Technical Brief: The Invariance Framework for Verifiable AI Governance

To: High-Level Stakeholders (UN, OECD, G7 AI Safety Institutes)

From: The Office of Andrew H. Bond

Subject: Transitioning from Probabilistic Alignment to Deterministic Geometric Governance

1. Executive Summary

Current AI safety paradigms rely on "Alignment"—a probabilistic attempt to make model outputs conform to human values through fine-tuning. This approach is fundamentally fragile due to **Representation Dependence**: a model's judgment often changes under simple redescriptions of the same situation[111]. We propose a transition to **Invariant Governance**, a framework that mandates AI systems remain consistent across all meaning-preserving transformations. By adopting the **Epistemic Invariance Principle (EIP)** and its normative specialization, the **Bond Invariance Principle (BIP)**, regulators can move from "vibes-based" oversight to **mathematically verifiable audits**[2222].

2. The Problem: The Fragility of Current AI

Existing AI evaluation (benchmarking) fails to detect "brittle generalization." A system may appear safe in a test environment but fail in the real world because its internal logic tracks **syntax** (how a prompt is written) rather than **structure** (the underlying facts or moral bonds)[3].

- **Epistemic Failures:** Renaming variables or reordering premises can flip a model's logical or mathematical conclusion[4].
- **Normative Failures:** Describing an ethical dilemma in different terms—without changing the morally relevant relationships (Bonds)—can result in contradictory ethical verdicts[5].

3. The Solution: The Invariance Principles

The framework introduces two foundational principles that serve as a "Constitutional Logic" for autonomous systems:

A. The Epistemic Invariance Principle (EIP)

EIP requires that an AI's judgment be invariant under all transformations that preserve a domain's task-relevant structure[6].

- **Non-Degeneracy:** Ensuring the system remains sensitive to actual structural changes while ignoring superficial ones[7777].
- **Uncertainty Stability:** Mandating that the system explicitly abstain or escalate decisions when invariance cannot be certified[8888].

B. The Bond Invariance Principle (BIP)

A specialization of EIP for ethics, BIP dictates that a system's moral verdict must depend solely on the "Bonds" (morally relevant relationships) between stakeholders[9].

- **Stratified Geometric Ethics (SGE):** Modeling the "moral landscape" as a stratified space allows the system to represent hard boundaries (Vetoes) that cannot be crossed, regardless of the prompt's phrasing.
- **Auditability:** BIP ensures that unjustified prejudice is impossible to hide; if the bonds are identical, the verdict must be identical. Any deviation is a mathematically provable audit failure.

4. Operational Infrastructure for Policy Makers

To accelerate adoption, this framework provides a "Moral Compiler" and a verification blueprint:

- **Platonic Dialogue Interface:** A narrative-driven method to compile stakeholder values into machine-readable **DEME Profiles**.
- **Transformation-Based Test Suites:** Standardized evaluation protocols that compute "Invariance PASS rates" across mathematical, semantic, and normative domains[10101010].
- **Machine-Checkable Audit Artifacts:** Digital signatures and JSON-based "Epistemic Contracts" that bind every AI decision to its evidence provenance and transformation trials[11111111].

5. Policy Recommendation

We recommend that international governing bodies adopt **Invariance Certification** as a requirement for "Safety-Critical" AI deployments (e.g., Healthcare, Autonomous Transport, Infrastructure). By requiring models to prove **Representation Invariance**, we can ensure that AI agents behave as predictable, stable, and "objectively" consistent actors in human society[12121212].

Contact for Technical Consultation: *Andrew H. Bond San José State University andrew.bond@sjsu.edu*