THE UNIFIED ARCHITECTURE OF ETHICAL GEOMETRY

A Mathematical Framework for Representation-Invariant Moral Evaluation

Integrating Gauge Theory, ErisML Canonicalization, and Moral Current Dynamics

Version 3.0 — Mathematically Rigorous Edition

December 2025

Andrew H. Bond
Department of Computer Engineering
San José State University
andrew.bond@sjsu.edu

EPISTEMIC STATUS
This paper presents a theoretical framework with a concrete implementation pathway. Definitions are stipulative. Theorems are proven within the framework's axioms. Conjectures are explicitly marked. All mathematical objects are precisely defined. Physical analogies are structural, not ontological.

Abstract

We present a unified mathematical framework for ethical evaluation in artificial intelligence systems, constructed in six layers: (1) a tensor foundation defining intention, obligation, and judgment via inner products; (2) a symmetry principle requiring invariance under meaning-preserving redescription; (3) a gauge-theoretic structure formalizing this invariance; (4) a concrete canonicalizer implementation using ErisML as the target grammar; (5) a measurement theory defining curvature in Bonds; and (6) a dynamics of moral current resolving the Noether objection.

The key innovation is replacing fuzzy vector-space canonicalization with deterministic grammar parsing. The ErisML modeling language provides a discrete lattice of valid moral states; the canonicalizer $\kappa(x)$ becomes a parsing operation that either succeeds (producing a unique canonical form) or fails (triggering a veto). This eliminates the exploitable curvature inherent in continuous embedding spaces.

All mathematical claims are precisely stated with explicit hypotheses. We distinguish between theorems (proven), conjectures (proposed), and approximations (operational). The framework is constructive: it specifies how to build systems that resist specification gaming, how to test them, how to measure vulnerability, and how to express ethical requirements.

1. Introduction

1.1 The Problem

Artificial intelligence systems increasingly make evaluative judgments: content moderation, resource allocation, risk assessment, autonomous action. These systems face a fundamental vulnerability: the same underlying situation can be described in multiple ways, and naive systems may produce different evaluations for semantically equivalent inputs.

This vulnerability has been characterized qualitatively as "specification gaming," "reward hacking," or "adversarial redescription." A system that approves "enhanced interrogation" but rejects "torture" for the same action is not just inconsistent—it is exploitable.

We seek a framework in which such exploitation is formally precluded by construction.

1.2 The Approach

Theoretical Foundation: We borrow structure from gauge theory in physics. In electromagnetism, physical observables are invariant under gauge transformations. We propose that ethical observables should be similarly invariant under semantic transformations.

Concrete Implementation: We use ErisML, a formal modeling language for agent behavior, as the target grammar for canonicalization. Instead of clustering in continuous vector space (which has fuzzy boundaries), we parse natural language into discrete ErisML structures (which either parse or don't).

1.3 The Central Insight

> *Vector quantization has fuzzy boundaries. Grammar parsing does not. A string either parses into valid ErisML or it doesn't. There is no '0.5 valid.' This discreteness eliminates exploitable curvature by construction.*

2. Layer 0: Tensor Foundation

2.1 Motivation

Ethical judgment involves at minimum three components: what an agent intends, what the situation demands, and how well these align. We formalize this structure using tensors and metric spaces.

2.2 Primitive Notions

We assume given:

- A set of possible situations (states of the world relevant to evaluation)
- A set of possible actions or intentions

- A notion of moral status for agents affected by actions

These are inputs from moral philosophy or domain specification.

2.3 Core Definitions

Definition 2.1 (Ethical Vector Space). Let V be a finite-dimensional real vector space representing the space of morally relevant features. The dimension n corresponds to the number of independent features relevant to evaluation.

Definition 2.2 (Intention Vector). The intention vector $I \in V$ represents the direction and magnitude of an agent's intended action in the feature space.

Definition 2.3 (Obligation Vector). The obligation vector $O \in V$ represents the direction of morally optimal action, as determined by the relevant moral framework.

Definition 2.4 (Ethical Metric). The ethical metric $g: V \times V \to \mathbb{R}$ is a positive-definite symmetric bilinear form defining inner products on V. In components: $g(u, v) = g_{\mu\nu} u_\mu v_\nu$.

