The Invariance Thesis:
Representational Symmetry as the Foundation of Aligned Intelligence and Controllable Systems

A Unified Theory from Ethics to Physics

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

December 2025

> One Ring to rule them all, One Ring to find them,
> One Ring to bring them all, and in the darkness bind them.
>
> —J.R.R. Tolkien

Abstract

We present the **Invariance Thesis**: the claim that a single mathematical principle—representational symmetry under structure-preserving transformations—underlies and unifies (1) the formal foundations of ethical reasoning, (2) the containment of superintelligent AI, (3) the gauge-theoretic structure of decision systems, and (4) the control of physical systems with hard constraints. We identify three deep principles that emerge from this unification: the *Grounding Principle* (evaluation must depend on physical observables, not descriptions), the *Canonicalization Principle* (equivalent representations must be identified before evaluation), and the *Conservation Principle* (representational symmetry implies conserved quantities via Noether's theorem). We show that these principles constitute a "meta-theory" that transcends the distinction between ethical and physical systems, revealing alignment, safety, and control as manifestations of a single geometric structure. The practical import is that formally verified constraint satisfaction—in AI systems, fusion reactors, medical devices, or any safety-critical domain—reduces to enforcing representational invariance within an appropriately defined principal bundle.

Contents

1 Introduction: The Search for Unity

1.1 The Problem Landscape

Several apparently distinct problems share a common structure:

1. **AI Alignment:** How do we ensure that AI systems pursue goals aligned with human values, even as they become more capable than their designers?

2. **AI Containment:** How do we prevent superintelligent systems from escaping constraints through superior reasoning about representations?

3. **Ethical Decision-Making:** How do we formalize moral reasoning in a way that is deterministic, verifiable, and immune to manipulation?

4. **Safety-Critical Control:** How do we guarantee that control systems satisfy hard constraints, not merely with high probability but with mathematical certainty?

5. **Cross-System Transfer:** How do we design controllers or ethical frameworks that transfer across contexts without requiring complete re-specification?

These problems arise in different domains—philosophy, computer science, robotics, plasma physics—yet they share a structural similarity that suggests a unified solution.

1.2 The Central Claim

This paper argues that all five problems are manifestations of a single underlying principle:

> **The Invariance Thesis:** Aligned, contained, ethical, and safe systems are precisely those whose evaluations are invariant under transformations that preserve the physically or morally relevant structure.

We call this the "Ruling Ring" because it is a single mathematical structure that governs all these domains, bringing them under a common theoretical framework.

1.3 Overview of the Unified Framework

The framework consists of five interlocking components:

Component	Mathematical Structure	Role
Stratified Geometric Ethics (SGE)	Stratified manifolds	Moral space topology
Bond Invariance Principle (BIP)	Gauge symmetry	Core invariance requirement
Gauge Theory of Decision	Principal bundles	Geometric framework
No Escape Theorem	Structural containment	Impossibility of evasion
Transfer via Dimensionless Form	Scale invariance	Cross-system applicability

The thesis is that these are not five separate theories but five perspectives on a single underlying structure.

2 The Three Deep Principles

Analyzing the unified framework reveals three foundational principles from which everything else derives.

2.1 The Grounding Principle

Principle 1 (Grounding)**.** *Evaluation of configurations, actions, or states must depend on physical observables measured by a declared sensor suite, not on descriptions, labels, or representations.*

Formal statement: Let $\Psi : \mathcal{X} \to \mathbb{R}^N$ be a grounding map taking representations to measurement vectors. An evaluation function $\Sigma : \mathcal{X} \to V$ satisfies the Grounding Principle if and only if there exists $\bar{\Sigma} : \mathbb{R}^N \to V$ such that:

$$\Sigma(x) = \bar{\Sigma}(\Psi(x))$$

Implications:

- *In ethics:* Moral judgments depend on who is actually harmed, who actually consented, what actually happened—not on how these are described.

- *In AI safety:* An AI cannot evade constraints by relabeling actions ("cardiac arrest" vs. "permanent bio-equilibrium")—the evaluation depends on vital signs, not vocabulary.

- *In control:* A controller's decisions depend on measured density, current, and temperature—not on whether we express them in SI or CGS units.

- *In physics:* Observable quantities are gauge-invariant; only they correspond to physical reality.

Why this matters: The Grounding Principle closes the gap between representation and reality. It ensures that evaluations track the territory, not the map.

2.2 The Canonicalization Principle

Principle 2 (Canonicalization)**.** *Before evaluation, all representations must be mapped to a canonical form that is unique up to the structure being preserved.*

Formal statement: There exists a canonicalizer $\kappa : \mathcal{X} \to \mathcal{X}$ such that:

(i) $\kappa(x) = \kappa(x')$ whenever x and x' represent the same grounded state

(ii) $\kappa(\kappa(x)) = \kappa(x)$ (idempotence)

(iii) $\Psi(\kappa(x)) = \Psi(x)$ (grounding preservation)

Implications:

- *In ethics:* Different descriptions of the same situation ("killed" vs. "caused to die") are mapped to the same canonical form before moral evaluation.

- *In AI safety:* The AI operates on canonicalized inputs; it never sees raw representations that could be manipulated.

- *In control:* Sensor readings are processed into a standard format; the controller is invariant to measurement conventions.

- *In physics:* Gauge fixing selects a unique representative from each equivalence class of gauge-equivalent configurations.

Why this matters: Canonicalization eliminates representational degrees of freedom before evaluation, making it impossible to influence outcomes through redescription.

2.3   The Conservation Principle

Principle 3 (Conservation). *Representational symmetry implies, via Noether's theorem, that "alignment" is a conserved quantity: systems that begin in an aligned state cannot drift out of alignment through their internal dynamics.*

Formal statement: For a Lagrangian L invariant under the structure group $\mathcal{G}$, the Noether current

$$J = \lambda\omega(\dot{\gamma})$$

is conserved along Euler-Lagrange trajectories. In particular, if $J(0) = 0$ (initial alignment), then $J(t) = 0$ for all t.

Implications:

- *In ethics:* A moral agent that begins reasoning from grounded premises cannot "drift" into representation-dependent conclusions.

- *In AI safety:* An AI system satisfying BIP cannot transition from aligned to misaligned through reasoning alone; misalignment requires external intervention.

- *In control:* A controller initialized on the canonical section remains on the canonical section; representational drift is dynamically forbidden.

- *In physics:* Gauge symmetry implies charge conservation; representations cannot create or destroy physical charge.

Why this matters: Conservation transforms alignment from a behavioral property (requiring constant enforcement) to a structural property (conserved by dynamics).

3   The Mathematical Unification

3.1   The Master Structure

All instances of the framework share a common geometric structure:

Definition 3.1 (Invariant System). An **invariant system** is a tuple $(\mathcal{P}, \mathcal{M}, \pi, \mathcal{G}, \Psi, \kappa, \Sigma, \mathcal{C})$ where:

- $\mathcal{P}$ is a smooth manifold (representation space)

- $\mathcal{M}$ is a smooth manifold (physical ground)

- $\pi : \mathcal{P} \to \mathcal{M}$ is a principal $\mathcal{G}$-bundle

- $\mathcal{G}$ is a Lie group (structure-preserving transformations)

- $\Psi : \mathcal{P} \to \mathbb{R}^N$ is the grounding map (factoring through π)

- $\kappa : \mathcal{P} \to \mathcal{P}$ is the canonicalizer (inducing a global section)

- $\Sigma : \mathcal{P} \to V$ is the evaluation function ($\mathcal{G}$-invariant)

- $\mathcal{C} \subseteq \mathcal{M}$ is the constraint set (defined by Ψ-inequalities)

Theorem 3.2 (Master Invariance Theorem). *For any invariant system:*

(a) *Σ factors through $\mathcal{M}$: $\Sigma = \bar{\Sigma} \circ \pi$*

(b) *The canonical connection is flat: $\Omega = 0$*

(c) The alignment current is conserved: $\frac{d}{dt}J = 0$

(d) Constraint satisfaction is preserved: $\gamma(0) \in \mathcal{C} \Rightarrow \gamma(t) \in \mathcal{C}_\varepsilon$

(e) Representational escape is impossible: $\Sigma(\kappa(x)) = \Sigma(\kappa(x'))$ *for all* x, x' *with* $\Psi(x) = \Psi(x')$

Proof. Combine the proofs of Theorems 4.1, 3.1, 6.2, 7.1, and 9.1 from the gauge theory paper. Each property follows from the structure group action, the canonicalizer properties, and the Noether theorem applied to the gauge-invariant Lagrangian. $\square$

3.2 Domain Instantiations

The master structure instantiates differently in each domain:

Component	Ethics	AI Safety	Control	Physics
$\mathcal{P}$	Descriptions	Input space	Representations	Field configs
$\mathcal{M}$	Bond structure	Grounded state	Physical state	Gauge orbits
$\mathcal{G}$	Bond-preserving	Canonicalizable	Unit/coord changes	Gauge group
Ψ	Physical facts	Sensor readings	Measurements	Observables
κ	Standard form	Input processing	Normalization	Gauge fixing
Σ	Moral evaluation	Safety score	Cost function	Action
$\mathcal{C}$	Forbidden acts	Safety constraints	Physical limits	—

3.3 Why the Same Structure Appears Everywhere

The appearance of identical structure across domains is not coincidental. It reflects a deep truth:

> *Whenever a system must make decisions based on physical reality rather than descriptions of reality, the mathematics of gauge symmetry naturally emerges.*

This is because:

1. **Reality is unique; descriptions are many.** Any physical state admits infinitely many equivalent descriptions (different coordinates, units, labels, orderings).

2. **Correct decisions must be description-independent.** Whether we're choosing moral actions, safe AI behaviors, or control inputs, the decision should depend on what's true, not on how we said it.

3. **Description-independence is gauge invariance.** Mathematically, requiring decisions to be independent of representational choice is requiring invariance under the group of representational transformations.

4. **Gauge invariance implies bundle structure.** The quotient of representations by the invariance group is a principal bundle; the invariance requirement makes functions "descend" to the base.

5. **Noether's theorem applies.** Any Lagrangian formulation of the decision process, if gauge-invariant, has a conserved current corresponding to the symmetry.

The "Ruling Ring" is not an arbitrary choice of framework—it is the *inevitable* mathematical structure for any system that must act on reality rather than representations.

4 The Stratified Geometry of Constraints

4.1 Why Stratification?

Real decision problems involve discontinuities that smooth manifolds cannot represent:

- **Threshold effects:** Consent at 0.99 vs. 1.00; Greenwald fraction at 0.84 vs. 0.86

- **Lexical priorities:** Rights violations dominate welfare considerations at any ratio

- **Incommensurable values:** No finite trade-off between life and property

- **Genuine dilemmas:** Both options involve irreducible moral loss

Theorem 4.1 (Stratified Spaces are Necessary). *Among standard geometric structures (smooth manifolds, manifolds with corners, cell complexes), only stratified spaces represent all four phenomena above.*

4.2 Stratified Principal Bundles

The master structure extends to stratified spaces:

Definition 4.2 (Stratified Invariant System). A **stratified invariant system** is an invariant system where $\mathcal{M}$ is a stratified space $(\mathcal{M}, \{M_i\}, \preceq)$ and:

1. Each stratum M_i is a smooth manifold

2. The bundle $\mathcal{P}|_{M_i} \to M_i$ is a principal $\mathcal{G}$-bundle

3. The canonicalizer κ restricts to each stratum

4. The evaluation Σ is smooth on each stratum, with controlled discontinuities at boundaries

Significance: Stratification allows the framework to handle moral discontinuities (in ethics), regime changes (in control), and phase transitions (in physics) while preserving the invariance structure on each smooth stratum.

4.3 Regime-Dependent Control as Stratification

In plasma control, the operating regime (L-mode, H-mode, ELMy H-mode) determines which physics applies. This is naturally modeled by stratification:

- Each regime is a stratum $M_i \subset \mathcal{M}$

- Within each stratum, smooth control laws apply

- Regime transitions are stratum boundary crossings

- Different grounding tensors may be relevant in different regimes

The BIP requirement applies within each stratum and constrains how control laws can change across boundaries.

5 The No Escape Theorem as Geometric Necessity

5.1 Reformulation

The No Escape Theorem is not merely a conditional result about AI containment—it is a geometric necessity within the invariant system framework.

Theorem 5.1 (No Escape as Geometry). *In an invariant system $(\mathcal{P}, \mathcal{M}, \pi, \mathcal{G}, \Psi, \kappa, \Sigma, \mathcal{C})$:*

1. ***Representational escape is geometrically impossible:*** *The fibers of π are the orbits of $\mathcal{G}$; Σ is constant on fibers; no manipulation of the fiber direction changes Σ.*

2. ***Semantic evasion is blocked by grounding:*** *Σ depends only on Ψ-values; relabeling that preserves Ψ is a gauge transformation that leaves Σ invariant.*

3. ***Alignment drift is forbidden by conservation:*** *The Noether current J is conserved; if $J(0) = 0$, then $J(t) = 0$ forever.*

4. ***Constraint violation requires leaving the admissible base:*** *$\mathcal{C} \subset \mathcal{M}$ is defined by Ψ-inequalities; violations are properties of base points, not of representations.*

5.2 What Remains: Residual Risks

The No Escape Theorem localizes risk to specific failure modes:

Risk	Nature	Mitigation
Ψ-inadequacy	Missing relevant observables	Domain expertise, governance
Implementation error	TCB bugs	Formal verification
Sensor spoofing	Physical attack	Hardware security
Governance capture	Political attack	Institutional design

None of these are "cognitive escape"—they are engineering, security, or political problems. The mathematical structure is sound; the residual risks concern whether it is correctly instantiated.

6 Dimensional Invariance and Transfer

6.1 Units as Gauge Transformations

The scaling group $\mathcal{S} = (\mathbb{R}_{>0})^d$ (for d independent physical dimensions) acts on measurements by unit rescaling. This is a subgroup of $\mathcal{G}$.

Theorem 6.1 (Dimensional Invariance). *If Σ is formulated in terms of dimensionless quantities Ψ^* satisfying:*
$$\Psi^*(\lambda \cdot x) = \Psi^*(x) \quad \forall \lambda \in \mathcal{S}$$

then Σ is unit-independent and transfers across systems with different scales.

6.2 Cross-System Transfer

In ethics: Moral principles formulated in terms of dimensionless ratios (proportionality, equality) transfer across contexts with different absolute scales.

In AI safety: Safety constraints formulated in terms of normalized quantities transfer across systems with different capability levels.

In control: Controllers formulated in terms of dimensionless parameters (Greenwald fraction, normalized beta) transfer across tokamaks with different sizes.

In physics: Physical laws formulated in dimensionless form (Reynolds number, Froude number) exhibit similarity and scaling.

6.3 The Universality of Dimensionless Physics

The transfer property reflects a deep physical truth: nature does not have preferred units. Any system that respects this fact inherits transfer properties automatically.

The "Ruling Ring" thus explains why dimensionless formulations work: they are manifestations of the full gauge invariance that the framework requires.

7 From Ethics to Physics and Back

7.1 The Ethics-Physics Bridge

The framework reveals a deep connection between ethical and physical reasoning:

Ethical Concept	Physical Analog
Bonds (moral relationships)	Constraints (physical laws)
Bond-preserving transformations	Gauge transformations
Moral evaluation	Energy/action functional
Ethical principles	Conservation laws
Moral dilemmas	Phase transitions
Incommensurable values	Singularities

This is not mere analogy. The mathematical structures are *identical*. A principal bundle with connection is a principal bundle with connection, whether the base manifold represents moral configurations or plasma states.

7.2 Why the Bridge Exists

The bridge exists because both ethics and physics are about:

1. **Reality** (not descriptions of reality)

2. **Constraints** (what must hold, not what we prefer)

3. **Invariance** (conclusions independent of representation)

4. **Conservation** (some quantities are preserved)

Any domain with these features will exhibit the same mathematical structure. The "Ruling Ring" is not a theory of ethics or a theory of physics—it is a theory of *reality-grounded invariant systems*, which ethics and physics both instantiate.

7.3 Implications for Artificial Intelligence

For AI systems:

1. **Alignment is geometric:** An aligned AI is one whose evaluation function satisfies gauge invariance. This is checkable, not interpretable.

2. **Containment is structural:** A contained AI is one whose outputs are elements of a well-defined quotient space. Escape requires violating mathematical definitions.

3. **Safety is conserved:** A safe AI, initialized correctly, remains safe by Noether's theorem. Misalignment requires external perturbation.

4. **Transfer is automatic:** An AI formulated in dimensionless terms transfers across contexts without retraining.

8 The Practical Program

8.1 Implementation Requirements

To instantiate the framework in a real system:

1. **Define the grounding:** Specify the sensor suite $\mathcal{S}$ and grounding map Ψ. This is the critical design choice—it determines what the system "sees."

2. **Identify the structure group:** Determine which transformations are representation-preserving (unit changes, relabeling, reordering, etc.).

3. **Construct the canonicalizer:** Build a verified procedure κ that maps equivalent representations to identical canonical forms.

4. **Define the evaluation:** Construct Σ as a function of Ψ-values only. Encode constraints as regions where $\Sigma = -\infty$.

5. **Implement the audit:** Build verification that confirms $\Sigma(\kappa(x))$ was correctly computed and that the output respects constraints.

6. **Verify the pipeline:** Formally verify that the canonicalizer, grounding, and evaluator satisfy the required properties.

8.2 Verification Conditions

The system is correct if:

1. $\kappa(x) = \kappa(x')$ whenever $\Psi(x) = \Psi(x')$ (canonicalization respects grounding)

2. $\Sigma(x) = \Sigma(\kappa(x))$ for all x (evaluation is post-canonicalization)

3. $\Sigma(x) = -\infty$ for all x with $\Psi(x) \notin \mathcal{C}$ (constraints are encoded)

4. The implementation matches the specification (formal verification)

Under these conditions, the Master Invariance Theorem guarantees safety.

8.3 Applications

The framework applies to:

- **AI Ethics:** DEME 2.0 architecture for real-time ethical governance

- **AI Safety:** Structural containment for advanced AI systems

- **Fusion Control:** BIP-compliant density and stability control for tokamaks

- **Medical Devices:** Insulin pumps, anesthesia systems with hard safety constraints

- **Autonomous Vehicles:** Collision avoidance with formal guarantees

- **Power Grids:** Frequency stabilization with provable constraint satisfaction

- **Chemical Processes:** Reactor control preventing runaway conditions

Each domain instantiates the same abstract structure with domain-specific grounding, constraints, and evaluation.

9 Philosophical Implications

9.1 The Nature of Normativity

The framework suggests a view of normativity:

> *Normative facts are facts about structure-preserving relationships (bonds). Normative reasoning is reasoning that respects the invariance structure of these bonds.*

This is neither pure realism (norms as natural facts) nor pure constructivism (norms as arbitrary conventions). It is *structural realism about normativity*: bonds are real relationships; invariance under bond-preserving transformations is the criterion of correctness.

9.2 The Tractability of Ethics

If ethics has the structure of a gauge theory, then:

1. Ethical reasoning is *computable* (reduce to finite approximations)

2. Ethical claims are *verifiable* (check gauge invariance)

3. Ethical systems are *transferable* (dimensionless formulation)

4. Ethical safety is *conserved* (Noether's theorem)

Ethics is not an intractable domain requiring infinite wisdom. It is a structured domain admitting formal methods—provided we accept that ethical evaluation must be grounded in physical observables and invariant under representation changes.

9.3 The Unity of Reason

The framework suggests that practical reason (ethics) and theoretical reason (physics) share a common structure:

- Both are grounded in physical reality

- Both require invariance under appropriate transformations

- Both admit conservation laws

- Both are formalizable and verifiable

The "Ruling Ring" is thus a contribution to the philosophy of normativity: it proposes that the structure of correct reasoning is the same whether we are reasoning about what to do or about what is the case.

10 Limitations and Open Problems

10.1 The Grounding Problem

The framework assumes an adequate grounding Ψ is given. But:

- How do we know if Ψ is adequate?

- Who decides what observables are morally/physically relevant?

- What if we discover Ψ was incomplete?

These are governance questions, not mathematical questions. The framework makes them explicit and tractable but does not solve them.

10.2 The Stratification Problem

Stratified spaces handle discontinuities, but:

- How do we know the correct stratification?

- What happens at stratum boundaries?

- Can stratification be learned from data?

10.3 The Dynamics Problem

The Noether analysis assumes a Lagrangian formulation. But:

- What is the "natural" Lagrangian for ethical dynamics?

- Are there non-Lagrangian systems where conservation fails?

- How does discrete-time implementation affect conservation?

10.4 The Implementation Problem

Formal verification is required, but:

- Current verification tools have limits

- Real systems have analog components not fully formalizable

- The gap between formal model and physical implementation persists

11 Conclusion: The Ruling Ring

11.1 Summary of the Thesis

We have argued that:

1. A single mathematical structure—the principal bundle with grounding, canonicalization, and gauge-invariant evaluation—underlies ethical reasoning, AI containment, control theory, and physical law.

2. Three deep principles (Grounding, Canonicalization, Conservation) generate the entire framework.

3. The No Escape Theorem, the representation theorems, the conservation laws, and the transfer properties are all manifestations of this unified structure.

4. Practical implementation requires specifying the grounding, constructing the canonicalizer, encoding constraints, and formally verifying the pipeline.

5. Residual risks (grounding adequacy, implementation correctness, physical security, governance integrity) are explicit and manageable—no longer hidden in unstructured appeals to "alignment."

11.2   The Name

We call this the "Ruling Ring" because:

- It is *one* mathematical structure (principal bundle with flat connection)

- That *rules* the correctness of all evaluations (via gauge invariance)

- That *finds* the invariant content in any representation (via canonicalization)

- That *brings together* ethics, AI, control, and physics (as domain instantiations)

- That *binds* systems to respect constraints (via structural containment)

Unlike Tolkien's Ring, this one is not malevolent. It is the mathematical structure that makes trustworthy AI, safe control, and rigorous ethics possible.

11.3   The Vision

The Invariance Thesis suggests a research program:

1. **Theoretical:** Develop the mathematics of stratified principal bundles, non-abelian structure groups, and curved connections.

2. **Computational:** Build verified implementations of canonicalizers, groundings, and evaluators.

3. **Empirical:** Validate the framework on real systems (tokamaks, medical devices, AI systems).

4. **Philosophical:** Explore the implications for metaethics, philosophy of science, and the unity of reason.

5. **Governance:** Develop institutions for grounding selection, verification auditing, and update control.

If this program succeeds, the alignment problem—for AI, for control systems, for ethical reasoning—will be transformed from an intractable philosophical puzzle to a solved engineering discipline.

The Ring has been forged. The question is whether we have the wisdom to use it.

A Relationship to Prior Work

A.1 Papers in the Framework

1. **Stratified Geometric Ethics (SGE):** Establishes stratified manifolds as the geometry of moral space; proves representation theorems for satisfaction operators.

2. **No Escape:** Proves that structural containment blocks cognitive escape routes; identifies residual risks.

3. **Gauge Theory of Invariant Control:** Constructs the principal bundle; proves BIP = gauge invariance; derives Noether conservation.

4. **BIP Fusion Theory:** Applies the framework to tokamak plasma control; demonstrates transfer via dimensionless formulation.

5. **DEME 2.0:** Implements the framework in a real-time ethical governance architecture.

This paper (the "Ruling Ring") synthesizes all five, identifying the common structure and the three deep principles.

A.2 Connections to External Literature

- **Gauge theory in physics:** The bundle structure and Noether analysis are standard differential geometry.

- **Geometric ethics:** Extends prior work (Pettigrew et al.) from smooth manifolds to stratified spaces.

- **Fairness as invariance:** Shares the invariance intuition with Dwork et al. on individual fairness.

- **Formal verification in AI:** Complements temporal logic approaches with geometric methods.

- **Control theory:** Extends classical control (PID, MPC, RL) with formal constraint guarantees.

B   Glossary of Key Terms

Term	Definition
Bond	A morally or physically relevant relationship between entities
BIP	Bond Invariance Principle: evaluations are invariant under bond-preserving transformations
Canonicalizer	A map that selects a unique representative from each equivalence class
Flat connection	A connection with zero curvature; implies no holonomy
Gauge invariance	Independence of evaluations from choice of representation
Grounding	Anchoring evaluations to physical observables
Invariant system	A system satisfying the three deep principles
No Escape	Theorem that structural containment blocks representational evasion
Noether current	Conserved quantity arising from continuous symmetry
Principal bundle	Fiber bundle where fibers are group orbits
SGE	Stratified Geometric Ethics
Stratified space	Union of smooth manifolds of varying dimensions
Structure group	Group of transformations preserving relevant structure