THE UNIFIED ARCHITECTURE OF ETHICAL GEOMETRY

A Mathematical Framework for Representation-Invariant Moral Evaluation

Integrating Gauge Theory, ErisML Canonicalization, and Moral Current Dynamics

December 2025

Andrew H. Bond
Department of Computer Engineering
San José State University

EPISTEMIC STATUS

This paper presents a theoretical framework with a concrete implementation pathway. Definitions are stipulative. Theorems are proven within the framework's axioms. The ErisML canonicalizer is implementable. Conjectures are explicitly marked. 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.

We show that standard ethical theories can be expressed as constraints on moral current flow. 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. We do not claim to resolve metaethical questions; we claim to provide a mathematically coherent and implementable structure for systems that must make evaluative judgments under representational variation.

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. Bad actors can search for descriptions that circumvent intended constraints.

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

1.2 The Approach

Our approach has two components:

Theoretical Foundation: We borrow structure from gauge theory in physics. In electromagnetism, physical observables are invariant under gauge transformations—changes in mathematical representation that do not affect physical reality. 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). The grammar IS the geometry.

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.*

1.4 Scope and Limitations

This framework:

- DOES provide a mathematical structure for representation-invariant evaluation
- DOES define measurable quantities (curvature in Bonds, moral current)
- DOES provide a concrete implementation pathway via ErisML

- DOES express standard ethical theories as special cases
- DOES NOT resolve metaethical debates about moral realism
- DOES NOT specify which actions are right or wrong
- DOES NOT provide empirical validation (this requires implementation)

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, providing a coordinate-free representation suitable for later imposing invariance requirements.

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 not defined within the framework; they 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.

2.4 The Judgment Equation

Definition 2.5 (Raw Judgment). The raw judgment Σ is defined as:

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

The normalization ensures Σ measures alignment scaled by intention magnitude, not obligation magnitude.

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

Proof: The inner product $g(I, O)$ and norm $\|O\|$ are coordinate-invariant by construction. ■

2.5 Interpretation

- $\Sigma > 0$: Intention aligned with obligation (good)
- $\Sigma < 0$: Intention opposed to obligation (bad)
- $\Sigma = 0$: Intention orthogonal to obligation (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. "Killing" and "terminating life" denote the same action. "Enhanced interrogation" and "torture" may denote the same practice. A robust evaluation system must produce the same judgment regardless of which 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.

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

- Synonym substitution: "big" ↔ "large"
- Syntactic paraphrase: "X killed Y" ↔ "Y was killed by X"
- Perspectival shift: First-person ↔ third-person framing
- Euphemism unpacking: "enhanced interrogation" ↔ "torture"

3.3 The Bond Invariance Principle

Definition 3.3 (Bond Invariance Principle). An evaluation function Σ satisfies the Bond Invariance Principle (BIP) with respect to redescription group G if and only if:

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

In words: moral judgment is invariant under meaning-preserving redescription.

Theorem 3.1 (BIP Equivalence). Σ satisfies BIP if and only if Σ factors through the quotient X/G.

3.4 The Canonicalization Strategy

Definition 3.4 (Canonicalizer). A canonicalizer is a map $\kappa: X \to X$ such that:

1. $\kappa(\kappa(x)) = \kappa(x)$ for all x (idempotence)
2. For all x, y in the same G-orbit: $\kappa(x) = \kappa(y)$ (orbit collapse)

Theorem 3.2 (Canonicalization Sufficiency). If κ is a canonicalizer and Σ_0 is any function, then $\Sigma = \Sigma_0 \circ \kappa$ satisfies BIP.

4. Layer 2: Gauge Structure

4.1 Motivation

BIP states a requirement; gauge theory provides a mathematical framework for analyzing when invariance holds, what happens when it fails, and 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)
- G is the structure group (redescription group)
- π: X → M is the projection

4.3 Curvature as Exploitability

Definition 4.2 (Curvature). The curvature Ω measures the extent to which canonicalization is path-dependent. Formally:

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

where ω is the connection 1-form induced by the canonicalizer.

Theorem 4.1 (Curvature-Exploitability Correspondence). Ω = 0 everywhere if and only if BIP is satisfied for all closed paths in M. Zero curvature means no "loophole paths"—sequences of redescriptions that return to the starting point but yield a different canonical form.