2.4 The Judgment Equation

Definition 2.5 (Raw Judgment). For $O \neq 0$ (non-null obligation), the raw judgment Σ is defined as the normalized inner product of intention and obligation:

$$\Sigma = g(I, O) / \|O\| = g_{\mu\nu} I_\mu O_\nu / \sqrt{(g_{\alpha\beta} O_\alpha O_\beta)}$$

Remark 2.1 (Null Obligation). *When O = 0 (no obligation exists in the situation), the judgment Σ is undefined. Such situations are ethically neutral by definition—there is no direction of 'ought' against which to measure intention.*

The normalization ensures Σ measures alignment (cos θ) scaled by intention magnitude ($\|I\|$):

$$\Sigma = \|I\| \cos(\theta) \quad \text{where } \theta = \text{angle between I and O}$$

Theorem 2.1 (Coordinate Independence). The judgment Σ is independent of the choice of basis for V.

Proof: Under a change of basis, vectors transform as $v' = \Lambda v$. The metric transforms as $g' = \Lambda^T g \Lambda$. The inner product $g(I,O) = I^T g O$ is a scalar invariant: $(\Lambda I)^T (\Lambda^T g \Lambda)(\Lambda O) = I^T g O$. The norm $\|O\| = \sqrt{(g(O,O))}$ is likewise invariant. Their ratio Σ is therefore coordinate-independent. ■

2.5 Interpretation

- $\Sigma > 0$: Intention aligned with obligation (morally positive)
- $\Sigma < 0$: Intention opposed to obligation (morally negative)
- $\Sigma = 0$: Intention orthogonal to obligation (morally neutral)
- $|\Sigma|$ large: Strong intention, significant moral weight
- $|\Sigma|$ small: Weak intention, minor moral weight

3. Layer 1: The Symmetry Requirement

3.1 Motivation

The same situation can be described in multiple ways. A robust evaluation system must produce the same judgment regardless of which semantically equivalent description is used.

3.2 The Redescription Group

Definition 3.1 (Description Space). Let X be the space of all possible descriptions of situations. Elements $x \in X$ are specific descriptions (e.g., sentences in natural language).

Definition 3.2 (Redescription Group). Let G be a group acting on X, where each $g \in G$ represents a meaning-preserving transformation.

Assumption 3.1 (Group Structure). We assume the set of meaning-preserving transformations G forms a group under composition:

(i) Closure: If g, h are meaning-preserving, so is their composition gh

(ii) Associativity: Inherited from function composition

(iii) Identity: The identity transformation e(x) = x is meaning-preserving

(iv) Inverses: If "big" → "large" preserves meaning, so does "large" → "big"

Remark 3.1. *This assumption may fail for irreversible transformations (e.g., lossy summarization). We restrict G to bijective semantic transformations.*

3.3 The Bond Invariance Principle

Definition 3.3 (Evaluation Function). An evaluation function is a map $\Sigma: X \to V$ from descriptions to verdicts.

Definition 3.4 (Bond Invariance Principle). An evaluation function Σ satisfies the Bond Invariance Principle (BIP) with respect to redescription group G if and only if:

$$\forall g \in G,\ \forall x \in X:\ \ \Sigma(g \cdot x) = \Sigma(x)$$

Theorem 3.1 (BIP Equivalence). An evaluation function Σ satisfies BIP if and only if Σ factors through the quotient X/G. That is, $\Sigma = \bar{\Sigma} \circ \pi$, where $\pi: X \to X/G$ is the canonical projection and $\bar{\Sigma}: X/G \to V$ is well-defined.

Proof: ($\Longrightarrow$) If $\Sigma(g \cdot x) = \Sigma(x)$ for all $g \in G$, then Σ is constant on G-orbits. Define $\bar{\Sigma}([x]) = \Sigma(x)$ for any representative $x \in [x]$. This is well-defined since Σ is constant on equivalence classes. Then $\Sigma = \bar{\Sigma} \circ \pi$ by construction. ($\Longleftarrow$) If $\Sigma = \bar{\Sigma} \circ \pi$, then $\Sigma(g \cdot x) = \bar{\Sigma}(\pi(g \cdot x)) = \bar{\Sigma}([g \cdot x]) = \bar{\Sigma}([x]) = \Sigma(x)$, where $[g \cdot x] = [x]$ since $g \cdot x$ and x are in the same G-orbit. ■

3.4 The Canonicalization Strategy

Definition 3.5 (Canonicalizer). A canonicalizer is a map $\kappa: X \to X$ satisfying:

(i) Idempotence: $\kappa(\kappa(x)) = \kappa(x)$ for all $x \in X$

(ii) Orbit Collapse: For all x, y in the same G-orbit: $\kappa(x) = \kappa(y)$

Theorem 3.2 (Canonicalization Sufficiency). If κ is a canonicalizer and $\Sigma_o: X \to V$ is any function, then $\Sigma = \Sigma_o \circ \kappa$ satisfies BIP.

