huayan framework serves as an explanatory vocabulary, not an ontological claim, and its link to technical design appears approximate. Given that the evidence could demonstrate that this framework generates design requirements such as equitable allocation and mutual constitution, the results may suggest that purely systems-theoretic approaches do not produce these requirements. Furthermore, the findings could indicate that this paper presents a design proposal, not a deployed system, and the evaluation plan depends on implementation.

9. Conclusion

This article demonstrates that effective global AI governance requires infrastructure, not just principles. Despite broad agreement on transparency, fairness, accountability, and safety, the findings could indicate that this consensus has not prevented ethics washing, computing power concentration, or harmful system deployment. Moreover, the evidence may suggest that the core problem appears to

be a lack of mechanisms to convert shared commitments into verifiable, operational steps. In light of these significant findings, MACPS could address this gap through five interconnected layers that might improve coordination, provide key capabilities, and ensure credibility.

However, the results may suggest that infrastructure-based cooperation could demonstrate important differences from traditional multilateralism, since conventional agreements depend on shared values or external threats. Given that the evidence could indicate that infrastructure-based cooperation appears self-enforcing, the findings may suggest that defection means immediate loss of shared resources, while participation might yield benefits unavailable to any actor alone. Furthermore, the results could demonstrate that Huayan's perfect interfusion captures this logic, as each participant may grow through the shared system while retaining its identity.

Additionally, the evidence could indicate that research confirms that infrastructure embeds values, and the findings may suggest that treating this as an opportunity for intentional design appears consequential. Therefore, the results could demonstrate that since technical systems inevitably carry values, the key questions might concern which values are embedded, how institutions embed them, and who is accountable. Moreover, the findings may suggest that constitutional infrastructure could make equity structural rather than optional, helping close the persistent gap between agreed principles and actual resource distribution.

In light of significant evidence, MACPS might not require universal participation, and a coalition of willing partners could establish an initial operational system. Furthermore, the results could indicate that network externalities might increase incentives to join, as exclusion from shared infrastructure may grow costlier over time.

Given that the findings could demonstrate that this paper contributes on two fronts, the evidence may suggest that demonstrating how institutional and technical integration creates cooperative gains appears critical. The study may suggest that the main contribution could demonstrate a five-layer architecture, a framework for infrastructure-mediated multilateralism, and governance mechanisms that might balance operational effectiveness with legitimacy.

Declarations

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The author declares no conflicts of interest.

Data Availability: Not applicable (design proposal). Pilot evaluations will generate publicly available datasets subject to privacy constraints described in Section 4.3.

Ethics Approval: Not applicable. Pilot evaluations will undergo independent ethics review prior to implementation as specified in Section 7.

References

[1] Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1:389–399. https://doi.org/10.1038/s42256-019-0088-2

[2] Epoch AI (2024) Training compute of frontier AI models grows by 4-5x per year. Epoch AI Research. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year. Accessed 15 Jan 2025

[3] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, pp 33–44. https://doi.org/10.1145/3351095.3372873

[4] McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: the AI Incident Database. In: Proceedings of the AAAI Conference on Artificial Intelligence 35(17):15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

