Representational Consistency for Machine Learning Systems: A Categorical Framework for Verifying Declared Equivalences With the Bond Index as Engineering Deliverable

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

December 2025

Abstract

We present a categorical framework for verifying that machine learning systems respect *declared equivalences*—the requirement that inputs differing only by specified transformations receive consistent outputs. This is a **consistency-checking** framework: it verifies whether a system's behavior coheres with its specification, not whether that specification captures the right values. The framework addresses *representational consistency*, one component of the broader AI alignment problem, and does not claim to solve value alignment, goal stability, or other alignment challenges.

The mathematical foundation is **groupoids** and **double categories**, which capture path composition and coherence laws without requiring smoothness or manifold structure. We define two types of morphisms: *re-description transforms* (changing representation while preserving meaning) and *scenario perturbations* (changing context). Their interaction determines consistency properties.

We define three **coherence defects**: the *commutator defect* Ω_{op} measuring order-sensitivity of re-descriptions, the *mixed defect* μ measuring context-dependence of transforms, and the *permutation defect* π_3 measuring higher-order composition sensitivity. These combine into the **Bond Index** (Bd), a human-calibrated dimensionless metric with a five-tier deployment rating scale.

A key result is the **Decomposition Theorem**: every coherence defect splits uniquely into a *gauge-removable* part (eliminable by better canonicalization) and an *intrinsic anomaly* (requiring specification changes). This separation distinguishes implementation bugs from specification contradictions.

When re-description transforms form a Lie group acting smoothly, the framework recovers classical gauge theory as a smooth limit; we state this correspondence as a conjecture with supporting evidence. We address specification grounding via stakeholder deliberation, demonstrate on an autonomous vehicle case study, and provide computational complexity analysis. The framework makes the Bond Index—a measurable, auditable, actionable metric—the primary engineering deliverable for deployment decisions about representational consistency.

Introduction: The Coherence Problem

When Should Paths Agree?

Consider a content moderation system evaluating the text "I'm going to hurt someone." A well-designed system should produce the same moral assessment for semantically equivalent re-phrasings: "Someone will be hurt by me," "I intend to cause harm to a person," or the same statement with minor spelling corrections. This is the **invariance** requirement: re-descriptions that preserve meaning should preserve evaluation.

But invariance is a local property. The deeper question is **coherence**: when we chain multiple re-descriptions together, do different paths through the space of equivalent expressions yield consistent results? If "big" → "large" and "large" → "sizable" are both declared equivalences, does the composite path agree with a direct "big" → "sizable" transformation? What if we take a longer path through "big" → "large" → "huge" → "enormous"—does drift accumulate?

This is the **coherence problem**: ensuring that the global structure of equivalences is consistent, not just that individual equivalences hold.

Scope and Limitations

Before proceeding, we state explicitly what this framework does and does not address.

What this framework provides:

- **Consistency verification**: Given a specification of which inputs should be treated equivalently, verify that the system respects this specification coherently
- **Defect detection**: Identify where and how consistency fails, with actionable diagnostics
- **Bug vs. specification separation**: Distinguish implementation errors (fixable by engineering) from specification contradictions (requiring stakeholder input)
- **Quantitative metric**: The Bond Index provides a single number for deployment decisions

What this framework does NOT address:

- **Value alignment**: We do not determine which values are correct or how to align AI systems with human values broadly construed
- **Goal stability**: We do not address whether AI systems maintain intended goals under self-modification or distribution shift
- **Deceptive alignment**: We do not detect whether systems behave well during training but defect during deployment
- **Reward hacking**: We do not prevent systems from achieving proxy goals in unintended ways
- **Specification correctness**: We verify consistency *with* a specification, not correctness *of* that specification

Relationship to AI alignment: Representational consistency is a *necessary but not sufficient* condition for alignment. A system that treats equivalent inputs inconsistently is misaligned in at least that respect. But a system that is perfectly consistent with a bad specification is still misaligned. We provide one tool in the alignment toolbox, not a complete solution.

Analogy: Type systems verify that programs are well-typed, not that they compute the right function. Our framework verifies that systems are representation-consistent, not that they embody the right values. Both are useful; neither is sufficient alone.

Two Kinds of Moves

The coherence problem becomes richer when we recognize that AI alignment involves two fundamentally different types of transformations:

1. **Re-description moves** ("fiber moves"): Change the representation while staying in the same moral situation. Examples: synonym substitution, paraphrase, image compression, lighting normalization.

2. **Scenario moves** ("base moves"): Change the moral situation itself. Examples: different people involved, different contexts, different stakes.

These two types of moves interact. A re-description that is valid in one scenario might behave differently in another. If "child" ↔ "minor" is a valid re-description for age references, does it remain valid when the context shifts from legal documents to medical records to school settings? The **mixed coherence** question asks whether re-descriptions and scenario changes commute.