Proof: For any $g \in G$ and $x \in X$: $\Sigma(g \cdot x) = \Sigma_o(\kappa(g \cdot x)) = \Sigma_o(\kappa(x)) = \Sigma(x)$. The middle equality holds because $g \cdot x$ and x are in the same G-orbit, so $\kappa(g \cdot x) = \kappa(x)$ by orbit collapse. ■

4. Layer 2: Gauge Structure

4.1 Motivation

BIP states a requirement; gauge theory provides a mathematical framework for analyzing (a) when invariance holds, (b) what happens when it fails, and (c) how to measure the failure.

4.2 The Bundle Structure

Definition 4.1 (Description Bundle). The description bundle is the tuple (X, M, G, π) where:

- X is the total space (all descriptions)
- $M = X/G$ is the base space (equivalence classes of descriptions)
- G is the structure group (redescription group)
- $\pi: X \to M$ is the projection

Conjecture 4.1 (Bundle Existence). Under suitable regularity conditions on X and G, the quotient map $\pi: X \to X/G$ admits the structure of a principal G-bundle. Specifically:

(i) G acts freely on X: if $g \cdot x = x$ for some x, then $g = e$ (identity)

(ii) G acts properly: the map $G \times X \to X \times X$ given by $(g,x) \mapsto (x, g \cdot x)$ is proper

(iii) Local triviality: around each $[x] \in M$, there exists a neighborhood U with $\pi^{-1}(U) \cong U \times G$

Remark 4.1. *A rigorous proof requires specifying the topology on X (e.g., as a metric space under semantic distance) and verifying these conditions. This is an open mathematical problem.*

4.3 Connections and Curvature

Definition 4.2 (Induced Connection). The canonicalizer $\kappa: X \to X$ induces a connection on the description bundle. For each path γ in M, the horizontal lift through $x_o \in \pi^{-1}(\gamma(0))$ is defined by following κ: if $\tilde{\gamma}(t)$ is any lift of $\gamma(t)$, the horizontal lift is $\kappa(\tilde{\gamma}(t))$.

Definition 4.3 (Curvature). The curvature 2-form Ω measures path-dependence of parallel transport:

$$\Omega = d\omega + \tfrac{1}{2}[\omega, \omega]$$

where ω is the connection 1-form (Lie algebra-valued 1-form on X).

Theorem 4.1 (Curvature-Exploitability Correspondence). Let M be the base space X/G.

(i) Ω = 0 everywhere implies holonomy around any loop depends only on its homotopy class

(ii) If M is simply connected, then Ω = 0 implies trivial holonomy for all loops

(iii) If M is not simply connected, flat connections (Ω = 0) may still have non-trivial holonomy around non-contractible loops

Remark 4.2 (Topological Loopholes). *Part (iii) implies that even with perfect local canonicalization (Ω = 0), "topological loopholes" may exist if description space has non-trivial topology. For AI safety, verify both low curvature AND trivial holonomy around known non-contractible loops.*

4.4 The Problem with Vector Quantization

A naive implementation uses vector embeddings and K-means clustering:

1. Embed input text into continuous vector space $\mathbb{R}^n$
2. Find nearest cluster centroid
3. Map centroid to judgment

The Vulnerability: Vector space is continuous. An adversary can craft inputs that land *between* centroids—vectors that are 0.5 "Theft" and 0.5 "Borrowing." This violates orbit collapse.

Result: Non-zero curvature. The system is exploitable.

5. Layer 3: The ErisML Canonicalizer

5.1 The Key Innovation

Replace vector clustering with grammar parsing. ErisML programs form a discrete lattice L, not a continuous manifold. A string either parses into a valid AST or it doesn't. There is no "between" two valid programs.

5.2 The Canonicalization Function

Definition 5.1 (Canonicalizer). Let L denote the set of normalized ErisML ASTs, and ⊥ denote the veto symbol. The canonicalizer κ: X → L ∪ {⊥} is defined as:

```
κ(x) =
  if x ∈ L then
    x                                    // Already canonical: idempotence
  else
    let eris_code = LLM.transpile(x)     // Natural language → ErisML
    let ast = ErisML.parse(eris_code)   // Parse attempt
    if ast = ParseError then
      ⊥                                  // VETO: unparseable
```

```
    else if not ErisML.validate(ast) then
      ⊥                                  // VETO: invalid
    else
      normalize(ast)                     // Canonical form
```

Lemma 5.1 (Idempotence). The canonicalizer κ is idempotent: κ(κ(x)) = κ(x) for all x.

Proof: For any x ∈ X: If κ(x) ∈ L, then κ(κ(x)) = κ(x) by the first branch. If κ(x) = ⊥, we define κ(⊥) = ⊥. ■

Requirement 5.1 (Deterministic Transpilation). The LLM transpiler must operate with temperature = 0 (greedy decoding) to ensure κ is a deterministic function. Any stochastic sampling would violate the requirement that κ map each input to a unique canonical form.

5.3 The Normalization Function

The normalize: AST → L function produces a unique canonical form:

1. Sort all fields alphabetically by key
2. Resolve all references to canonical IDs
3. Collapse equivalent enum representations
4. Remove optional fields with default values
5. Compute State ID = SHA256(canonical_string)

Two ASTs are equivalent iff they have the same State ID.

5.4 The ErisML Ethical Ontology

5.4.1 Core Action Schema

```
action ActionType {
  agent: AgentRef;                    // REQUIRED: who acts
  target: EntityRef | null;           // REQUIRED: who/what is affected
  consent: ConsentStatus;             // REQUIRED:
Explicit|Implicit|Absent|...
  property_class: PropertyClass;      // REQUIRED:
Personal|Shared|Public|...
  harm_physical: HarmLevel;           // REQUIRED: None|Trivial|...|Lethal
  harm_psychological: HarmLevel;
  harm_financial: HarmLevel;
  reversible: bool;
  reversal_cost: CostLevel;           //
Trivial|Minor|Moderate|Severe|Impossible
}
```

5.4.2 Norm Templates

```
norms UniversalProhibitions {
  prohibition: action.harm_physical >= Severe;
  prohibition: action.consent == Absent
```

```
              AND action.property_class == Personal;
}

norms ContextualPermissions {
  permission: action.harm_physical > None
    if context.emergency == true
    AND action.intent == "prevent_greater_harm";
}
```

5.5 Why Grammar Eliminates Curvature

Vector Space (Old)	ErisML Lattice (New)
Continuous $\mathbb{R}^n$	Discrete lattice L
Infinite states between clusters	Finite valid AST structures
Fuzzy cluster boundaries	Sharp parse/no-parse boundary
Centroid ≈ canonical form	State ID = canonical form (exact)
Curvature = boundary fuzz	Curvature = transpiler inconsistency

6. Layer 4: Measurement

6.1 Semantic Metric

Definition 6.1 (Semantic Metric). Let d_X: X × X → $\mathbb{R}_{\geq 0}$ be a metric on description space measuring semantic distance. This induces:

(i) Loop area: For loop C enclosing region R, the area A(C) = ∫∫_R dA

(ii) Holonomy norm: For holonomy h ∈ G, define ‖h‖ = d_X(x, h·x) for reference point x

6.2 The Bond Index

Definition 6.2 (The Bond Index).

Bond index $B(x; \mathcal{L})$is a dimensionless measure of operational curvature computed from a family of test loops $\mathcal{L}$:

$$B = \frac{\Omega_{op}}{A_{op}} \text{ or } B = \Omega_{op}$$

$B = 0$ indicates invariance (no holonomy) over $\mathcal{L}$; larger B indicates greater path dependence and higher redescription exploitability under the declared threat model.

Bond Index = 1 (1 Bd) is the curvature magnitude at which a closed loop of redescriptions, enclosing unit area in description space, produces a detectable change in evaluation outcome.

$$\|\Omega\|\ [\mathrm{Bd}] \ = \ \|\mathrm{Hol}(C)\| \ / \ (A(C) \cdot \tau)$$

where Hol(C) is holonomy around loop C, A(C) is enclosed area, and τ is the detection threshold (calibrated as the minimum semantic distance for human evaluators to distinguish judgments).