[5] European Parliament (2024) Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Off J Eur Union L 186:1–144. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[6] UNESCO (2021) Recommendation on the ethics of artificial intelligence. UNESCO, Paris. https://unesdoc.unesco.org/ark:/48223/pf0000381137
[7] G7 (2023) Hiroshima Process international code of conduct for advanced AI systems. G7, Hiroshima. https://www.mofa.go.jp/ecm/ec/page5e_000076.html. Accessed 15 Jan 2025
[8] Png M (2022) At the tensions of South and North: critical roles of Global South stakeholders in AI governance. In: Proceedings of the ACM conference on fairness, accountability, and transparency. ACM, New York, pp 1434–1445. https://doi.org/10.1145/3531146.3533200
[9] Effoduh JO (2024) A Global South perspective on explainable AI. Carnegie Endowment for International Peace, Washington, DC. https://carnegieendowment.org/research/2024/04/a-global-south-perspective-on-explainable-ai. Accessed 15 Jan 2025
[10] Bietti E (2020) From ethics washing to ethics bashing: a view of tech ethics from within moral philosophy. In: Proceedings of the ACM conference on fairness, accountability, and transparency. ACM, New York, pp 210–219. https://doi.org/10.1145/3351095.3372860
[11] Cook FH (1977) Hua-yen Buddhism: the jewel net of Indra. Pennsylvania State University Press, University Park
[12] Winner L (1980) Do artifacts have politics? Daedalus 109(1):121–136. https://www.jstor.org/stable/20024652
[13] Star SL (1999) The ethnography of infrastructure. Am Behav Sci 43(3):377–391. https://doi.org/10.1177/00027649921955326
[14] Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven
[15] Ostrom E (1990) Governing the commons: the evolution of institutions for collective action. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511807763
[16] Cihon P, Maas MM, Kemp L (2020) Fragmentation and the future: investigating architectures for international AI governance. Glob Policy 11(5):545–556. https://doi.org/10.1111/1758-5899.12890
[17] Taeihagh A (2021) Governance of artificial intelligence. Policy Soc 40(2):137–157. https://doi.org/10.1080/14494035.2021.1928377
[18] Ulnicane I, Knight W, Leach T, Stahl BC, Wanjiku WG (2021) Framing governance for a contested emerging technology: insights from AI policy. Policy Soc 40(2):158–177. https://doi.org/10.1080/14494035.2020.1855800
[19] Maas MM, Villalobos JJ (2023) International AI institutions: a literature review. AI Foundations Report 1. https://www.governance.ai/research-paper/international-ai-institutions-a-literature-review. Accessed 15 Jan 2025
[20] Roberts H, Hine E, Taddeo M, Floridi L (2024) Global AI governance: barriers and pathways forward. Int Aff 100(3):1275–1286. https://doi.org/10.1093/ia/iiae073
[21] Ho L, Barnhart J, Trager R, Bengio Y, Brundage M, Carnegie A, Chowdhury R, Dafoe A, Hadfield GK, Levi M, Snidal D (2023) International institutions for advanced AI. arXiv:2307.04699. https://arxiv.org/abs/2307.04699
[22] Trager RF, Harack B, Reuel A, Carnegie A, Heim L, Ho L, Kreps S, Leng R, Schuett J, Simchowitz M, Tallberg J (2023) International governance of civilian AI: a jurisdictional certification approach. arXiv:2308.15514. https://arxiv.org/abs/2308.15514
[23] Benkler Y (2006) The wealth of networks: how social production transforms markets and freedom. Yale University Press, New Haven. https://www.benkler.org/Benkler_Wealth_Of_Networks.pdf
[24] Couldry N, Mejias UA (2019) The costs of connection: how data is colonizing human life and appropriating it for capitalism. Stanford University Press, Stanford
[25] Benjamin R (2019) Race after technology: abolitionist tools for the New Jim Code. Polity Press, Cambridge
[26] Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. PublicAffairs, New York
[27] Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407. https://doi.org/10.1561/0400000042
[28] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the ACM SIGSAC conference on computer and communications security. ACM, New York, pp 1175–1191. https://doi.org/10.1145/3133956.3133982
[29] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. arXiv:1606.06565. https://arxiv.org/abs/1606.06565
[30] Ngo R, Chan L, Mindermann S (2024) The alignment problem from a deep learning perspective. In: Proceedings of the International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=fh8EYKFKns
[31] Veale M, Zuiderveen Borgesius F (2021) Demystifying the Draft EU Artificial Intelligence Act. Comput Law Rev Int 22(4):97–112. https://doi.org/10.9785/cri-2021-220402

[32] Smuha NA, Yeung K (2025) The European Union's AI Act: beyond motherhood and apple pie? In: Smuha NA (ed) The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence. Cambridge University Press, Cambridge, pp 228–258. https://doi.org/10.1017/9781009367783.015
[33] Cleary T (1983) Entry into the inconceivable: an introduction to Hua-yen Buddhism. University of Hawaii Press, Honolulu
[34] W3C (2012) OWL 2 Web Ontology Language document overview. W3C Recommendation. https://www.w3.org/TR/owl2-overview/. Accessed 6 Jan 2026
[35] GPAI (2023) Global Partnership on AI: 2023 annual report. GPAI Secretariat, Paris. https://gpai.ai/. Accessed 15 Jan 2025
[36] OECD (2024) OECD.AI Policy Observatory. https://oecd.ai. Accessed 1 Jan 2026
[37] Cederman L-E, Weidmann NB (2017) Predicting armed conflict: time to adjust our expectations? Science 355(6324):474–476. https://doi.org/10.1126/science.aal4483
[38] Buchanan A, Keohane RO (2006) The legitimacy of global governance institutions. Ethics Int Aff 20(4):405–437. https://doi.org/10.1111/j.1747-7093.2006.00043.x
[39] Global Fund (2023) Operational policy note: grant implementation. The Global Fund, Geneva. https://www.theglobalfund.org/en/operational-policy-notes/. Accessed 15 Jan 2025