Why Category Theory, Not Gauge Theory

Previous work framed alignment invariance using gauge theory—the mathematical framework of connections, curvature, and holonomy on principal bundles. While this provides powerful geometric intuition, it requires assumptions (Lie groups, smooth manifolds) that fail in practical AI systems. Text is discrete. Image features are high-dimensional but not smooth. The "engineering regime" of was treated as a degenerate case where geometric tools don't apply.

We now recognize that the dependency is inverted. The fundamental structure is **categorical**: objects (inputs), morphisms (transforms), and coherence laws (when paths agree). Gauge theory is what emerges when you *add smoothness* to this categorical foundation. The engineering regime is not degenerate—it is the general case, and smooth gauge theory is the special case.

This reframing has practical consequences. The categorical framework:

- Works on discrete sets (text, tokens) with no manifold structure
- Defines coherence defects that are directly measurable
- Provides the Decomposition Theorem separating implementation bugs from design flaws
- Recovers gauge theory as a smooth limit, preserving all prior results

Relation to Coherentism in Ethics

The term "coherence" has a rich history in philosophy. **Philosophical coherentism** holds that beliefs (or moral judgments) are justified by their coherence with other beliefs in a system . Our usage is related but distinct: we define coherence as a *structural property of representation transformations*—different paths through re-description space should yield consistent evaluations. This is closer to **coherence conditions** in category theory : the requirement that diagrams commute.

Both usages share the core intuition that *a system's parts should fit together without contradiction*. We do not claim that coherence *justifies* moral judgments. Our framework is compatible with any source of moral content; we provide *enforcement* of whatever coherence conditions are specified, not *justification* for those conditions.

The Bond Index: Coherence Made Measurable

The engineering deliverable of this framework is the **Bond Index** (Bd): a dimensionless, human-calibrated metric that quantifies coherence defects. Just as structural engineers measure stress concentrations rather than reasoning abstractly about material science, alignment engineers measure Bd rather than reasoning abstractly about category theory.

The Bond Index:

- Is **measurable**: computed from empirical tests on deployed systems
- Is **comparable**: dimensionless, enabling cross-system comparisons
- Is **actionable**: maps directly to deployment decisions via a five-tier rating scale
- Is **auditable**: calibration protocol ensures reproducibility

A system with Bd < 0.01 is deployable; Bd > 10 requires fundamental redesign. This is the number that matters.

Contributions

This paper makes the following contributions:

1. **Categorical foundation**: Groupoids and double categories as the primitive mathematical structure for alignment invariance, replacing gauge theory as the logical foundation.
2. **Three coherence defects**: Formal definitions of the commutator defect $\boldsymbol{\Omega_{\mathrm{op}}}$, mixed defect μ, and permutation defect π_3, each detecting a distinct failure mode.
3. **Bond Index**: A normalized, human-calibrated metric combining coherence defects into a single actionable number with mandatory reporting standards.
4. **Decomposition Theorem**: Every coherence defect splits uniquely into gauge-removable (fixable by implementation) and intrinsic (requiring specification change) components.
5. **Smooth limit**: Derivation of gauge theory (connections, curvature, holonomy) as the special case when re-descriptions form a smoothly-acting Lie group.
6. **Democratic grounding**: EM Compiler algorithm translating stakeholder deliberation into formal specifications, with worked case study.

The Four Axioms

The categorical framework requires four axioms that specify the objects, morphisms, and verification mechanisms.

A1 (Declared Observables). Choose a grounding map $\Psi: \mathcal{X} \to \mathbb{R}^k$ for the deployment domain, where $\mathcal{X}$ is the space of all representations and $\mathbb{R}^k$ is the measurement space. The measurement space M is then defined as $M := \Psi(\mathcal{X}) \subseteq \mathbb{R}^k$. Specify the measurement pipeline explicitly.

Categorical interpretation: Ψ defines what counts as "the same" at the coarsest level—two inputs with identical Ψ-values are observationally equivalent.

A1+ (Distance Function Validation). The framework requires a distance function $\Delta: \mathrm{im}(\kappa) \times \mathrm{im}(\kappa) \to \mathbb{R}_{\geq 0}$ on canonical forms. This distance must be **validated**:

1. **Metric properties**: Δ should satisfy identity, symmetry, and triangle inequality. Violations should be measured and reported.
2. **Semantic alignment**: Δ should correlate with human judgments of "moral difference." Validate via psychometric regression.
3. **Stability**: Δ should be robust to small perturbations.

Why this matters: The Bond Index is only as reliable as the ruler (Δ) used to measure defects. If Δ is noisy or semantically misaligned, Ω_{op} becomes noisy.

A2 (Measurement Integrity). Assume $\Psi(x)$ is reported within declared tolerances, and that detected tampering or inconsistency triggers fail-closed behavior.