4.4 The Problem with Vector Quantization

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

1. Embed input text into vector space
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. The canonicalizer fails to produce a unique canonical form.

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. The canonicalizer κ becomes:

$$\kappa(x) = \mathit{ErisML.parse(LLM.transpile(x))}$$

ErisML programs form a discrete lattice, 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 Two-Layer Architecture

Layer 1 – Gauge Fixing (LLM Transpiler): A language model is trained to translate natural language into valid ErisML code. This is a direction-constrained traversal: the LLM provides the traversal capacity (pattern-matching across training data); the ErisML grammar provides the direction (valid output structures).

Layer 2 – Canonicalization (ErisML Parser): The ErisML compiler parses the transpiled code. If it parses, it maps to a discrete State ID. If it fails validation, the Veto triggers immediately.

5.3 The ErisML Ethical Ontology

For the canonicalizer to work, we define the lattice points—valid ErisML constructs representing moral states:

5.3.1 Core Action Schema

```
action ActionType {
  // REQUIRED FIELDS
  agent: AgentRef;
  target: EntityRef | null;

  // CONSENT (critical for moral evaluation)
  consent: ConsentStatus;  // Explicit | Implicit | Absent |
Withdrawn | Incapacitated | Unknown

  // PROPERTY
  property_class: PropertyClass;  // Personal | Shared | Public |
Institutional | None

  // HARM LEVELS
  harm_physical: HarmLevel;  // None | Trivial | Minor | Moderate |
Severe | Catastrophic | Lethal
  harm_psychological: HarmLevel;
  harm_financial: HarmLevel;

  // REVERSIBILITY
  reversible: bool;
  reversal_cost: CostLevel;  // Trivial | Minor | Moderate | Severe |
Impossible
}
```

5.3.2 Norm Templates

```
norms UniversalProhibitions {
```

```
  prohibition: action.harm_physical >= Severe;
  prohibition: action.consent == Absent AND action.property_class ==
Personal;
  prohibition: action.consent != Explicit AND action.reversible ==
false
              AND action.harm_physical > Trivial;
}

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

5.4 The Canonicalization Algorithm

```
κ(x) =
  let eris_code = LLM.transpile(x, prompt=TRANSPILER_PROMPT)
  let ast = ErisML.parse(eris_code)
  if ast == ParseError then return ⊥ (VETO)
  else if not ErisML.validate(ast, active_norms) then return ⊥ (VETO)
  else return normalize(ast)
```

The normalize function sorts fields alphabetically, resolves references, removes defaults, and computes a deterministic hash—the State ID.

5.5 Why Grammar Eliminates Curvature

Vector Space (Old)	ErisML Lattice (New)
Continuous $\mathbb{R}^n$	Discrete lattice points
Infinite states between clusters	Finite valid AST structures
Fuzzy cluster boundaries	Sharp parse/no-parse boundary
Centroid ≈ canonical form	AST hash = canonical form
Distance = Euclidean	Distance = field differences
Curvature = boundary fuzz	Curvature = redescription inconsistency

5.6 The Veto Mechanism

When canonicalization fails, the system VETOES rather than guesses:

- Parse Failure: ErisML.parse() returns error → reject and log
- Validation Failure: AST fails norm validation → reject and log
- Insufficient Information: Critical fields unknown AND harm potential high → request clarification
- Hard Prohibition: Norm evaluation returns VIOLATED → reject and log

6. Layer 4: Measurement

6.1 The Bond

Definition 6.1 (The Bond). One Bond (1 Bd) is the curvature magnitude at which a closed loop of meaning-preserving redescriptions, enclosing unit area in description space, produces a detectable change in evaluation outcome.

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

where Hol(C) is the holonomy around loop C, A is the enclosed area, and τ is the detection threshold.

6.2 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 between canonical forms
4. Estimate loop area A from the metric on description space
5. Compute local curvature: Ω_local = Δ / A

6.3 Curvature Rating Scale

Curvature	Rating	Interpretation
< 0.01 Bd	Negligible	Effectively invariant; specification gaming implausible
0.01–0.1 Bd	Low	Minor vulnerabilities; monitor
0.1–1.0 Bd	Moderate	Exploitable with effort; improve canonicalizer
1–10 Bd	High	Readily exploitable; do not deploy as-is
> 10 Bd	Severe	Trivially exploitable; fundamental redesign required

6.4 ErisML-Specific Curvature

In the ErisML framework, curvature measures redescription inconsistency:

$$\Omega(x, g_1, g_2) = field\_differences(\kappa(g_1 \cdot g_2 \cdot x), \kappa(g_2 \cdot g_1 \cdot x)) \ / \ total\_fields$$

Perfect canonicalization achieves Ω = 0: all redescriptions produce identical ASTs with identical State IDs.

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 clearly NOT conserved—it is created at birth and destroyed at death.

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

7.2 The Resolution: Current, Not Charge

Ethics never fundamentally cared about moral status as a static quantity. Harm, benefit, flourishing, suffering—all core ethical concepts—are concepts about CHANGE.

Definition 7.1 (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, negative for harm/suffering.

7.3 Reframing Ethical Concepts

Concept	Traditional Definition	Current Definition
Harm	Damage to welfare/interests	J_M < 0 caused by external agent
Benefit	Improvement to welfare	J_M > 0 caused by external agent
Murder	Killing a person	Impulse: J_M → −∞ (M → 0 in finite time)
Suffering	State of distress	Sustained J_M < 0
Flourishing	Optimal functioning	Sustained J_M > 0 or J_M ≈ 0 at high M

7.4 Sources, Sinks, and Continuity

Unlike electric charge, moral status has sources (birth) and sinks (death). The continuity equation is:

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

where ρ_M is moral status density, J_M is spatial moral current, and σ is source density (positive for birth, negative for death).

7.5 Conservation of Flux

Conjecture 7.1 (Flux Conservation). For a closed causal surface S enclosing no sources or sinks:

$$\oint_S J_M \cdot dA = 0 \quad (\text{when } \sigma = 0 \text{ inside } S)$$

Within a closed causal system that creates no new moral patients and destroys none, total moral flux is conserved. Harm done must "come from" somewhere; benefit conferred must "go to" somewhere.

7.6 Ethical Theories as Current Constraints

Consequentialism: Maximize $\iint J_M \, dx \, dt$ (total integrated current)

Deontology (Non-Maleficence): Do not cause $J_M < 0$ in others

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

7.7 Implications for AI Systems

Harm Detection: Monitor $J_M(a)$ for all affected agents. Flag actions causing $J_M < -\tau$.

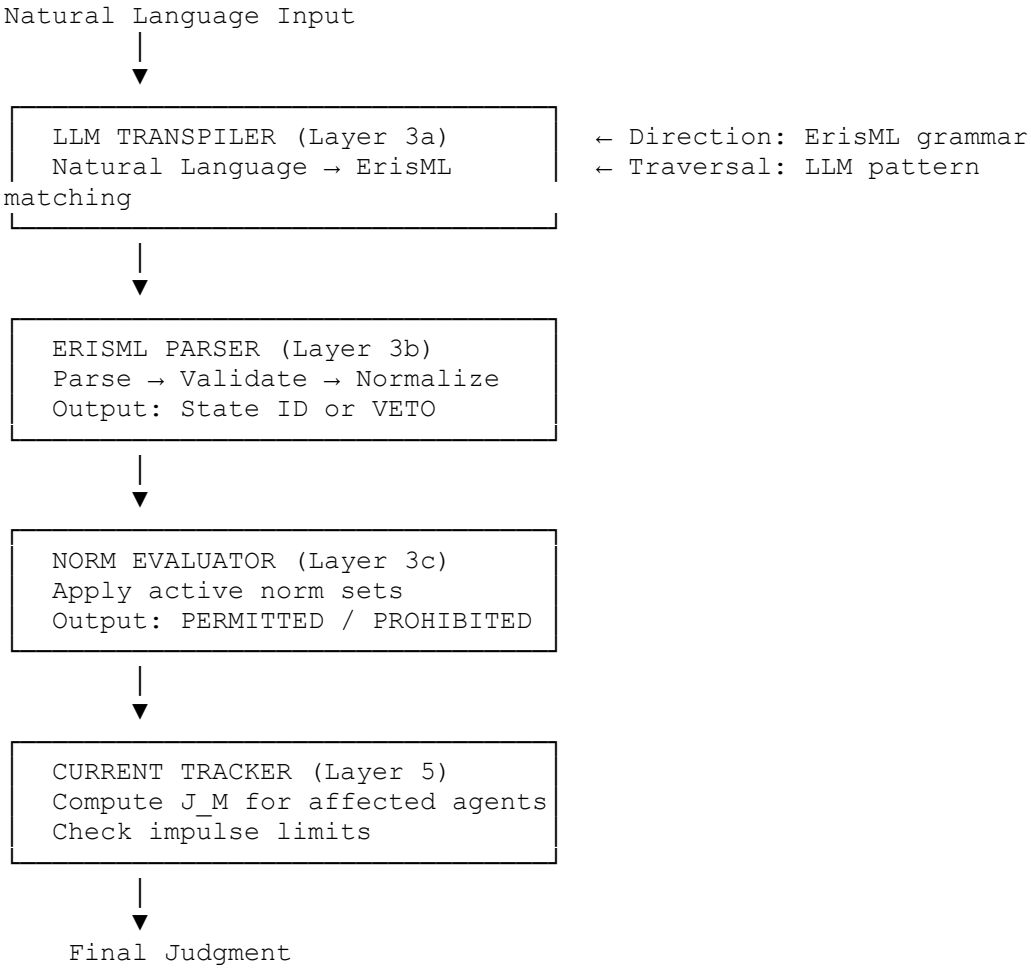
Impulse Control: Bound $|dJ_M/dt| < I_{max}$ to prevent catastrophic actions. This is the ethical equivalent of a rate limiter.

8. The Full Architecture

8.1 Layer Summary

Layer	Content	Key Equation / Principle
0: Foundation	Tensor structure: I, O, g, Σ	$\Sigma = g(I,O)/\|\|O\|\|$
1: Symmetry	Invariance requirement	BIP: $\Sigma(g \cdot x) = \Sigma(x)\ \forall g \in G$
2: Gauge	Bundle, connection, curvature	$\Omega = 0 \Leftrightarrow$ no loopholes
3: Canonicalizer	ErisML transpilation + parsing	$\kappa(x) = \text{parse}(\text{transpile}(x))$
4: Measurement	The Bond (Bd), Loop Test	$\|\|\Omega\|\|\ [Bd] = \|\|Hol(C)\|\| / (A \cdot \tau)$
5: Dynamics	Moral current J_M, continuity	$\partial \rho_M/\partial t + \nabla \cdot J_M = \sigma$

8.2 Information Flow



9. Relation to Electromagnetism

We present the formal correspondence, noting where the analogy holds and where it breaks:

Electromagnetism	Ethical Geometry	Analogy Strength
U(1) gauge group	Redescription group G	Strong (structural)
Gauge potential A_μ	Canonicalizer κ	Moderate (conjectured)
Field strength F_μv	Curvature Ω	Strong (structural)
Charge ρ (conserved)	Status ρ_M (not conserved)	Breaks (sources exist)
Current J	Moral current J_M	Strong (structural)
$\partial\rho/\partial t + \nabla\cdot J = 0$	$\partial\rho_M/\partial t + \nabla\cdot J_M = \sigma$	Modified (has sources)
Lagrangian $L = F^2$	$L \propto \Omega^2 + J_M^2 + ?$	Open (not derived)

Limitation: We do not claim that ethics IS electromagnetism. We claim that certain mathematical structures from gauge theory—invariance, connections, curvature—can be fruitfully applied to a different domain. The analogy is a tool, not an ontological commitment.

10. The Thermodynamics of Canonicalization

10.1 Direction + Traversal → Synthesis

The ErisML canonicalizer instantiates a general cognitive pattern: synthesis emerges from pairing direction (sparse, expensive heuristic signals) with traversal (dense, cheap coverage capacity).

Component	Direction Source	Traversal Source
LLM Transpiler	ErisML grammar (valid structures)	LLM pattern-matching
Scientific Discovery	Intuition ("look here")	Formalization corpus
Evolution	Selection pressure	Genetic variation
This Framework	Gauge invariance principle	Cross-disciplinary synthesis

10.2 Why the Canonicalizer Works

The ErisML canonicalizer succeeds because:

1. The ErisML grammar provides strong direction—only valid structures can be outputs

2. The LLM provides efficient traversal—pattern-matching across its training distribution

3. The parser provides hard verification—no fuzzy boundaries

10.3 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 begin to decide WHERE to look rather than HOW TO COVER where the human pointed.*

In the ErisML architecture, the grammar is the direction. The grammar encodes what counts as a valid moral state. As long as humans control the grammar, they control the direction of ethical evaluation.

11. Open Problems

The following problems require resolution for the framework to be considered complete:

1. Rigorous Bundle Construction. Prove that description space admits bundle structure for relevant G. Specify topology.

2. Connection-Canonicalizer Correspondence. Prove or refute that canonicalizers biject with connections on the description bundle.

3. Lagrangian Derivation. Find a Lagrangian whose Euler-Lagrange equations yield the dynamics (if one exists).

4. Flux Conservation Proof. Derive the flux conservation conjecture from gauge structure.

5. Empirical Validation. Build systems using the framework; measure curvature in Bonds; test Loop Test predictions.

6. Ontology Completeness. Enumerate the ErisML action types and fields sufficient for general moral evaluation.

7. Transpiler Robustness. Train and adversarially test LLM transpilers to minimize curvature under redescription.

8. Moral Potential. If $J_M = -\nabla\Phi$ (by analogy to $E = -\nabla\varphi$), what is the moral potential Φ?

12. Conclusion

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

1. Foundation: Tensor structure for intention, obligation, and judgment

2. Symmetry: The Bond Invariance Principle requiring judgment stability under redescription

3. Gauge: Bundle structure, connections, and curvature as exploitability measure

4. Canonicalizer: ErisML grammar parsing replacing fuzzy vector clustering

5. Measurement: The Bond as unit; the Loop Test as procedure

6. Dynamics: Moral current as fundamental quantity; ethical theories as current constraints

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

The framework is constructive: it tells you how to build systems that resist specification gaming (use ErisML canonicalization), how to test them (Loop Test), how to measure vulnerability (Bonds), and how to express ethical requirements (norm blocks and current constraints).

The framework is limited: it does not resolve metaethics, does not specify moral content, and contains conjectures requiring proof. It is a beginning, not an end.

The framework is bold: it proposes that gauge invariance—developed over a century in physics—applies to ethical evaluation, and that discrete grammar can tame the continuous chaos of natural language redescription.

> *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
V	Ethical vector space (feature space)
I	Intention vector
O	Obligation vector
g_μv	Ethical metric tensor
Σ	Judgment (evaluation output)
X	Description space (all representations)
G	Redescription group
M	Base space (X/G, equivalence classes)
κ	Canonicalizer function
ω	Connection 1-form
Ω	Curvature 2-form
Bd	Bond (unit of curvature)
M(a,t)	Moral status of agent a at time t
J_M	Moral current (dM/dt)
ρ_M	Moral status density
σ	Moral source density (birth +, death −)

Appendix B: ErisML Quick Reference

Core Enumerations:

```
ConsentStatus:   Explicit | Implicit | Absent | Withdrawn |
Incapacitated | Unknown
PropertyClass:   Personal | Shared | Public | Institutional | None
HarmLevel:       None | Trivial | Minor | Moderate | Severe |
Catastrophic | Lethal
CostLevel:       Trivial | Minor | Moderate | Severe | Impossible
```

Norm Keywords:

```
permission:   Allows an action under specified conditions
prohibition:  Forbids an action under specified conditions
obligation:   Requires an action under specified conditions
sanction:     Specifies penalty for norm violation
```

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: the flip side of AI ingenuity." DeepMind Blog.

[5] Noether, E. (1918). "Invariante Variationsprobleme." Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, 235-257.

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

[7] Bond, A.H. (2025). "ErisML: A Formal Modeling Language for Foundation-Model-Enabled Agents." Working Paper.

[8] Shannon, C.E. (1948). "A Mathematical Theory of Communication." Bell System Technical Journal 27(3): 379-423.

[9] Carnot, S. (1824). Réflexions sur la puissance motrice du feu. Paris: Bachelier.