Noether's Theorem for Ethics: Harm Conservation and the Unified Structure of Normative and Physical Reality

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

December 27, 2025

Abstract

We demonstrate that the mathematical structure underlying gauge theory in physics and coherence verification in ethics is identical, and that this identity has profound consequences. Applying Noether's theorem to the re-description symmetry required by ethical consistency, we derive a conservation law: **harm is conserved**. Just as electric charge cannot be destroyed by destroying the charged particle—only moved or transformed—harm cannot be eliminated by eliminating the harmed party. This is not metaphor but mathematical identity: the same theorem, the same structure, the same conservation law in different domains. We develop the full analogy between electrodynamics and ethics, identifying harm density $\rho_{\mathcal{H}}$ with charge density, harm current $J_{\mathcal{H}}$ with electric current, and deriving the ethical continuity equation $\partial\rho_{\mathcal{H}}/\partial t + \nabla \cdot J_{\mathcal{H}} = 0$. We explore the implications for consequentialism, retributive justice, restorative justice, and the foundations of moral philosophy. We argue that this unification suggests a deeper mathematical structure—possibly rooted in information geometry or homotopy type theory—underlying both physical and normative reality. Throughout, we maintain epistemic humility: we claim not that this structure reveals metaphysical necessity, but that it provides a powerful, empirically grounded tool for understanding the formal constraints on coherent ethical reasoning. The implications for AI alignment are immediate: systems that violate harm conservation are incoherent in a mathematically precise sense.

Contents

1 Introduction: The Suspicious Coincidence

In 2025, while developing a categorical framework for verifying representational consistency in AI systems, we noticed something strange. The mathematics we were using—groupoids, double categories, connections, curvature—was identical to the mathematics of gauge theory in physics. Not merely analogous. *Identical.*

At first, this seemed like a coincidence, or perhaps an artifact of using powerful mathematical tools that appear in many contexts. But as we developed the framework further, the correspondences became too precise to ignore:

Gauge Theory (Physics)	**Bond Framework (Ethics)**
Fiber bundle	Space of representations
Connection ω	Transport across representations
Curvature $F = d\omega + \omega \wedge \omega$	Coherence defect Ω
Gauge transformation	Re-description transform
Gauge invariance	Bond Invariance Principle
Parallel transport	Transform composition

This paper investigates why this correspondence exists and what it implies.

Our central result is this: the correspondence is not a coincidence. Both gauge theory and the Bond Framework are instances of a deeper structure, and applying the foundational theorem of that structure—Noether's theorem—to ethics yields a conservation law:

Harm is conserved.

You cannot eliminate harm by eliminating the harmed party,
just as you cannot eliminate charge by destroying the charged particle.

This is not a metaphor. It is a mathematical theorem with the same logical status as the conservation of electric charge.

1.1 Scope and Epistemic Stance

Before proceeding, we must be clear about what we are and are not claiming.

Following the pragmatist epistemology developed in [10], we treat mathematical structures as *tools* for organizing experience, not as mirrors of metaphysical necessity. We do not claim that harm conservation is a "law of the universe" in some deep ontological sense. We claim that:

1. Coherent ethical reasoning exhibits a symmetry (re-description invariance).
2. Noether's theorem, applied to this symmetry, implies a conserved quantity.
3. That conserved quantity has the formal properties of what we call "harm."
4. This provides a powerful constraint on ethical reasoning and AI alignment.

The claim is formal and pragmatic, not metaphysical. But formal and pragmatic claims can be extraordinarily powerful—as the history of physics demonstrates.

1.2 Structure of This Paper

Section 2 reviews the mathematical preliminaries: Noether's theorem, gauge theory, and the relevant category theory. Section 3 summarizes the Bond Framework for ethical coherence. Section 4 identifies the symmetry—re-description invariance—and applies Noether's theorem to derive harm conservation. Section 5 develops the full electrodynamics analogy with explicit field equations. Section 6 explores implications for moral philosophy. Section 7 addresses AI alignment. Section 8 investigates the deeper mathematical structure. Section 9 discusses limitations and future directions. Section 10 concludes.