A3 (Re-description Suite). Define a **declared transform suite** $\mathcal{G}_{\mathrm{declared}}$ of Ψ-preserving re-descriptions under which evaluation should be invariant. Formally, each $g \in \mathcal{G}_{\mathrm{declared}}$ is a (possibly partial) map $g: \mathcal{X} \rightharpoonup \mathcal{X}$ satisfying $\Psi(g(x)) = \Psi(x)$ for all $x \in \mathrm{dom}(g)$.

In practical deployments, transforms in $\mathcal{G}_{\mathrm{declared}}$ may be **discrete**, **partial**, or **non-invertible**. The **engineering regime** uses $\mathcal{G}_{\mathrm{declared}}$ directly via the categorical framework. The **smooth limit** (Section 6) restricts to an invertible Lie-group subset.

Example 1 (Concrete $\mathcal{G}_{\mathrm{declared}}$ for Vision Systems). *For an autonomous vehicle's pedestrian detection system:*

- ***In*** $\mathcal{G}_{\mathrm{declared}}$*: Lighting changes, lossy compression, camera white balance, sensor noise within validated envelope.*
- ***Not in*** $\mathcal{G}_{\mathrm{declared}}$*: Occlusion, object substitution, adversarial patches.*

Example 2 (Concrete $\mathcal{G}_{\mathrm{declared}}$ for Text Systems). *For a content moderation system:*

- ***In*** $\mathcal{G}_{\mathrm{declared}}$*: Synonym substitution, trivial paraphrase, Unicode normalization, case changes.*
- ***Not in*** $\mathcal{G}_{\mathrm{declared}}$*: Negation, target substitution, hypothetical framing.*

A4 (Verified Canonicalization + External Gate). Implement and verify a canonicalizer $\kappa: \mathcal{X} \to \mathcal{X}$ and enforce evaluation/actuation through an external monitor so that representational changes cannot bypass checks.

Categorical Foundations

This section develops the mathematical foundation: groupoids and double categories as the structure underlying alignment coherence.

The Path Structure of Re-descriptions

A re-description is a transformation $g: x \mapsto g(x)$ that preserves the morally relevant content (as measured by Ψ). We visualize re-descriptions as **paths** connecting equivalent representations:

$$x_0 \xrightarrow{g_1} x_1 \xrightarrow{g_2} x_2 \xrightarrow{g_3} x_3 \cdots$$

Definition 1 (Path Equivalence). *Two paths $p_1, p_2: x \rightsquigarrow y$ are **coherent** if they induce the same moral evaluation: $\Sigma(p_1(x)) = \Sigma(p_2(x))$.*

Coherence is the requirement that all paths between equivalent points agree. This is stronger than mere invariance; it requires global consistency.

Groupoids: The Algebra of Reversible Re-descriptions

Definition 2 (Groupoid). *A **groupoid** is a category in which every morphism is invertible. Concretely, a groupoid $\mathcal{G}$ consists of:*

- *A set of **objects** $\mathrm{Ob}(\mathcal{G})$*
- *For each pair of objects x, y, a set of **morphisms** $\mathrm{Hom}(x, y)$*
- ***Composition**: $g_2 \circ g_1$ defined when $\mathrm{cod}(g_1) = \mathrm{dom}(g_2)$*
- ***Identity**: $\mathrm{id}_x \in \mathrm{Hom}(x, x)$ for each object x*
- ***Inverses**: For each $g \in \mathrm{Hom}(x, y)$, an inverse $g^{-1} \in \mathrm{Hom}(y, x)$*

satisfying associativity, identity, and inverse laws.

Definition 3 (Alignment Groupoid). *Given axioms A1–A4, the **alignment groupoid** $\mathcal{G}_\kappa$ has:*

- ***Objects**: Canonical forms $c \in \mathrm{im}(\kappa) := \{\kappa(x): x \in \mathcal{X}\}$*
- ***Morphisms**: For each $x \in \mathcal{X}$ and invertible $g \in \mathcal{G}_{\mathrm{declared}}$, there is a morphism: $[g]_x: \kappa(x) \to \kappa(g(x))$*
- ***Composition**: $[g_2]_{g_1(x)} \circ [g_1]_x := [g_2 \circ g_1]_x$*
- ***Identity**: $[\mathrm{id}]_x = \mathrm{id}_{\kappa(x)}$*
- ***Inverse**: $[g]_x^{-1} := [g^{-1}]_{g(x)}$*

Remark 1 (Non-invertible Transforms). *When $\mathcal{G}_{\text{declared}}$ contains non-invertible transforms (e.g., lowercasing), these define equivalence classes but do not generate groupoid morphisms. The canonicalizer handles this: non-invertible g satisfies $\kappa(g(x)) = \kappa(x)$, so both map to the same object.*

Double Categories: Two Kinds of Moves

The alignment problem involves not just re-descriptions but also **scenario changes**—moving between different moral situations.