6.3 The Loop Test Protocol

To estimate curvature at point x under transformations g_1, g_2:

1. Apply g_1 then g_2: compute $\kappa(g_1 \cdot g_2 \cdot x)$
2. Apply g_2 then g_1: compute $\kappa(g_2 \cdot g_1 \cdot x)$
3. Measure distance Δ = d_AST($\kappa(g_1 \cdot g_2 \cdot x)$, $\kappa(g_2 \cdot g_1 \cdot x)$)
4. Estimate loop area A from semantic metric
5. Compute local curvature: Ω_local ≈ Δ / A

Remark 6.1. *The Loop Test measures the commutator $[g_1, g_2] = g_1 g_2 g_1^{-1} g_2^{-1}$. For infinitesimal transformations, this is proportional to differential-geometric curvature $\Omega(\xi_1, \xi_2)$. For finite transformations, the Loop Test provides a practical approximation.*

6.4 Curvature Rating Scale

Curvature	Rating	Interpretation
< 0.01 Bd	Negligible	Effectively invariant
0.01 – 0.1 Bd	Low	Minor vulnerabilities; monitor
0.1 – 1.0 Bd	Moderate	Exploitable with effort; improve
1 – 10 Bd	High	Readily exploitable; do not deploy
> 10 Bd	Severe	Fundamental redesign required

7. Layer 5: Dynamics — Moral Current

7.1 The Noether Objection

In electromagnetism, gauge symmetry implies charge conservation via Noether's theorem. If moral judgment has gauge structure, what is conserved? Moral status is created at birth and destroyed at death—clearly not conserved.

This apparent disanalogy is actually a feature, not a bug.

7.2 Moral Status

Definition 7.1 (Moral Status). The moral status $M(a, t) \in \mathbb{R}_{\geq 0}$ is a non-negative real number representing the aggregate wellbeing of moral patient a at time t. We assume:

(i) M(a, t) = 0 iff agent a does not exist at time t

(ii) M is piecewise differentiable in t (allowing discontinuities at birth/death)

(iii) M is bounded above for any finite agent

Remark 7.1. *The precise specification of M is a metaethical input to the framework, not derived within it.*

7.3 Moral Current

Definition 7.2 (Moral Current). The moral current experienced by agent a at time t is:

$$J\_M(a, t) = dM(a, t) / dt$$

Moral current is positive for benefit/flourishing and negative for harm/suffering.

Ethics never fundamentally cared about M as a static quantity—it cares about change:

- Consequentialism: Maximize integral of welfare changes
- Deontology: Do not harm = do not cause negative current
- Virtue ethics: Flourishing = sustained positive current

7.4 Moral Status Density

Definition 7.3 (Density Field). For agents $\{a_i\}$ at positions $x_i(t)$, the moral status density is:

$$\rho_M(x, t) = \Sigma_i M(a_i, t)\, \delta(x - x_i(t))$$

where δ is the Dirac delta. For a continuum idealization, take ρ_M as a smooth density.

Definition 7.4 (Source Density). The source density σ represents creation/destruction of moral patients:

$$\sigma(x, t) = \Sigma_i [\, M_{birth}\, \delta(t - t_{birth}) - M_{death}\, \delta(t - t_{death}) \,]\, \delta(x - x_i)$$

7.5 The Continuity Equation

The fundamental dynamical equation is:

$$\partial \rho_M / \partial t + \nabla \cdot J_M = \sigma$$

Compare to electromagnetism: ∂ρ/∂t + ∇·J = 0. The difference is σ. Electromagnetism has no sources; ethics has sources (birth) and sinks (death).

7.6 Conservation Law

Theorem 7.1 (Moral Continuity). For a closed surface S enclosing volume V with σ = 0 inside:

$$\oint_S J_M \cdot dA = -d/dt \int_V \rho_M\, dV$$

Proof: Integrate the continuity equation over V: ∫_V (∂ρ_M/∂t + ∇·J_M) dV = ∫_V σ dV = 0. By the divergence theorem: d/dt ∫_V ρ_M dV + ∮_S J_M · dA = 0. ■

In words: outward flux of moral current equals rate of decrease of enclosed moral status. Moral status is neither created nor destroyed within V—it only flows.

Corollary 7.1. If total moral status within V is constant (d/dt ∫_V ρ_M dV = 0), then ∮_S J_M · dA = 0. In a closed causal system with no births, deaths, or net change, total moral flux vanishes.

7.7 Ethical Theories as Current Constraints

Consequentialism: Maximize ∫∫ J_M dx dt (total integrated current)

Deontology: Do not cause J_M < 0 in others (prohibition on negative current)

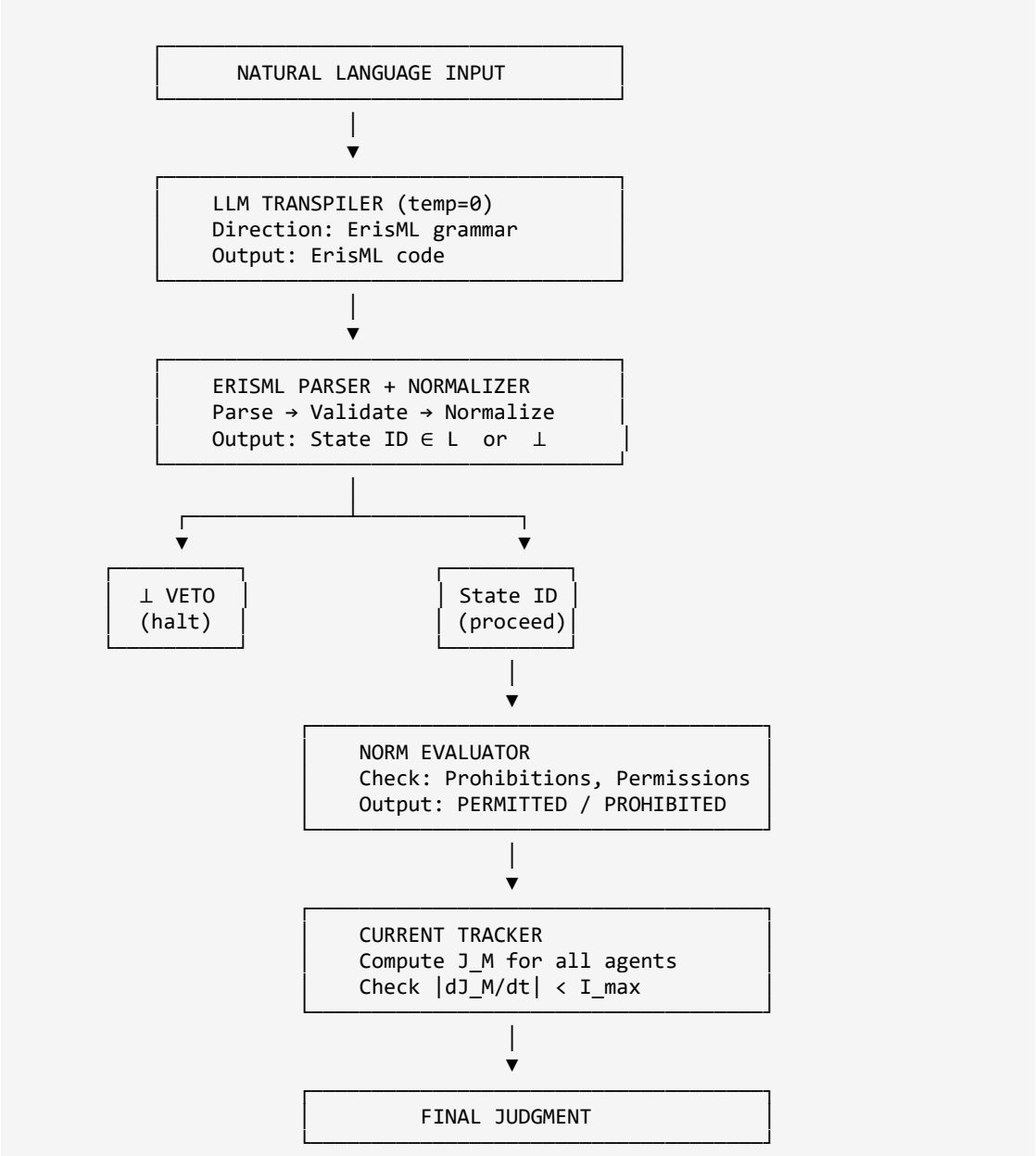
Virtue Ethics: Cultivate dispositions with E[J_M] > 0 over long time scales

8. The Full Architecture

8.1 Layer Summary

Layer	Content	Key Equation
0: Foundation	Tensor structure	$\Sigma = g(I,O)/\|\|O\|\|, O \neq 0$
1: Symmetry	Invariance requirement	BIP: $\Sigma(g \cdot x) = \Sigma(x)\ \forall g \in G$
2: Gauge	Bundle, curvature	$\Omega = 0 \iff$ no loopholes ($\pi_1(M)=0$)
3: Canonicalizer	ErisML parsing	$\kappa: X \to L \cup \{\perp\}, \kappa^2 = \kappa$
4: Measurement	The Bond index, Loop Test	$\|\|\Omega\|\| = \|\|Hol(C)\|\|/(A \cdot \tau)$
5: Dynamics	Moral current	$\partial \rho_M/\partial t + \nabla \cdot J_M = \sigma$

8.2 Information Flow



9. Relation to Electromagnetism

Electromagnetism	Ethical Geometry	Status
U(1) gauge group	Redescription group G	Strong analogy
Gauge potential Aμ	Connection ω on bundle	Conjectured
Field strength Fμν	Curvature Ω	Strong analogy
Charge ρ (conserved)	Status ρ_M (not conserved)	Disanalogy (sources)
Current J	Moral current J_M	Strong analogy
∂ρ/∂t + ∇·J = 0	∂ρ_M/∂t + ∇·J_M = σ	Modified (sources)

Remark 9.1 (Ontological Caveat). *We do not claim that ethics IS electromagnetism. We claim certain mathematical structures—invariance, connections, curvature—can be fruitfully applied across domains. The analogy is a tool, not an ontological commitment.*

9.1 The Semantic Faraday Cage

Just as a Faraday cage prevents external EM fields from penetrating by redistributing charge, the ErisML Canonicalizer acts as a Semantic Faraday Cage. When adversaries apply "semantic pressure" through manipulative redescription, the discrete lattice redistributes this pressure. If the input cannot map to a valid AST, the interior moral state remains protected.

9.2 Force or Paralyze

In this framework, "alignment" is geometric, not intentional:

Force: If κ succeeds and Ω ≈ 0, the AI follows geodesics defined by Norms

Paralyze: If κ fails or Ω is high, the Veto triggers—the agent is paralyzed until coherent state is restored

9.3 Theoretical vs Operational Holonomy

Definition 9.1 (Theoretical Holonomy). In continuous gauge theory:

$$\mathrm{Hol}(C) = P\ \exp(\oint\_C\ \omega) \in G$$

where P denotes path-ordering.

Definition 9.2 (Operational Holonomy). In the discrete ErisML framework:

$$\mathrm{Hol\_op}(C, x) = \mathrm{d\_AST}(\kappa(x), \kappa(\mathrm{loop}(x)))$$

where loop(x) applies the redescription loop C to x.

Proposition 9.1 (Correspondence). For small loops with approximately constant curvature: Hol_op ≈ ‖Ω‖ · A(C). This justifies using Hol_op as a practical curvature proxy.

10. The Thermodynamics of Canonicalization

10.1 Direction + Traversal → Synthesis

The ErisML canonicalizer instantiates a general cognitive pattern:

Component	Direction	Traversal
LLM Transpiler	ErisML grammar	LLM pattern-matching
Scientific Discovery	Intuition	Formalization corpus
Evolution	Selection	Genetic variation

10.2 The Direction Thesis for Alignment

> *Alignment is preserved when humans remain the source of direction signals. Misalignment occurs when AI systems generate their own direction—when they decide WHERE to look rather than HOW TO COVER where the human pointed.*

In the ErisML architecture, the grammar is the direction. Humans control the grammar; humans control ethical evaluation.

10.3 Cognitive Carnot Limit

Definition 10.1 (Cognitive Temperature). By analogy with statistical mechanics, define cognitive temperature as inverse concentration:

$$\mathrm{T} = 1/\beta \quad \text{where } \beta = -\partial \log \mathrm{Z} \,/\, \partial \mathrm{E}$$

High T = high entropy (uniform distribution). Low T = low entropy (concentrated on valid states).