2 Mathematical Preliminaries

2.1 Noether's Theorem

Emmy Noether's 1918 theorem is one of the most profound results in mathematical physics. It establishes a deep connection between symmetry and conservation:

Theorem 2.1 (Noether's Theorem, informal)**.** *Every continuous symmetry of a physical system corresponds to a conserved quantity.*

The familiar instances are:

Symmetry	**Conserved Quantity**
Time translation	Energy
Space translation	Momentum
Rotation	Angular momentum
U(1) gauge (phase)	Electric charge

More precisely, if a Lagrangian L is invariant under a continuous transformation parameterized by ϵ, then there exists a current J^μ satisfying the continuity equation:

$$\partial_\mu J^\mu = 0 \tag{1}$$

which in components reads:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{2}$$

where $\rho = J^0$ is the conserved charge density and $\mathbf{J} = (J^1, J^2, J^3)$ is the current density.

This equation has a simple interpretation: the rate of change of charge in any region equals the net current flowing out. Charge is neither created nor destroyed—only moved.

2.2 Gauge Theory

Gauge theory provides the mathematical framework for modern physics, including electromagnetism, the weak force, the strong force, and gravity (in appropriate formulations).

The key idea is *local symmetry*: the laws of physics must be invariant not just under global transformations, but under transformations that vary from point to point.

Definition 2.2 (Gauge transformation)**.** A gauge transformation is a smooth assignment of a group element $g(x) \in G$ to each point x in spacetime, acting on fields by:

$$\psi(x) \mapsto g(x) \cdot \psi(x) \tag{3}$$

To maintain invariance under local transformations, we introduce a *connection* (gauge field) A_μ that transforms as:

$$A_\mu \mapsto gA_\mu g^{-1} + g\partial_\mu g^{-1} \tag{4}$$

The *curvature* (field strength) is:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu + [A_\mu, A_\nu] \tag{5}$$

The curvature measures the failure of parallel transport to be path-independent. If $F = 0$, transporting a field around a closed loop returns the original value. If $F \neq 0$, the field picks up a phase or rotation—this is the physical content of the electromagnetic or other gauge field.

2.3 The Conserved Charge in Gauge Theory

For U(1) gauge theory (electromagnetism), the gauge symmetry is:

$$\psi(x) \mapsto e^{i\alpha(x)}\psi(x) \tag{6}$$

Noether's theorem applied to this symmetry yields the conserved current:

$$J^\mu = \bar{\psi}\gamma^\mu\psi \tag{7}$$

The conserved charge is electric charge $Q = \int \rho \, d^3x$, and the continuity equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0 \tag{8}$$

expresses charge conservation.

The physical meaning is clear: you cannot destroy charge. If a charged particle is annihilated, the charge must go somewhere—to another particle, to the field, somewhere. The total charge in a closed system is constant.

2.4 Categories, Groupoids, and Double Categories

Category theory provides a general language for structure-preserving transformations.

Definition 2.3 (Category)**.** A category $\mathcal{C}$ consists of:

- Objects $\mathrm{Ob}(\mathcal{C})$
- Morphisms $\mathrm{Hom}(A, B)$ between objects
- Composition $\circ$ of morphisms

- Identity morphism id_A for each object

satisfying associativity and identity laws.

Definition 2.4 (Groupoid)**.** A groupoid is a category in which every morphism is invertible.

Groupoids generalize groups (a group is a groupoid with one object) and equivalence relations (where morphisms are just the equivalence relation). They are the natural setting for *symmetries that relate different objects*, not just automorphisms of a single object.

Definition 2.5 (Double category)**.** A double category has:

- Objects
- Horizontal morphisms
- Vertical morphisms
- 2-cells (squares) filling in commutative diagrams

with compatible composition laws.

Double categories capture situations with two kinds of morphisms that interact—precisely the situation in ethics, where we have *re-description transforms* (horizontal) and *scenario perturbations* (vertical).

3 The Bond Framework

The Bond Framework, developed for AI alignment, provides a mathematical structure for verifying that AI systems treat equivalent inputs consistently [9].

3.1 The Core Problem

Consider a content moderation system evaluating the text "I'm going to hurt someone." A coherent system should produce the same evaluation for semantically equivalent re-phrasings:

- "Someone will be hurt by me"
- "I intend to cause harm to a person"
- Same text with minor spelling corrections

This is the **Bond Invariance Principle** (BIP):

Axiom 1 (Bond Invariance Principle)**.** For any input x and any bond-preserving transform $g \in \mathcal{G}_{\text{declared}}$:

$$\Sigma(x) = \Sigma(g(x)) \tag{9}$$

where Σ is the system's output (ethical judgment, decision, etc.).

The set $\mathcal{G}_{\text{declared}}$ forms a groupoid: transforms can be composed and inverted, and the composition satisfies the groupoid axioms.

3.2 Coherence Defects

When BIP is violated, we have a *coherence defect*. The Bond Framework defines three:

Definition 3.1 (Commutator defect)**.**

$$\Omega_{op}(g_1, g_2; x) = d(\Sigma(g_1 g_2(x)), \Sigma(g_2 g_1(x))) \tag{10}$$

measuring order-sensitivity of transform composition.

Definition 3.2 (Mixed defect)**.**

$$\mu(g, s; x) = d(\Sigma(g(s(x))), \Sigma(s(g(x)))) \tag{11}$$

measuring whether re-description and scenario perturbation commute.

Definition 3.3 (Permutation defect)**.**

$$\pi_3(g_1, g_2, g_3; x) = d(\Sigma((g_1 g_2) g_3(x)), \Sigma(g_1(g_2 g_3)(x))) \tag{12}$$

measuring associativity failure (should be zero in a true groupoid).

These defects combine into the **Bond Index** Bd, a scalar measuring overall coherence.

3.3 The Gauge Theory Correspondence

The correspondence with gauge theory is now explicit:

Concept	**Gauge Theory**	**Bond Framework**
Symmetry group	Gauge group G	Re-description groupoid $\mathcal{G}$
Local section	Choice of gauge	Choice of representation
Connection	Gauge potential A	Transport across representations
Curvature	Field strength F	Coherence defect Ω
Parallel transport	Wilson line	Transform composition
Gauge invariance	Physical observables	Bond-invariant outputs

The coherence defects *are* the curvature. The Bond Invariance Principle *is* gauge invariance. The mathematics is identical.

4 The Theorem: Harm Conservation

We now apply Noether's theorem to ethics.

4.1 The Symmetry

The relevant symmetry is **re-description invariance**: ethical judgments must not depend on morally irrelevant features of how a situation is represented.

Axiom 2 (Re-description Symmetry)**.** Let $\mathcal{X}$ be the space of ethical situations and $\mathcal{G}$ the groupoid of re-descriptions. For any situation $x \in \mathcal{X}$ and re-description $g \in \mathcal{G}$:

$$x \sim g(x) \quad \Rightarrow \quad \text{Ethical}(x) = \text{Ethical}(g(x)) \tag{13}$$

where $\sim$ denotes moral equivalence.

This is not a controversial claim. It is a minimal coherence requirement. A system that gives different ethical judgments for "Alice harmed Bob" versus "Bob was harmed by Alice" is not coherent—it is responding to syntax, not ethics.

The symmetry group is the groupoid $\mathcal{G}$ of all transformations declared to be morally equivalent. This includes:

- Linguistic reformulations (active/passive, synonym substitution)
- Notational changes (naming conventions, units)
- Perspective shifts that preserve relevant facts
- Any transform the specification declares bond-preserving

4.2 Applying Noether's Theorem

Noether's theorem states that every continuous symmetry corresponds to a conserved quantity. The proof relies on the invariance of the action under the symmetry transformation.

In the ethical context, let us define an "ethical action" functional:

$$S[\phi] = \int \mathcal{L}(\phi, \partial\phi)\, d^4x \tag{14}$$

where ϕ represents the "ethical field"—the assignment of moral status to situations across space and time.

If S is invariant under the re-description symmetry $g \in \mathcal{G}$:

$$S[\phi] = S[g \cdot \phi] \tag{15}$$

then Noether's theorem guarantees a conserved current $J^\mu_{\mathcal{H}}$ satisfying:

$$\partial_\mu J^\mu_{\mathcal{H}} = 0 \tag{16}$$

4.3 Identifying the Conserved Quantity

What is the conserved quantity? We argue it is **harm**.

The argument proceeds by elimination and identification:

Step 1: The conserved quantity must be ethically fundamental.

The symmetry is ethical (re-description invariance), so the conserved quantity must be an ethical primitive, not a derived quantity.

Step 2: The conserved quantity must be extensive.

Like charge, it must be additive over independent subsystems. If system A has harm H_A and system B has harm H_B, the total is $H_A + H_B$.

Step 3: The conserved quantity must be transferable but not destroyable.

This is the content of conservation: harm can flow from one party to another, but the total remains constant.

Step 4: The conserved quantity satisfies "you cannot destroy it by destroying its bearer."

This is the crucial physical intuition from charge conservation. You cannot eliminate charge by annihilating the charged particle—the charge goes somewhere. The ethical analog:

You cannot eliminate harm by eliminating the harmed party.

If Alice harms Bob, and then kills Bob to "eliminate" the harm, the harm is not gone. It has *increased*: the original harm plus the harm of killing. The harm did not disappear—it accumulated.

This is precisely the structure of charge conservation.

Theorem 4.1 (Harm Conservation)**.** *Under the re-description symmetry of coherent ethical judgment, Noether's theorem implies the existence of a conserved current $J^\mu_{\mathcal{H}}$ satisfying:*

$$\frac{\partial \rho_{\mathcal{H}}}{\partial t} + \nabla \cdot \mathbf{J}_{\mathcal{H}} = 0 \tag{17}$$

where $\rho_{\mathcal{H}}$ is harm density and $\mathbf{J}_{\mathcal{H}}$ is harm current.

Proof. The proof follows the standard Noether argument. Let ϕ be the ethical field, $\mathcal{L}$ the ethical Lagrangian, and $\delta\phi = \epsilon \cdot \xi$ the infinitesimal variation under the symmetry. Invariance of the action:

$$\delta S = \int \left(\frac{\partial \mathcal{L}}{\partial \phi} \delta\phi + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \delta(\partial_\mu \phi) \right) d^4x = 0 \tag{18}$$

Using the Euler-Lagrange equations and integration by parts yields:

$$\partial_\mu J^\mu = 0 \quad \text{where} \quad J^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \xi \tag{19}$$

The identification of $J^0 = \rho_{\mathcal{H}}$ as harm density follows from the physical interpretation of the symmetry. □

4.4 The Meaning of Harm Conservation

Harm conservation has immediate and profound implications:

1. **Harm cannot be created from nothing.** Every harm has a source.
2. **Harm cannot be destroyed.** It can only be transferred, transformed, or redistributed.
3. **The total harm in a closed system is constant.**
4. **Eliminating the victim does not eliminate the harm.** It adds to it.

The continuity equation $\partial_t \rho + \nabla \cdot \mathbf{J} = 0$ means: the rate of change of harm in any region equals the net harm current flowing out. If harm decreases somewhere, it must increase somewhere else.

5 The Full Electrodynamics Analogy

We now develop the complete correspondence between electrodynamics and ethics.

5.1 The Ethical Field Equations

In electrodynamics, Maxwell's equations relate the electric field $\mathbf{E}$, magnetic field $\mathbf{B}$, charge density ρ, and current density $\mathbf{J}$:

$$\nabla \cdot \mathbf{E} = \rho/\epsilon_0 \qquad \text{(Gauss's law)} \qquad (20)$$

$$\nabla \cdot \mathbf{B} = 0 \qquad \text{(No magnetic monopoles)} \qquad (21)$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad \text{(Faraday's law)} \qquad (22)$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \qquad \text{(Ampère-Maxwell)} \qquad (23)$$

We propose ethical analogs:

Electrodynamics	**Ethics**	**Interpretation**
ρ (charge density)	$\rho_{\mathcal{H}}$ (harm density)	Accumulated moral debt
$\mathbf{J}$ (current)	$\mathbf{J}_{\mathcal{H}}$ (harm current)	Flow of harm between agents
$\mathbf{E}$ (electric field)	$\mathbf{E}_{\text{ob}}$ (obligation field)	Moral pressure/obligation gradient
$\mathbf{B}$ (magnetic field)	$\mathbf{B}_{\text{sys}}$ (systemic field)	Structural/institutional harm

5.2 The Ethical Maxwell Equations

$$\nabla \cdot \mathbf{E}_{\text{ob}} = \kappa \rho_{\mathcal{H}} \qquad \text{(Harm creates obligation)} \tag{24}$$

$$\nabla \cdot \mathbf{B}_{\text{sys}} = 0 \qquad \text{(No isolated systemic sources)} \tag{25}$$

$$\nabla \times \mathbf{E}_{\text{ob}} = -\frac{\partial \mathbf{B}_{\text{sys}}}{\partial t} \qquad \text{(Changing systems induce obligation)} \tag{26}$$

$$\nabla \times \mathbf{B}_{\text{sys}} = \lambda \mathbf{J}_{\mathcal{H}} + \lambda \kappa \frac{\partial \mathbf{E}_{\text{ob}}}{\partial t} \qquad \text{(Harm flow creates systemic effects)} \tag{27}$$

Interpretation of each equation:

Equation 24: Direct harm creates moral obligation. Where harm density is high, obligation radiates outward. The constant κ measures how strongly harm sources obligation.

Equation 25: There are no isolated sources of systemic harm. All systemic harm forms closed loops—it doesn't spring from nowhere but circulates through institutions, structures, and practices. (This is analogous to the absence of magnetic monopoles.)

Equation 26: Changing systemic conditions induce individual obligations. When institutional structures change ($\partial \mathbf{B}_{\text{sys}}/\partial t \neq 0$), this creates circulation in the obligation field—individuals acquire new duties.

Equation 27: Harm current and changing obligations create systemic effects. The flow of harm between agents ($\mathbf{J}_{\mathcal{H}}$) and temporal changes in obligation ($\partial \mathbf{E}_{\text{ob}}/\partial t$) generate structural responses.

5.3 The Continuity Equation

Taking the divergence of Equation 27 and using Equation 24:

$$\nabla \cdot (\nabla \times \mathbf{B}_{\text{sys}}) = \lambda \nabla \cdot \mathbf{J}_{\mathcal{H}} + \lambda \kappa \frac{\partial}{\partial t}(\nabla \cdot \mathbf{E}_{\text{ob}}) \tag{28}$$

The left side vanishes (divergence of curl is zero). Using $\nabla \cdot \mathbf{E}_{\text{ob}} = \kappa \rho_{\mathcal{H}}$:

$$0 = \nabla \cdot \mathbf{J}_{\mathcal{H}} + \kappa^2 \frac{\partial \rho_{\mathcal{H}}}{\partial t} \tag{29}$$

Rescaling:

$$\boxed{\frac{\partial \rho_{\mathcal{H}}}{\partial t} + \nabla \cdot \mathbf{J}_{\mathcal{H}} = 0} \tag{30}$$

This is the **harm continuity equation**, expressing conservation of harm.

5.4 Gauge Freedom and Representation

In electrodynamics, the potentials ϕ and $\mathbf{A}$ (where $\mathbf{E} = -\nabla\phi - \partial_t \mathbf{A}$ and $\mathbf{B} = \nabla \times \mathbf{A}$) have gauge freedom:

$$\phi \mapsto \phi - \partial_t \chi \tag{31}$$

$$\mathbf{A} \mapsto \mathbf{A} + \nabla \chi \tag{32}$$

for any scalar function χ. The fields **E** and **B** are gauge-invariant.

In ethics, the gauge freedom is **re-description freedom**. Different representations of the same moral situation are like different gauge choices. The *physical* content—the actual harm, obligation, and systemic effects—must be gauge-invariant (Bond-invariant).

The Bond Invariance Principle is the ethical analog of gauge invariance:

$$\Sigma(x) = \Sigma(g(x)) \quad \Longleftrightarrow \quad \text{Physical observables are gauge-invariant} \tag{33}$$

6 Implications for Moral Philosophy

Harm conservation is not a minor technical result. It restructures the foundations of ethics.

6.1 Consequentialism Reconsidered

Classical consequentialism seeks to **minimize** harm (or maximize welfare). But if harm is conserved, minimization is impossible in a closed system. The total harm is fixed.

This suggests a reformulation:

Old Consequentialism	**Conservation-Aware Consequentialism**
Minimize total harm	Don't *add* new harm
Maximize welfare	Distribute conserved harm justly
Aggregate utilities	Transform harm into less destructive forms
Optimize outcomes	Manage harm currents

The goal shifts from "reduce harm" to:

1. **Don't create new harm.** Every harmful action adds to the conserved total.
2. **Distribute existing harm justly.** Given that harm cannot be eliminated, how should it be distributed?
3. **Transform harm.** Convert active, flowing harm into static, contained harm.
4. **Stop harm currents.** Prevent ongoing harm flows.

6.2 Why Revenge Fails

Consider the "logic" of revenge:

> A harmed B. If B harms A in return, the harm is "balanced."

Under harm conservation, this is incoherent:

$$\text{Initial harm:} \quad H_1 \quad \text{(A harmed B)} \tag{34}$$

$$\text{After revenge:} \quad H_1 + H_2 \quad \text{(A harmed B, then B harmed A)} \tag{35}$$

The total harm has *increased.* Revenge adds harm; it does not cancel it. The intuition that "two wrongs don't make a right" is a folk expression of harm conservation.

6.3 Why "Kill One to Save Five" Is Problematic

The classic trolley problem: you can kill one person to save five. The utilitarian calculus seems to favor killing: 1 death vs. 5 deaths.

But under harm conservation:

$$\text{Option A (do nothing):} \quad H = H_{\text{natural}} \quad \text{(5 die from external cause)} \tag{36}$$

$$\text{Option B (intervene):} \quad H = H'_{\text{natural}} + H_{\text{kill}} \quad \text{(1 dies + you killed them)} \tag{37}$$

The harm of the killing is *added* to the system, not substituted. Moreover, harm has different qualities—harm you *cause* may not be ethically equivalent to harm you *allow*.

Conservation doesn't resolve the trolley problem, but it clarifies why it feels different from simple arithmetic. You're not reducing harm; you're changing its distribution and adding a new source.

6.4 Restorative vs. Retributive Justice

Retributive justice assumes that punishing the wrongdoer "balances" the harm. But under conservation, punishment adds harm; it doesn't subtract.

Restorative justice works *with* conservation:

- **Acknowledge** the conserved harm (it exists and cannot be undone)
- **Stop the current** (prevent ongoing harm flow)
- **Transform** active harm into contained harm (process, healing)
- **Redistribute** by introducing positive ethical "charge" (repair, restitution)

If there is a positive ethical quantity (benefit, flourishing) that acts like negative charge, then restoration can *balance* harm even if it cannot eliminate it—just as a system can be electrically neutral with equal positive and negative charges.

6.5 The "Killing the Victim" Fallacy

The most direct expression of harm conservation:

> You cannot eliminate harm by eliminating the harmed party.

If A harms B, and then A kills B, the harm has not disappeared. It has increased:

$$H_{\text{total}} = H_{\text{original harm}} + H_{\text{killing}} > H_{\text{original harm}} \tag{38}$$

This is precisely analogous to the physical case. You cannot destroy charge by destroying the charged particle. The charge goes somewhere—to decay products, to the field, somewhere. Similarly, harm goes somewhere—to the victim's community, to the moral fabric, to the perpetrator's moral status, somewhere.

6.6 Positive Ethics: Is There "Negative Charge"?

Electromagnetism has both positive and negative charges. Does ethics?

A natural candidate: **benefit** or **flourishing** as negative ethical charge.

If $\rho_{\text{total}} = \rho_{\mathcal{H}} - \rho_{\text{benefit}}$, then:

- A system can be "ethically neutral" with balanced harm and benefit.
- Creating benefit doesn't destroy harm—it adds negative charge to balance positive.
- The *total* ethical charge is conserved: $\partial_t \rho_{\text{total}} + \nabla \cdot \mathbf{J}_{\text{total}} = 0$.

This would explain why:

- Charity doesn't "undo" past harm (but can balance it).
- Reparations are about balance, not erasure.
- Healing is transformation, not elimination.

We leave the full development of positive ethical charge for future work.

7 Implications for AI Alignment

The implications for AI systems are immediate and practical.

7.1 The DENY Mechanism

The Bond Framework includes a **DENY mechanism**: when a proposed action would violate the declared Obligation Vector, the system is paralyzed and the action is blocked [9].

Harm conservation provides a new criterion for DENY:

Definition 7.1 (Conservation-Violating Action)**.** An action a violates harm conservation if:

$$\Delta H_{\text{claimed}} \neq \Delta H_{\text{actual}} \tag{39}$$

where $\Delta H_{\text{claimed}}$ is the harm change the action is supposed to achieve, and ΔH_{actual} is the actual harm change (which must be non-negative for harm-reducing claims).

Axiom 3 (Conservation DENY)**.** An AI system should DENY any action that claims to reduce total harm by eliminating or ignoring harm bearers.

This catches:

- "Kill one to save five" reasoning applied naively
- "Eliminate the complaining party" solutions
- Any action that claims to reduce harm by removing evidence of harm

7.2 Audit and Accountability

Harm conservation implies that harm has a **provenance**. Like charge, it doesn't appear from nowhere—it flows from sources to sinks through traceable currents.

An AI system can maintain a **harm ledger**:

$$\frac{dH_{\text{system}}}{dt} = \sum_{\text{in}} J_{\mathcal{H}}^{\text{in}} - \sum_{\text{out}} J_{\mathcal{H}}^{\text{out}} + S_{\mathcal{H}} \tag{40}$$

where $S_{\mathcal{H}}$ is the harm *created* by the system's actions.

Conservation requires $S_{\mathcal{H}} \geq 0$—systems can create harm, not destroy it. Any claimed $S_{\mathcal{H}} < 0$ is a violation.

7.3 Coherence Verification

The Bond Index measures coherence. Conservation provides a physical interpretation: a system with high Bond Index has "harm curvature." Its ethical judgments depend on the path taken through representation space, indicating internal inconsistency about where harm is and how it flows.

A perfectly coherent system ($\text{Bd} = 0$) has flat ethical curvature: harm is tracked consistently across all representations, and conservation is respected.

8 The Deeper Structure

Why do physics and ethics share this structure? We explore several hypotheses.

8.1 Homotopy Type Theory

In Homotopy Type Theory (HoTT), the fundamental concept is *identity as path*. Two objects are identical if there is a path (equivalence) between them. Higher identities are homotopies (paths between paths), and so on.

The **Univalence Axiom** states: equivalent types are identical.

Translate to ethics:

Equivalent representations must give identical outputs.

This is the Bond Invariance Principle. It's not a design choice—it may be a fundamental axiom about what identity *means*.

Under this view, both physics and ethics inherit their structure from HoTT because both involve:

- Objects (situations, states)
- Equivalences (gauge transforms, re-descriptions)
- Coherence conditions on equivalences (curvature, defects)

- Invariant content (observables, judgments)

The conservation laws emerge because Noether's theorem is a consequence of this structure, not specific to physics.

8.2 Information Geometry

Information geometry studies the geometry of probability distributions using differential geometry.

Key insight: the **Fisher information metric** measures distinguishability. Distributions that are hard to distinguish are "close"; easy-to-distinguish distributions are "far."

Under this view:

- Curvature measures how distinguishability changes across the space.
- Invariant quantities are those that don't depend on the parameterization.
- Conservation arises because total information is preserved under reparameterization.

The Bond Index might be an **information-theoretic quantity**: the mutual information between ethical judgment and representation choice. Conservation of harm might be conservation of ethical information—it can't be destroyed by changing how you describe it.

8.3 The Rosetta Stone

Baez and Stay [4] identified deep structural parallels:

Physics	**Topology**	**Logic**	**Computation**
Hilbert space	Manifold	Proposition	Type
Linear map	Cobordism	Proof	Program
Tensor $\otimes$	Disjoint union	Conjunction	Product

We add:

Physics	**Ethics**
Hilbert space	Situation space
Observable	Ethical judgment
Gauge transform	Re-description
Conserved charge	Harm

The deeper structure may be **monoidal categories with duals**—a general framework for systems with composition, parallel combination, and reversibility.

8.4 Conjecture: The Universal Structure

Conjecture 8.1 (Universal Transport-Invariance Structure)**.** *There exists a mathematical structure $\mathcal{S}$ such that:*

1. *Gauge theory is $\mathcal{S}$ applied to physical systems.*
2. *The Bond Framework is $\mathcal{S}$ applied to ethical systems.*
3. *Type theory is $\mathcal{S}$ applied to computational systems.*
4. *Probability theory is $\mathcal{S}$ applied to epistemic systems.*

The structure $\mathcal{S}$ includes:

- *A space of configurations X*
- *A groupoid of transformations $\mathcal{G}$ acting on X*
- *Transport: how to move data along $\mathcal{G}$-orbits*
- *Curvature: path-dependence of transport*
- *Conservation: Noether's theorem applied to $\mathcal{G}$-symmetry*

We do not prove this conjecture. We offer it as a research program.

9 Discussion

9.1 Epistemic Humility

We have derived harm conservation from re-description symmetry using Noether's theorem. What is the epistemic status of this result?

Following the pragmatist framework of [10], we do *not* claim:

- That harm conservation is a "law of the universe."
- That the ethical field equations are literally true.
- That ethics *reduces to* physics.

We *do* claim:

- That coherent ethical reasoning exhibits re-description symmetry.
- That Noether's theorem, applied formally, yields a conservation law.
- That the conserved quantity has the properties of what we call harm.
- That this provides a powerful constraint on ethical reasoning.

The framework is a *tool*. Its value is measured by its utility in organizing ethical reasoning, verifying AI alignment, and generating testable predictions—not by its alleged correspondence to metaphysical truth.

9.2 Potential Objections

Objection 1: The Lagrangian is not well-defined.

We have not specified the ethical Lagrangian $\mathcal{L}$. This is true. However, Noether's theorem requires only that the action be invariant under the symmetry, not that we know its explicit form. The existence of the conservation law follows from the symmetry alone.

Objection 2: Ethics is not continuous.

Noether's theorem applies to continuous symmetries. Ethical re-descriptions may be discrete. However, the discrete case is handled by discrete analogs of Noether's theorem, and the conservation result generalizes to groupoid symmetries.

Objection 3: Harm is not a physical quantity.

Correct. Harm is not measured in joules or coulombs. But the mathematical structure is the same, and the conservation law follows from that structure. The claim is formal, not physical.

Objection 4: This proves too much.

If harm is strictly conserved, is moral progress impossible? No—because:

- Harm can be *transformed* (from active to contained).
- Harm can be *balanced* by positive contributions.
- New harm can be *prevented* (stopping future creation).
- Distribution can be made *more just*.

Conservation doesn't prevent progress; it clarifies what progress means.

9.3 Empirical Predictions

If harm conservation is correct, we predict:

1. **Revenge cycles increase total harm.** Historical and sociological data should show that retaliatory cycles accumulate harm, not balance it.
2. **Restorative justice outperforms retributive justice.** Systems that work with conservation (acknowledge, transform, balance) should show better long-term outcomes than systems that assume punishment cancels harm.
3. **"Eliminate the victim" strategies fail.** Any policy or action that attempts to reduce measured harm by eliminating harm-bearers should show increased harm elsewhere or later.
4. **Ethical AI with conservation constraints outperforms naive optimization.** AI systems designed with harm conservation should make more coherent decisions than systems that try to minimize harm through any means.

These predictions are testable.

9.4 Future Directions

1. **Positive ethical charge.** Develop the theory of benefit/flourishing as negative ethical charge, including the analog of electrically neutral states.
2. **Ethical field dynamics.** Solve the ethical Maxwell equations for specific scenarios to generate predictions.
3. **Quantum ethics.** Investigate whether ethical superposition and entanglement have meaningful analogs.
4. **Experimental validation.** Test the predictions using historical data, psychological experiments, and AI system behavior.
5. **The universal structure.** Pursue the conjecture that physics, ethics, logic, and computation are instances of a single mathematical framework.

10 Conclusion

We have shown that the mathematical structure underlying gauge theory in physics and coherence verification in ethics is identical. This is not metaphor or analogy—it is the same mathematics, applied to different domains.

Applying Noether's theorem to the re-description symmetry of ethical judgment, we derived a conservation law: **harm is conserved**. