

Foundations of Invariant Evaluation: A Conservative Framework for Representation-Independent Reasoning

Grounding, Canonicalization, and Invariance as
Design Principles for Trustworthy Systems

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

December 2025

Abstract

We develop a framework for evaluation systems—whether ethical, safety-critical, or physical—that must produce consistent outputs regardless of how inputs are represented. We do *not* claim to solve the alignment problem, prove metaphysical theses, or provide guarantees independent of implementation. Rather, we identify *necessary conditions* for representation-independence and show these conditions have the mathematical structure of gauge theory. Specifically: (1) evaluation must factor through a declared grounding in observables; (2) equivalent representations must be identified via canonicalization; (3) evaluation must be invariant under the resulting equivalence relation. When a system satisfies these conditions and the Lagrangian formulation applies, Noether’s theorem implies conservation of a quantity we call the “alignment current.” The framework’s guarantees are *conditional* on adequate grounding, correct implementation, and applicability of the continuous symmetry assumptions. We make explicit what the framework does and does not establish.

Contents

1	Introduction	3
1.1	The Problem	3
1.2	Scope and Limitations	3
1.3	Structure of the Paper	3
2	Motivation: Limitations of Purely Formal Approaches	3
2.1	Observation 1: Formal Systems Do Not Determine Their Own Interpretation . .	4
2.2	Observation 2: Gödel’s Theorems Limit Self-Certification	4
2.3	Observation 3: Multiple Logics Are Internally Coherent	4
2.4	The Upshot	4
3	The Formal Framework	4
3.1	Basic Structures	5
3.2	The Grounding Axiom	5
3.3	The Equivalence Relation	5
3.4	The Canonicalization Axiom	6

3.5	The Transformation Group	6
3.6	The Invariance Theorem	6
4	The Geometric Structure	6
4.1	Bundle Structure	6
4.2	Connection and Curvature	7
5	Conservation Results	7
5.1	The Lagrangian Setting	7
5.2	Noether's Theorem	8
5.3	Conservation of Alignment	8
6	Applications	8
6.1	Application to Ethics: The Bond Invariance Principle	8
6.2	Application to Control Systems	9
6.3	Application to AI Safety	9
7	Limitations and Non-Claims	10
7.1	The Framework Does Not Solve the Grounding Problem	10
7.2	The Framework Does Not Guarantee Implementation Correctness	10
7.3	Conservation Requires Continuous Symmetry and Lagrangian Dynamics	10
7.4	The Framework Does Not Establish Metaphysical Claims	11
7.5	Summary of Claims	11
8	Conclusion	11
8.1	What We Have Established	11
8.2	What Remains Open	11
8.3	The Conservative Claim	12
A	Proof of Theorem 4.2	12
B	Relationship to Physics Literature	12
C	Glossary	13

1 Introduction

1.1 The Problem

Consider systems that must evaluate configurations, actions, or states:

- An AI system evaluating whether an action is safe
- A control system evaluating whether a state satisfies constraints
- An ethical reasoning system evaluating whether an action is permissible

A common failure mode: the evaluation depends not only on the *situation* but also on how the situation is *represented*. Changing labels, reordering options, or reformulating descriptions changes the output, even when the underlying situation is unchanged.

Definition 1.1 (Representation-Dependence). An evaluation function $\Sigma : \mathcal{X} \rightarrow V$ is **representation-dependent** if there exist $x, x' \in \mathcal{X}$ such that:

1. x and x' represent the same underlying situation (by some criterion)
2. $\Sigma(x) \neq \Sigma(x')$

Goal: Identify conditions under which evaluation is *representation-independent*.

1.2 Scope and Limitations

Scope Limitation: This paper provides *sufficient conditions* for representation-independence within a specified formal framework. It does *not*:

- Prove that these conditions are the *only* way to achieve representation-independence
- Guarantee correct behavior if the grounding is inadequate or implementation is faulty
- Establish metaphysical claims about the nature of reality, logic, or causation
- Solve the “hard problem” of choosing which features are morally or physically relevant

The framework tells you: *if* you can specify adequate grounding and *if* you implement correctly, *then* certain guarantees follow.

1.3 Structure of the Paper

Section 2 motivates the framework by examining why purely formal approaches face limitations. Section 3 presents the core definitions and axioms. Section 4 develops the mathematical structure. Section 5 states and proves the main results. Section 6 discusses applications. Section 7 addresses limitations explicitly.

2 Motivation: Limitations of Purely Formal Approaches

We present three observations that motivate the need for grounding. These are not metaphysical claims but observations about the structure of formal systems.

2.1 Observation 1: Formal Systems Do Not Determine Their Own Interpretation

Claim 1 (Interpretation Underdetermination). A formal system F consisting of symbols and inference rules does not, by itself, determine what its symbols refer to or whether its theorems are true of any particular domain.

Support: This is a standard result in mathematical logic. Any consistent first-order theory with infinite models has models of every infinite cardinality (Löwenheim-Skolem). The same axioms can be satisfied by vastly different structures. The formal system alone does not pick out the “intended” model.

Consequence: Derivability within a formal system ($F \vdash \phi$) does not establish truth in any particular domain without an additional specification of how the system is interpreted.

Scope Limitation: We are not claiming formal systems are useless or that logic is arbitrary. We are claiming that formal derivability alone is insufficient to establish claims about specific domains without interpretation.

2.2 Observation 2: Gödel’s Theorems Limit Self-Certification

Theorem 2.1 (Gödel’s Second Incompleteness Theorem). *If F is a consistent formal system capable of expressing basic arithmetic, then F cannot prove its own consistency.*

Consequence for our purposes: A formal system cannot, from within, certify that its axioms are consistent or that its theorems correspond to truths about an intended domain. External verification or grounding is required.

Scope Limitation: Gödel’s theorems apply to sufficiently powerful formal systems (those capable of expressing arithmetic). They do not apply to all formal systems. We cite them to illustrate a general point: formal systems have inherent limitations on self-certification.

2.3 Observation 3: Multiple Logics Are Internally Coherent

Classical logic, intuitionistic logic, relevant logic, and various modal logics are all internally coherent formal systems. They differ in their theorems (e.g., excluded middle is a theorem of classical but not intuitionistic logic).

Consequence: There is no purely formal argument establishing that one logic is the “correct” logic for all domains. The choice of logic is, in part, a choice of tool for a purpose.

Scope Limitation: We are not endorsing global logical relativism. Classical logic may well be correct for most purposes. The point is that the choice cannot be made on purely formal grounds; it requires some criterion external to the formalisms being compared.

2.4 The Upshot

These observations suggest that trustworthy evaluation requires something beyond formal derivability: a specification of *what the formal system is about* and *how it connects to the domain of application*. We call this specification **grounding**.

3 The Formal Framework

We now present the framework precisely. All claims in this section are definitions or axioms, not empirical claims about the world.

3.1 Basic Structures

Definition 3.1 (Representation Space). A **representation space** is a set \mathcal{X} whose elements are representations (descriptions, encodings, formulations) of situations in some domain of interest.

Definition 3.2 (Value Space). A **value space** is a set V (typically \mathbb{R} , $\{0, 1\}$, or a partial order) representing possible evaluation outputs.

Definition 3.3 (Evaluation Function). An **evaluation function** is a map $\Sigma : \mathcal{X} \rightarrow V$ assigning evaluations to representations.

3.2 The Grounding Axiom

The key idea: evaluation should depend on *what is measured*, not on *how it is described*.

Axiom 1 (Grounding). There exists:

- (a) A **measurement space** $\mathcal{M} \subseteq \mathbb{R}^N$ for some $N \in \mathbb{N}$
- (b) A **grounding map** $\Psi : \mathcal{X} \rightarrow \mathcal{M}$
- (c) A **grounded evaluation** $\bar{\Sigma} : \mathcal{M} \rightarrow V$

such that for all $x \in \mathcal{X}$:

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

Interpretation:

- \mathcal{M} represents measurement outcomes (what sensors report)
- Ψ extracts measurements from representations
- $\bar{\Sigma}$ operates on measurements, not descriptions
- The axiom requires that Σ factors through Ψ

Remark 3.4 (What Grounding Does Not Guarantee). The Grounding Axiom does not guarantee that Ψ captures all relevant features. If Ψ omits morally or physically relevant aspects of the situation, evaluation may be “grounded” but still inadequate. The adequacy of Ψ is a domain-specific question outside the scope of the formal framework.

3.3 The Equivalence Relation

Grounding induces an equivalence relation on representations.

Definition 3.5 (Grounding Equivalence). Two representations $x, x' \in \mathcal{X}$ are **grounding-equivalent**, written $x \sim_{\Psi} x'$, if and only if $\Psi(x) = \Psi(x')$.

Proposition 3.6 (Invariance Under Equivalence). *If Axiom 1 holds, then Σ is constant on equivalence classes:*

$$x \sim_{\Psi} x' \implies \Sigma(x) = \Sigma(x')$$

Proof. $x \sim_{\Psi} x'$ means $\Psi(x) = \Psi(x')$. By Axiom 1, $\Sigma(x) = \bar{\Sigma}(\Psi(x)) = \bar{\Sigma}(\Psi(x')) = \Sigma(x')$. \square

3.4 The Canonicalization Axiom

A canonicalizer selects a unique representative from each equivalence class.

Axiom 2 (Canonicalization). There exists a **canonicalizer** $\kappa : \mathcal{X} \rightarrow \mathcal{X}$ satisfying:

- (a) **Idempotence:** $\kappa(\kappa(x)) = \kappa(x)$ for all $x \in \mathcal{X}$
- (b) **Equivalence-class collapse:** $x \sim_\Psi x' \implies \kappa(x) = \kappa(x')$
- (c) **Grounding preservation:** $\Psi(\kappa(x)) = \Psi(x)$ for all $x \in \mathcal{X}$

Proposition 3.7 (Canonicalization Yields Representative). *Under Axiom 2, $\kappa(x)$ is a well-defined representative of the equivalence class $[x]_\Psi$.*

Proof. By (b), κ maps all elements of $[x]_\Psi$ to the same element. By (c), this element is in $[x]_\Psi$. By (a), it is a fixed point of κ . \square

3.5 The Transformation Group

Definition 3.8 (Structure-Preserving Transformations). The **structure group** \mathcal{G} is the set of bijections $g : \mathcal{X} \rightarrow \mathcal{X}$ that preserve grounding:

$$\mathcal{G} := \{g : \mathcal{X} \rightarrow \mathcal{X} \mid g \text{ is a bijection and } \Psi(g(x)) = \Psi(x) \text{ for all } x\}$$

Proposition 3.9 (\mathcal{G} is a Group). *\mathcal{G} forms a group under composition.*

Proof. Identity: $\text{id} \in \mathcal{G}$ (trivially preserves Ψ). Closure: if $g, h \in \mathcal{G}$, then $\Psi(gh(x)) = \Psi(h(x)) = \Psi(x)$. Inverse: if $g \in \mathcal{G}$ and $y = g(x)$, then $\Psi(g^{-1}(y)) = \Psi(x) = \Psi(g(x)) = \Psi(y)$. \square

3.6 The Invariance Theorem

Theorem 3.10 (Invariance). *If Axiom 1 holds, then Σ is \mathcal{G} -invariant:*

$$\Sigma(g(x)) = \Sigma(x) \quad \forall g \in \mathcal{G}, x \in \mathcal{X}$$

Proof. Let $g \in \mathcal{G}$. By definition, $\Psi(g(x)) = \Psi(x)$. By Axiom 1, $\Sigma(g(x)) = \bar{\Sigma}(\Psi(g(x))) = \bar{\Sigma}(\Psi(x)) = \Sigma(x)$. \square

Remark 3.11 (Converse). The converse is also true under mild conditions: if Σ is \mathcal{G} -invariant and \mathcal{G} acts transitively on Ψ -equivalence classes, then Σ factors through Ψ .

4 The Geometric Structure

We now show that the framework has the structure of a principal fiber bundle. This section assumes familiarity with differential geometry; readers primarily interested in applications may skip to Section 5.

4.1 Bundle Structure

Definition 4.1 (Principal Bundle Conditions). The quadruple $(\mathcal{X}, \mathcal{M}, \Psi, \mathcal{G})$ satisfies **principal bundle conditions** if:

- (PB1) \mathcal{G} acts freely on \mathcal{X} : $g(x) = x \implies g = \text{id}$
- (PB2) \mathcal{G} acts transitively on fibers: $\Psi(x) = \Psi(x') \implies \exists g \in \mathcal{G} : g(x) = x'$
- (PB3) $\Psi : \mathcal{X} \rightarrow \mathcal{M}$ is surjective

(PB4) (Smoothness, if applicable): \mathcal{X}, \mathcal{M} are smooth manifolds; \mathcal{G} is a Lie group; Ψ is a smooth submersion

Theorem 4.2 (Bundle Structure). *Under conditions (PB1)–(PB3), $\Psi : \mathcal{X} \rightarrow \mathcal{M}$ is a principal \mathcal{G} -bundle in the topological sense. Under (PB1)–(PB4), it is a smooth principal bundle.*

Proof. (PB1) gives freeness. (PB2) implies orbits equal fibers. (PB3) gives surjectivity. The quotient $\mathcal{X}/\mathcal{G} \cong \mathcal{M}$ follows. For the smooth case, (PB4) plus standard results on Lie group actions give local triviality. \square

Scope Limitation: The bundle structure requires (PB1)–(PB2), which are substantive assumptions. Not all grounding setups satisfy them. When they fail, the geometric picture is only partially applicable.

4.2 Connection and Curvature

Definition 4.3 (Connection via Section). A global section $\sigma : \mathcal{M} \rightarrow \mathcal{X}$ satisfying $\Psi \circ \sigma = \text{id}_{\mathcal{M}}$ determines a connection on the bundle.

Theorem 4.4 (Canonicalization Gives Section). *Under Axioms 1 and 2, if the bundle conditions (PB1)–(PB2) hold, then κ induces a global section $\sigma : \mathcal{M} \rightarrow \mathcal{X}$ defined by:*

$$\sigma(m) := \kappa(x) \quad \text{for any } x \in \Psi^{-1}(m)$$

Proof. By Axiom 2(b), $\kappa(x) = \kappa(x')$ whenever $\Psi(x) = \Psi(x')$, so σ is well-defined. By Axiom 2(c), $\Psi(\sigma(m)) = \Psi(\kappa(x)) = \Psi(x) = m$, so $\Psi \circ \sigma = \text{id}_{\mathcal{M}}$. \square

Theorem 4.5 (Flatness). *The connection determined by a global section has zero curvature.*

Proof. A global section trivializes the bundle: $\mathcal{X} \cong \mathcal{M} \times \mathcal{G}$ via $(m, g) \mapsto \sigma(m) \cdot g$. In a trivial bundle, the canonical flat connection has curvature $\Omega = d\omega + \frac{1}{2}[\omega, \omega] = 0$ (Maurer-Cartan equation on the \mathcal{G} factor). \square

Remark 4.6 (Interpretation of Flatness). Flatness means parallel transport is path-independent: representations can be “moved” between fibers without holonomy. Physically, this reflects that representational choices have no intrinsic structure—they are “pure gauge.”

5 Conservation Results

This section shows that, under additional assumptions, invariance implies conservation. The results require a Lagrangian formulation, which may not apply in all contexts.

5.1 The Lagrangian Setting

Definition 5.1 (Decision Lagrangian). Suppose \mathcal{X} and \mathcal{M} are smooth manifolds. A **decision Lagrangian** is a function $L : T\mathcal{X} \rightarrow \mathbb{R}$ of the form:

$$L(x, \dot{x}) = K(\dot{x}) - U(\Psi(x)) - \frac{\lambda}{2} \|\omega(\dot{x})\|_{\mathfrak{g}}^2 \quad (1)$$

where:

- K is a kinetic term (cost of deliberation)
- $U : \mathcal{M} \rightarrow \mathbb{R}$ is a potential (negative of grounded evaluation)
- ω is the connection form

- $\lambda > 0$ is a coupling constant
- $\|\cdot\|_{\mathfrak{g}}$ is an \mathcal{G} -invariant norm on the Lie algebra \mathfrak{g}

Scope Limitation: The Lagrangian formulation assumes continuous dynamics on smooth manifolds. It applies to systems whose evolution can be described by differential equations. Discrete systems, non-smooth dynamics, and systems without natural “kinetic” interpretations require separate analysis.

5.2 Noether’s Theorem

Theorem 5.2 (Noether). *Let $L : TQ \rightarrow \mathbb{R}$ be a Lagrangian on configuration space Q . If a one-parameter group $\{\phi_s\}_{s \in \mathbb{R}}$ satisfies $L \circ d\phi_s = L$ (symmetry), then the quantity*

$$Q_\xi := \left\langle \frac{\partial L}{\partial \dot{q}}, \xi \right\rangle$$

is constant along solutions of the Euler-Lagrange equations, where $\xi = \frac{d}{ds}|_{s=0}\phi_s(q)$ is the infinitesimal generator.

Proof. Standard. See Arnold, *Mathematical Methods of Classical Mechanics*, Chapter 4. □

5.3 Conservation of Alignment

Theorem 5.3 (Alignment Current). *For the Lagrangian (1), the gauge symmetry $x \mapsto x \cdot g$ implies a conserved current:*

$$J := \lambda \omega(\dot{x}) \in \mathfrak{g}$$

satisfying $\frac{d}{dt}J = 0$ along Euler-Lagrange trajectories.

Proof. The Lagrangian is \mathcal{G} -invariant by construction: K depends only on velocity magnitude; U depends only on $\Psi(x)$ which is \mathcal{G} -invariant; $\|\omega(\dot{x})\|$ is \mathcal{G} -invariant. Apply Theorem 5.2 to the one-parameter subgroups of \mathcal{G} . □

Corollary 5.4 (Horizontal Trajectories are Preserved). *If $\omega(\dot{x}(0)) = 0$ (trajectory starts horizontal), then $\omega(\dot{x}(t)) = 0$ for all t .*

Proof. $J(t) = \lambda \omega(\dot{x}(t)) = J(0) = 0$ by conservation. □

Remark 5.5 (Interpretation). “Horizontal” means the trajectory has no component in the fiber direction—it moves only in the base (physical) directions. Corollary 5.4 says: *a system that starts responding only to physical states, not to representational artifacts, continues to do so.*

Scope Limitation: Conservation requires: (1) the Lagrangian formulation applies, (2) the symmetry is continuous (Lie group), (3) dynamics follow Euler-Lagrange equations. For discrete symmetry groups (e.g., permutations) or non-Lagrangian dynamics, conservation in this form does not apply. The invariance results (Section 4) still hold; only the dynamical conservation fails.

6 Applications

6.1 Application to Ethics: The Bond Invariance Principle

Definition 6.1 (Bond). A **bond** is a morally relevant relationship between entities, represented as a tuple $b = (a, p, r, c)$ where a is an agent, p is a patient, r is a relation type, and c is context.

Definition 6.2 (Bond Structure). The **bond structure** $B(x)$ of a representation x is the set of all bonds encoded in x .

Instantiation of the framework:

- $\Psi(x) := B(x)$ (grounding in bond structure)
- $\mathcal{G} :=$ bond-preserving transformations (relabeling, reordering that preserve B)
- $\Sigma :=$ moral evaluation function

Corollary 6.3 (Bond Invariance Principle). *If moral evaluation satisfies Axiom 1 with $\Psi = B$, then:*

$$B(x) = B(x') \implies \Sigma(x) = \Sigma(x')$$

Moral judgments are invariant under bond-preserving transformations.

Scope Limitation: This does not tell us *which* relationships are morally relevant (the content of B). It says: *whatever you count as morally relevant, evaluation should depend only on that, not on representational choices.* The choice of B is a substantive ethical question outside the formal framework.

6.2 Application to Control Systems

Instantiation for plasma control:

- $\Psi(x) := (\bar{n}, I_p, \beta_N, \dots)$ (measured plasma parameters)
- $\mathcal{G} :=$ unit changes, coordinate transformations
- $\Sigma :=$ control objective / safety score

Corollary 6.4 (Unit Invariance). *If a controller satisfies Axiom 1, its outputs are independent of unit conventions.*

Corollary 6.5 (Dimensionless Transfer). *If Ψ consists of dimensionless quantities, controllers transfer across systems with different physical scales.*

6.3 Application to AI Safety

Instantiation for AI containment:

- $\Psi(x) :=$ sensor readings, physical observables
- $\mathcal{G} :=$ representational manipulations (relabeling, redescription)
- $\Sigma :=$ safety/alignment score

Theorem 6.6 (Representation Attack Immunity). *If an AI system's evaluation satisfies Axiom 1, then for any Ψ -preserving manipulation $g \in \mathcal{G}$:*

$$\Sigma(g(x)) = \Sigma(x)$$

The manipulation does not change the evaluation.

Proof. Immediate from Theorem 3.10. □

Scope Limitation: Theorem 6.6 blocks *representational* attacks (relabeling, redescription). It does *not* block:

- Physical attacks (sensor spoofing)
- Attacks that change Ψ -values (actual physical manipulation)
- Attacks on the implementation of κ , Ψ , or Σ
- Inadequacy of Ψ (missing relevant features)

The theorem’s scope is precisely: *given correct implementation and adequate grounding, representational evasion fails.*

7 Limitations and Non-Claims

We explicitly state what the framework does *not* establish.

7.1 The Framework Does Not Solve the Grounding Problem

Claim 2 (Grounding is Assumed, Not Derived). The framework assumes a grounding map Ψ is given. It provides no procedure for:

- Determining which features are morally/physically relevant
- Verifying that Ψ is adequate for the domain
- Discovering that Ψ is missing important features

Consequence: A system can satisfy all formal requirements of the framework while being inadequately grounded. The framework makes explicit *what must be specified* (Ψ) and *what properties follow if it is specified correctly*, but the correctness of Ψ itself is a domain-specific judgment.

7.2 The Framework Does Not Guarantee Implementation Correctness

Claim 3 (Implementation is Assumed, Not Verified). The framework’s guarantees are conditional on:

- Correct implementation of κ , Ψ , and Σ
- Correct implementation of the trusted computing base
- Physical security of sensors and actuators

Consequence: Bugs, hardware failures, or security breaches can violate the guarantees. The framework is a design specification, not a correctness proof for any particular implementation.

7.3 Conservation Requires Continuous Symmetry and Lagrangian Dynamics

Claim 4 (Conservation Has Preconditions). The conservation results (Theorem 5.3, Corollary 5.4) require:

- \mathcal{G} is a Lie group (continuous symmetry)
- Dynamics are governed by a Lagrangian
- The system evolves according to Euler-Lagrange equations

For discrete \mathcal{G} (e.g., permutation groups) or non-Lagrangian dynamics, conservation in this form does not hold.

Consequence: The invariance results (Theorem 3.10) are more general than the conservation results. Invariance holds whenever Axiom 1 holds; conservation requires additional structure.

7.4 The Framework Does Not Establish Metaphysical Claims

Claim 5 (No Metaphysical Commitments). The framework does not establish:

- That reality has any particular structure
- That logic, causation, or modality work in any particular way
- That grounding in observables is the only way to achieve trustworthiness

The framework is a *conditional*: if you want representation-independence, and if you can specify adequate grounding, then these mathematical structures provide it. The framework is compatible with many metaphysical views.

7.5 Summary of Claims

Claim	Established?	Requires
Grounding \Rightarrow Invariance	Yes	Axiom 1
Bundle structure exists	Yes	(PB1)–(PB3)
Canonicalization gives section	Yes	Axioms 1, 2, (PB1)–(PB2)
Connection is flat	Yes	Global section exists
Alignment current is conserved	Yes	Lagrangian, Lie group \mathcal{G} , E-L dynamics
Ψ is adequate	No	Domain expertise
Implementation is correct	No	Verification
All attacks are blocked	No	Only representational attacks

8 Conclusion

8.1 What We Have Established

1. **Definition:** Representation-independence means evaluation factors through a grounding map.
2. **Sufficient conditions:** Axioms 1 and 2 suffice for representation-independence.
3. **Mathematical structure:** Under principal bundle conditions, the framework has gauge-theoretic structure with flat connection.
4. **Conservation:** Under Lagrangian dynamics with continuous symmetry, alignment is conserved.
5. **Applications:** The framework instantiates in ethics (BIP), control (unit invariance), and AI safety (representation attack immunity).

8.2 What Remains Open

1. **Grounding selection:** How to choose Ψ for a given domain.
2. **Adequacy verification:** How to determine if Ψ is adequate.
3. **Implementation:** How to correctly implement κ , Ψ , Σ .
4. **Discrete symmetries:** Conservation-like results for non-Lie groups.
5. **Non-Lagrangian systems:** Dynamical guarantees without variational structure.

8.3 The Conservative Claim

We make a conservative claim:

If you can specify a grounding map Ψ that captures the relevant features of your domain, **and if** you correctly implement canonicalization and grounded evaluation, **then** your evaluation will be representation-independent in the precise sense that \mathcal{G} -transformations do not change outputs.

Under additional assumptions (Lagrangian dynamics, continuous symmetry), alignment is conserved: systems that start responding only to grounded features continue to do so.

The framework does not tell you what to ground in, does not verify your implementation, and does not block non-representational attacks. It provides structure, not magic.

A Proof of Theorem 4.2

Proof. We verify the principal bundle axioms.

Free action: (PB1) states $g(x) = x \implies g = \text{id}$, which is the definition of free action.

Transitive action on fibers: (PB2) states that for any x, x' with $\Psi(x) = \Psi(x')$, there exists g with $g(x) = x'$. This means \mathcal{G} acts transitively on each fiber $\Psi^{-1}(m)$.

Orbits equal fibers: If $x' = g(x)$ for some $g \in \mathcal{G}$, then $\Psi(x') = \Psi(g(x)) = \Psi(x)$ (since g preserves Ψ). So orbits are contained in fibers. By (PB2), fibers are contained in orbits. Hence orbits equal fibers.

Base space: $\mathcal{M} = \mathcal{X}/\mathcal{G}$ follows from orbits equaling fibers. (PB3) ensures Ψ is surjective.

Local triviality (smooth case): Under (PB4), Ψ is a smooth submersion with \mathcal{G} a Lie group acting freely and properly. Standard results (e.g., Lee, *Introduction to Smooth Manifolds*, Theorem 21.10) give local triviality. \square

B Relationship to Physics Literature

The gauge-theoretic structure we describe is standard in differential geometry and mathematical physics:

- **Principal bundles:** Kobayashi & Nomizu, *Foundations of Differential Geometry*
- **Connections and curvature:** Nakahara, *Geometry, Topology and Physics*
- **Noether's theorem:** Arnold, *Mathematical Methods of Classical Mechanics*
- **Gauge theory in physics:** Bleecker, *Gauge Theory and Variational Principles*

Our contribution is not to the mathematics of gauge theory but to its application as a framework for representation-independent evaluation.

C Glossary

Term	Definition
Grounding map Ψ	Function extracting measurements from representations
Canonicalizer κ	Function selecting unique representative per equivalence class
Structure group \mathcal{G}	Group of Ψ -preserving transformations
Grounding equivalence \sim_Ψ	$x \sim_\Psi x'$ iff $\Psi(x) = \Psi(x')$
Principal bundle	Fiber bundle where fibers are group orbits
Connection	Specification of “horizontal” directions in bundle
Flat connection	Connection with zero curvature
Noether current	Conserved quantity from continuous symmetry
BIP	Bond Invariance Principle: moral invariance under bond-preserving transformations