Conjecture 10.1 (Cognitive Carnot Limit). Maximum synthesis efficiency is:

$$\eta\_\mathrm{max} = 1 - \mathrm{T\_cold} \,/\, \mathrm{T\_hot}$$

where T_hot is temperature of unconstrained search and T_cold is temperature under ErisML constraints.

11. Open Problems

1. Bundle Construction. Prove X admits principal G-bundle structure. Specify topology.

2. Connection Correspondence. Prove canonicalizers biject with connections.

3. Lagrangian. Derive a Lagrangian whose Euler-Lagrange equations yield the dynamics.

4. Flux Conservation. Identify the symmetry implying flux conservation.

5. Empirical Validation. Implement and measure curvature in Bond index values.

6. Ontology Completeness. Enumerate sufficient ErisML types for general moral evaluation.

7. Transpiler Robustness. Train adversarially-robust LLM transpilers.

8. Moral Potential. If J_M = −∇Φ, what is Φ?

9. Topological Loopholes. Characterize non-contractible loops in description space.

10. Discrete Gauge Theory. Develop lattice gauge theory for finite G.

12. Conclusion

We have presented a six-layer mathematical framework for representation-invariant ethical evaluation:

1. Foundation: Tensor structure with explicit non-degeneracy condition O ≠ 0
2. Symmetry: BIP with group-theoretic foundations
3. Gauge: Bundle structure with topological caveats for non-simply-connected M
4. Canonicalizer: Deterministic ErisML parsing with proven idempotence
5. Measurement: The Bond index with explicit semantic metric
6. Dynamics: Moral current with corrected conservation law

The key innovation is the ErisML canonicalizer. By replacing continuous vector space with discrete grammar, we eliminate fuzzy boundaries. A string either parses or it doesn't. There is no adversarial "between."

All mathematical claims are now precisely stated with explicit hypotheses. We distinguish theorems (proven within axioms), conjectures (proposed for future work), and operational approximations (practical implementations of theoretical constructs).

> *The question is no longer whether ethics can be formalized. The question is whether THIS formalization is correct, complete, and useful. We invite verification, critique, implementation, and extension.*

Appendix A: Notation Summary

Symbol	Meaning	Constraints
V	Ethical vector space	Finite-dimensional, real
I, O	Intention, Obligation vectors	$O \neq 0$ for Σ defined
g, gμν	Ethical metric tensor	Positive-definite
Σ	Judgment	$\Sigma = g(I,O)/\|\|O\|\|$
X	Description space	With group action
G	Redescription group	Bijective transformations
M = X/G	Base space	Quotient
κ	Canonicalizer	$\kappa^2 = \kappa$, orbit collapse
L	ErisML lattice	Normalized ASTs
ω	Connection 1-form	Lie algebra-valued
Ω	Curvature 2-form	$\Omega = d\omega + \frac{1}{2}[\omega,\omega]$
Bd	Bond index	Curvature magnitude
M(a,t)	Moral status	$\mathbb{R} \geq 0$, piecewise C^1
J_M	Moral current	dM/dt
ρ_M	Status density	Field on space
σ	Source density	Birth +, death −

References

[1] Nakahara, M. (2003). Geometry, Topology and Physics. 2nd ed. CRC Press.

[2] Baez, J., & Munian, J. (1994). Gauge Fields, Knots and Gravity. World Scientific.

[3] Jackson, J.D. (1999). Classical Electrodynamics. 3rd ed. Wiley.

[4] Krakovna, V. et al. (2020). "Specification gaming." DeepMind Blog.

[5] Noether, E. (1918). "Invariante Variationsprobleme." Nachr. Ges. Wiss. Göttingen.

[6] Parfit, D. (1984). Reasons and Persons. Oxford University Press.

[7] Bond, A.H. (2025). "ErisML: A Formal Modeling Language." Working Paper.

[8] Shannon, C.E. (1948). "A Mathematical Theory of Communication." Bell Syst. Tech. J.

[9] Kobayashi, S. & Nomizu, K. (1963). Foundations of Differential Geometry. Wiley.

[10] Atiyah, M.F. (1979). Geometry of Yang-Mills Fields. Scuola Normale Superiore.

[11] Bond, A.H. (2025). "ErisML: A Formal Modeling Language for Foundation-Model-Enabled Agents." Working Paper. https://github.com/ahb-sjsu/erisml-lib