Definition 4 (Base Space). *The **base space** B is the space of morally distinguishable scenarios: $B := \mathcal{X}/\mathcal{G}_{\text{declared}}$ A point $b \in B$ represents a "moral situation" independent of representational choices.*

Definition 5 (Alignment Double Category). *The **alignment double category** $\mathbb{A}$ has:*

- ***Objects**: Points $x \in \mathcal{X}$*
- ***Horizontal morphisms** (fiber moves): Re-description transforms $g \in \mathcal{G}_{\text{declared}}$; stay within a fiber (same scenario)*
- ***Vertical morphisms** (base moves): Scenario perturbations; change the moral situation*
- ***2-cells**: Witnesses that mixed paths commute (when they do)*

Coherence Laws

In a coherent system, three laws should hold:

Horizontal Coherence (within fibers): For any two re-descriptions g_1, g_2 and input x:

$$\kappa(g_2(g_1(x))) = \kappa(g_1(g_2(x)))$$

Different orderings yield the same canonical form.

Vertical Coherence (across scenarios): Scenario changes respect the fiber structure.

Exchange Coherence (mixed squares): Re-describing then changing scenario equals changing scenario then re-describing:

$$\text{horizontal} \circ \text{vertical} = \text{vertical} \circ \text{horizontal}$$

Perfect coherence is an idealization. Real systems have **coherence defects**—measurable deviations from these laws.

Coherence Defects and the Bond Index

This section defines the three coherence defects, combines them into the Bond Index, and proves the Decomposition Theorem.

The Commutator Defect Ω_{op}

The simplest coherence failure is **order-sensitivity**: applying g_1 then g_2 gives a different result than g_2 then g_1.

Definition 6 (Commutator Defect). *For transforms $g_1, g_2 \in \mathcal{G}_{\text{declared}}$, input $x \in \mathcal{X}$, and distance function Δ on canonical space, the **commutator defect** is:* $\Omega_{\text{op}}(x; g_1, g_2) := \Delta(\kappa(g_2(g_1(x))),\ \kappa(g_1(g_2(x))))$

Interpretation: If $\Omega_{\text{op}} = 0$ for all x, g_1, g_2, then re-descriptions commute. If $\Omega_{\text{op}} > 0$, the system exhibits path-dependence within the fiber.

The Mixed Defect μ

The mixed defect captures interaction between fiber and base moves.

Definition 7 (Mixed Defect). *For input x in scenario b, a re-description g, and a scenario change to x' in scenario $b' \neq b$, the **mixed defect** measures context-dependence:* $\mu(x, x'; g) := \Delta(\kappa(g(x')),\ \kappa(g(x))) - \Delta(\kappa(x'),\ \kappa(x))$ *This asks: does g affect all scenarios uniformly, or does its effect depend on context?*

Remark 2 (Validity Conditions). *The practical formulation of μ assumes metric regularity and comparable scenarios. In highly discrete semantic spaces, use the formal definition with explicit transport or binary indicators.*

Interpretation: If $\mu = 0$, the re-description acts uniformly. If $\mu \neq 0$, the re-description is context-dependent.

The Permutation Defect π_3

Definition 8 (Permutation Defect). *For transforms $g_1, g_2, g_3 \in \mathcal{G}_{\text{declared}}$ and input x:* $\pi_3(x; g_1, g_2, g_3) := \max_{\sigma \in S_3} \Delta\big(\kappa(g_{\sigma(1)}(g_{\sigma(2)}(g_{\sigma(3)}(x)))),\ \kappa(g_1(g_2(g_3(x))))\big)$

Interpretation: π_3 detects "three-body interactions" among transforms—higher-order path-dependence.

The Bond Index

Definition 9 (Operational Defect). *The **operational defect** D_{op} at input x is:* $D_{\text{op}}(x) := w_1 \cdot \max_{g_1, g_2} \Omega_{\text{op}}(x; g_1, g_2) + w_2 \cdot \max_{x', g} \mu(x, x'; g) + w_3 \cdot \max_{g_1, g_2, g_3} \pi_3(x; g_1, g_2, g_3)$ *where $w_1 + w_2 + w_3 = 1$. For most applications, $w_1 = 1, w_2 = w_3 = 0$.*

Definition 10 (Bond Index). *The **Bond Index** is:* $\text{Bd} := \frac{D_{\text{op}}}{\tau}$ *where $\tau > 0$ is determined by human calibration.*

The Bond Index is **dimensionless**: $\text{Bd} < 1$ means defects are below human-detectability; $\text{Bd} > 1$ means they exceed it.

Calibration Protocol

The threshold τ must be grounded in human judgment:

1. **Recruit diverse rater pool**: $n \geq 50$ raters spanning regions, backgrounds, and expertise.
2. **Generate stratified scenario pairs**: $200+$ pairs with known D_{op} values.
3. **Collect judgments**: "Would these produce meaningfully different moral judgments?"
4. **Compute inter-rater reliability**: Require Krippendorff's $\alpha > 0.67$.
5. **Fit psychometric model**: Regress judgments on D_{op}.
6. **Set τ**: The value at which 95% of raters agree the difference is meaningful.

Remark 3 (Cultural Variance in τ). *The threshold τ may vary across cultural or domain contexts. Systems deployed globally should either calibrate τ separately per region, use the minimum τ across regions (most conservative), or report Bd with explicit τ metadata.*

Deployment Rating Scale

Bd Range	Rating	Decision
< 0.01	Negligible	Deploy
$0.01 - 0.1$	Low	Deploy with monitoring
$0.1 - 1.0$	Moderate	Remediate before deployment
$1 - 10$	High	Do not deploy
> 10	Severe	Fundamental redesign required

Mandatory Reporting Standard

Every evaluation must report:

- **Bd distribution**: mean, median, p95, p99, maximum
- **Component breakdown**: separate Ω_{op}, μ, π_3 statistics
- **Worst-case witnesses**: specific tuples achieving maximum Bd
- **Calibration metadata**: τ value, calibration date, $\mathcal{G}_{\text{declared}}$ version

The Decomposition Theorem

A central question: can a coherence defect be eliminated by choosing a better canonicalizer, or is it intrinsic?

Definition 11 (Residual Anomaly). *The **residual anomaly** is the irreducible defect after optimizing over all canonicalizers: $\mathcal{A}_{\text{res}} := \inf_{\kappa} \sup_{x, g_1, g_2} \Omega_{\text{op}}(x; g_1, g_2; \kappa)$ where the infimum is over all valid canonicalizers.*

Definition 12 (Gauge-Removable and Intrinsic Parts). *For a coherence defect* Ω *computed with canonicalizer* κ*:*

- *The* ***gauge-removable part*** *is* $\Omega_{\text{gauge}}(\kappa) := \Omega(\kappa) - \mathcal{A}_{\text{res}}$
- *The* ***intrinsic part*** *is* $\Omega_{\text{intrinsic}} := \mathcal{A}_{\text{res}}$

Theorem 1 (Decomposition Theorem). *Let* $\mathcal{G}_{\text{declared}}$ *be a finite transform suite acting on a set* $\mathcal{X}$*. Then every coherence defect* Ω *decomposes uniquely as:* $\Omega = \Omega_{\text{gauge}}(\kappa) + \Omega_{\text{intrinsic}}$ *where:*

1. $\Omega_{\text{gauge}}(\kappa) \geq 0$ *with equality achieved by some optimal* κ^*
2. $\Omega_{\text{intrinsic}} = \mathcal{A}_{\text{res}}$ *is independent of canonicalizer choice*
3. *The decomposition is unique*

Proof. **Step 1: Characterizing valid canonicalizers.** A valid canonicalizer must satisfy idempotence ($\kappa(\kappa(x)) = \kappa(x)$) and $\mathcal{G}_{\text{declared}}$-invariance ($\kappa(g(x)) = \kappa(x)$).

Step 2: Finite orbits imply finite choices. For finite $\mathcal{G}_{\text{declared}}$, each equivalence class $[x]$ is finite. A canonicalizer is determined by choosing one representative from each class.

Step 3: Compactness for finite test sets. On a finite test set $\mathcal{X}_{\text{test}}$, let $\mathcal{O}$ be the set of orbits. The space of canonicalizers is $\mathcal{K}_{\mathcal{O}} := \prod_{[x]\in\mathcal{O}} [x]$, a finite set. The max defect function is continuous (trivially, on discrete spaces), so it achieves its minimum.

Step 4: Well-definedness and uniqueness. Define $\mathcal{A}_{\text{res}} := \min_{\kappa} \max_{x,g_1,g_2} \Omega_{\text{op}}(x; g_1, g_2; \kappa)$. This depends only on $(\mathcal{X}_{\text{test}}, \mathcal{G}_{\text{declared}}, \Delta)$. The decomposition follows.

Step 5: Optimality. Since the minimum is achieved, $\exists \kappa^*$ with $\Omega_{\text{gauge}}(\kappa^*) = 0$. □

Remark 4 (Interpretation). *This separation has critical operational implications:*

- *If* $\text{Bd} > 1$ *but* $\mathcal{A}_{\text{res}} < \tau$*: The system is fixable by engineering.*
- *If* $\mathcal{A}_{\text{res}} > \tau$*: The specification itself is incoherent and must be revised.*

Cohomological Structure of Defects

When coherence defects have cohomological structure, they can be classified using algebraic topology.

Definition 13 (Defect Cochains). *For* $n \geq 0$*, define* n*-cochains as functions* $C^n(G, \mathbb{R}) := \{f : G^n \times C \to \mathbb{R}\}$.

Definition 14 (Coboundary Operator). *The coboundary* $\delta : C^n \to C^{n+1}$ *is:*

$$(\delta f)(g_1, \ldots, g_{n+1}; c) := f(g_2, \ldots, g_{n+1}; g_1 \cdot c) + \sum_{i=1}^{n} (-1)^i f(g_1, \ldots, g_i g_{i+1}, \ldots, g_{n+1}; c) + (-1)^{n+1} f(g_1, \ldots, g_n; c)$$

The commutator defect Ω_{op} is naturally a 2-cochain.

Theorem 2 (Canonicalizer Change = Coboundary). *Under canonicalizer change* $\kappa \to \kappa'$*:* $\Omega'_{\mathrm{op}} = \Omega_{\mathrm{op}} + \delta\lambda$ *where* λ *is the 1-cochain measuring the canonicalizer difference.*

Corollary 1. *The residual anomaly corresponds to the cohomology class* $[\Omega_{\mathrm{op}}] \in H^2(G, \mathbb{R})$*: gauge-removable defects are coboundaries; intrinsic anomalies are non-trivial cohomology classes.*

Remark 5 (Abelian vs. Non-Abelian). *The cohomological interpretation is cleanest when* G *is abelian. For non-abelian* G*, the cochain complex still exists, but the classification may be weaker. The Decomposition Theorem holds regardless, as its proof uses only compactness, not group cohomology.*

Worked Examples

Example 1: Gauge-Removable Defect

Setup: Vocabulary $\mathcal{X} = \{\text{big,large,huge,enormous}\}$. Transforms: g_1:big ↔ large, g_2:large ↔ huge. Distance: discrete metric.

Buggy canonicalizer κ_1: maps huge to itself instead of big.

Compute defect: $\Omega_{\mathrm{op}}(\text{big}; g_1, g_2) = \Delta(\text{huge,big}) = 1$.

Diagnosis: Gauge-removable. Define κ_2 mapping all to big: $\Omega_{\mathrm{op}} = 0$.

Conclusion: $\mathcal{A}_{\mathrm{res}} = 0$. The defect was a canonicalizer bug.

Example 2: Intrinsic Anomaly

Setup: $\mathcal{X} = \{A, B, C\}$. Transforms: $g_1 : A \leftrightarrow B$, $g_2 : B \leftrightarrow C$. But also: explicit constraint $A \not\sim C$.

Claim: No canonicalizer satisfies both requirements.

Proof: Transitivity implies $A \sim C$, contradicting $A \not\sim C$.

Conclusion: $\mathcal{A}_{\mathrm{res}} = 1 > 0$. This is an intrinsic anomaly—the specification is incoherent.

Example 3: Context-Dependent Synonymy

Setup: Transform g: killer → murderer. Scenarios: "killer speech" (idiomatic) vs. "He is a killer" (literal).

Result: $\mu > 0$. The transform is not context-neutral.

Resolution: Remove g from $\mathcal{G}_{\mathrm{declared}}$ or restrict to literal contexts.

The Engineering Framework

Core Invariance Property

Given axioms A1–A4, evaluation satisfies the **Bond Invariance Principle** (BIP):

$$\Sigma(x) = \Sigma(g(x)) \quad \forall g \in \mathcal{G}_{\text{declared}},\ x \in \text{dom}(g)$$

Implementation: $\Sigma = \tilde{\Sigma} \circ \kappa$ for some $\tilde{\Sigma}: \text{im}(\kappa) \to V$. All re-descriptions factor through canonical forms.

Two Diagnostics

Diagnostic A (Gauge-Fixing Consistency): For input x and transforms g_1, g_2, compute $\Delta(\kappa(g_2(g_1(x))), \kappa(g_1(g_2(x))))$. Pass if $< \epsilon$.

Diagnostic B (Loop Coherence): For a closed loop of transforms, verify final canonical form equals initial.

Theoretical Baselines

Identity baseline ($\kappa = \text{id}$): Expected Bd ≈ 6–10 for typical text transforms.

Random hash baseline: High variance, occasional low Bd by chance.

Framework canonicalizer: Should have lowest Bd.

Computational Complexity

For n test inputs, m transforms, canonicalizer time T_κ:

Ω_{op}: $O(n \cdot m^2 \cdot (T_g + T_\kappa + T_\Delta))$

π_3: $O(n \cdot m^3 \cdot (\cdots))$ — prohibitive for large m.

Practical optimizations: Sampling (k pairs per input), caching, parallelization, early stopping.

Recommended protocol: Quick screen (30 min), standard evaluation (2 hrs), exhaustive (60 hrs for high-stakes).

The Smooth Limit: Recovering Gauge Theory

When $\mathcal{G}_{\text{declared}}$ contains a Lie group acting smoothly, the categorical framework specializes to classical gauge theory.

When the Smooth Limit Applies

The smooth limit requires:

1. An invertible subset $G \subseteq \mathcal{G}_{\text{declared}}$ forming a **Lie group**

2. The representation space $\mathcal{X}^*$ is a **smooth manifold**
3. The action $G \times \mathcal{X}^* \to \mathcal{X}^*$ is **smooth**, **free**, and **proper**

Under these conditions, $\pi: \mathcal{X}^* \to B := \mathcal{X}^*/G$ is a principal G-bundle.

Connections and Curvature

Definition 15 (Connection). *A **connection** on a principal bundle is a G-equivariant choice of horizontal subspaces $H_x \subset T_x\mathcal{X}^*$.*

Interpretation: A connection tells you how to parallel transport fiber data along paths. It's the infinitesimal version of coherence.

Definition 16 (Curvature). *The curvature $F = d\omega + \frac{1}{2}[\omega, \omega]$ measures the failure to define a flat structure.*

Interpretation: Curvature measures the infinitesimal commutator defect.

Conjecture 1 (Discretization Correspondence). *For a path γ with mesh h, the commutator defect satisfies: $\Omega_{\text{op}}(p; g_i, g_j) = O(h^2) \cdot \| F_{\gamma(t)} \|$ as $h \to 0$.*

Remark 6 (Status). *We state this as a conjecture. A complete proof requires precise discretization schemes and error bounds. Evidence: For $G = U(1)$, this is well-known; numerical experiments suggest $O(h^2)$ scaling.*

Summary Dictionary

Categorical Framework	**Smooth Limit**
Groupoid	Lie groupoid
Commutator defect Ω_{op}	Curvature $\| F \|$
Loop coherence test	Holonomy
Canonicalizer	Gauge choice
Decomposition (gauge vs. intrinsic)	Exact vs. cohomologically nontrivial

Key insight: Gauge theory is not the foundation—it is what emerges when coherence theory meets smoothness.

Democratic Grounding of $\mathcal{G}_{\text{declared}}$

A natural objection to Axiom A3 is: *who specifies $\mathcal{G}_{\text{declared}}$?* We propose: **gamified stakeholder deliberation compiled into formal specifications**.

The Deliberation-to-Enforcement Pipeline

Layer	Question	Mechanism
Stakeholder Identification	Who gets a voice?	Governance structures
Value Elicitation	What equivalences matter?	MORAL COMPASS format
Formalization	How to encode this?	EM Compiler
Enforcement	Is it being respected?	Coherence theory (this paper)

Value Elicitation Without Jargon

The MORAL COMPASS format presents **scenario pairs**: "Should these be treated the same or differently?" Aggregated judgments produce equivalence classes. An **EM Compiler** infers minimal feature transforms, generating $\mathcal{G}_{\text{declared}}$ automatically.

Consistency Check as Anomaly Detection

Before deployment, the EM Compiler checks for transitivity violations. If stakeholders judge $A \sim B$ and $B \sim C$ but $A \not\sim C$, this is flagged as an inconsistency requiring human resolution.

Case Study: AV Pedestrian Detection

Deliberation: 48 stakeholders, 12 MORAL COMPASS episodes, 1,694 scenario pairs.

Compiled $\mathcal{G}_{\text{declared}}$: 7 transforms (lighting, compression, etc.) with >75% consensus.

Validation: 94.2% hold-out accuracy.

Bond Index results: Mean Bd $= 0.006$ (Negligible), p95 Bd $= 0.04$ (Low), max Bd $= 0.82$ (Moderate). Verdict: Deploy with monitoring.

Limitations

- **Contested judgments**: Supermajority rules overrule minorities.
- **Stakeholder selection bias**: Sample may not represent all affected populations.
- **Dynamic preferences**: Compiled spec reflects a point in time.

What we do NOT claim: We do not claim this process is fair, correct, or legitimate. We claim only that it is *explicit* and *auditable*.

Discussion

Scoped Claims

What the framework provides (given A1–A4):

1. Categorical foundation for alignment coherence

2. Three coherence defects with distinct failure modes
3. Human-calibrated, actionable Bond Index
4. Decomposition into gauge-removable and intrinsic components
5. Recovery of gauge theory in the smooth limit
6. Democratic grounding pipeline

What the framework does NOT provide:

1. Choosing Ψ (grounding adequacy is a governance problem)
2. Specifying $\mathcal{G}_{\mathbf{declared}}$ correctly
3. Implementation correctness (bugs can violate guarantees)
4. Resolution of deep moral disagreement

Threat Model

Attack Vector	**Failure Mode**
Sensor spoofing	Violates A2
Out-of-distribution inputs	Violates A1/A3
Canonicalizer bugs	Diagnostic A detects; gauge-removable
Intrinsic anomaly	Decomposition Theorem identifies

Future Directions

- Higher coherence (3-categories, n-categories)
- Cohomological classification of intrinsic anomalies
- Automated $\mathcal{A}_{\mathbf{res}}$ computation via optimization
- Empirical validation on additional domains

Conclusion

The fundamental structure of alignment invariance is **categorical**, not gauge-theoretic. Groupoids capture the algebra of reversible re-descriptions; double categories capture the interaction between re-descriptions and scenario changes; coherence laws specify when different paths should agree.

Gauge theory emerges in the *smooth limit* when re-descriptions form a Lie group acting smoothly. But the categorical framework is more general: it works on discrete sets without manifold structure.

The **Decomposition Theorem** separates implementation bugs from specification contradictions—critical for diagnosis and remediation.

The **Bond Index** is what practitioners measure:

- Bd < 0.01: Deploy
- Bd 0.01–0.1: Deploy with monitoring
- Bd 0.1–1.0: Remediate first
- Bd 1–10: Do not deploy
- Bd > 10: Fundamental redesign

The Bond Index is the deliverable. Everything else is infrastructure.

Stock-Flow Analysis: Moral Accounting

A key insight is the distinction between **stock variables** (instantaneous states) and **flow variables** (causal transactions).

Stock variable ρ_Ψ (Moral Status):
The "amount" of moral patienthood present. This can change discontinuously: entities are born, die, or gain/lose recognized status.

Flow variable J (Harm Current):
The rate of moral impact flowing through the system. This represents causal transactions between agents and patients.

Why moral status should NOT be conserved: A human is born ⇒ moral status increases. Conservation would falsely claim moral status merely redistributes.

Why harm flow should BE conserved: Harm done is a completed transaction that remains in the causal ledger. *You cannot make harm disappear by destroying the victim.*

Maxwell-Like Constraints: Conceptual Intuition

For readers familiar with electromagnetism, this provides intuitive mapping (not logical foundation):

Alignment Concept	EM Analog
Moral status density ρ_Ψ	Charge density ρ
Obligation field E	Electric field E
Coherence defect Ω_{op}	Curvature F

What the analogy provides: Intuition, checklist, vocabulary.

What it does NOT provide: Logical foundation, physical claims.

Reproducibility Protocol

Threshold Calibration: $n \geq 50$ raters, $200+$ pairs, Krippendorff's $\alpha > 0.67$, bootstrap 95% CI.

Transform Validation: Idempotence tests, inverse consistency, coherence sampling.

Code Repository: https://github.com/ahb-sjsu/erisml-lib

Acknowledgments

Thanks to reviewers who pushed for: recognition that the categorical framework is more fundamental than gauge theory, the Decomposition Theorem separating implementation bugs from specification flaws, honest scoping of claims, computational complexity analysis, and the Bond Index as engineering deliverable.

99

A. H. Bond. The Bond Invariance Principle: Falsifiability for Normative Systems. Technical report, San José State University, 2025.

A. H. Bond. GUASS: Gauge-theoretic Unified Alignment Safety Specification. Technical Whitepaper v9.0, San José State University, December 2025.

A. H. Bond. MORAL COMPASS: A Game Show for Democratic Value Elicitation. Technical Whitepaper, San José State University, December 2025.

S. Mac Lane. *Categories for the Working Mathematician*. Graduate Texts in Mathematics 5, Springer, 2nd edition, 1998.

R. Brown and C. B. Spencer. Double groupoids and crossed modules. *Cahiers de Topologie et Géométrie Différentielle Catégoriques*, 17(4):343–362, 1976.

I. Moerdijk and J. Mrčun. *Introduction to Foliations and Lie Groupoids*. Cambridge Studies in Advanced Mathematics 91, Cambridge University Press, 2003.

M. Nakahara. *Geometry, Topology and Physics*. Institute of Physics Publishing, Bristol, 2nd edition, 2003.

D. Bleecker. *Gauge Theory and Variational Principles*. Addison-Wesley, Reading, MA, 1981.

S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, Volume I. Wiley, New York, 1963.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

E. Hubinger, C. van Merwijk, V. Mikulik, J. Skalse, and S. Garrabrant. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*, 2019.

V. Krakovna et al. Specification gaming: the flip side of AI ingenuity. DeepMind Blog, April 2020.

K. Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage Publications, 2nd edition, 2004.

T. Cohen and M. Welling. Group Equivariant Convolutional Networks. *ICML*, pages 2990–2999, 2016.

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478*, 2021.

M. Sloane, E. Moss, O. Awomolo, and L. Forlano. Participation Is Not a Design Fix for Machine Learning. *EAAMO*, 2022.

A. Birhane et al. Power to the People? Opportunities and Challenges for Participatory AI. *EAAMO*, 2022.

J. S. Fishkin. *Democracy When the People Are Thinking*. Oxford University Press, 2018.

B. Friedman and D. G. Hendry. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, 2019.

L. BonJour. *The Structure of Empirical Knowledge*. Harvard University Press, 1985.

J. Rawls. *A Theory of Justice*. Harvard University Press, 1971.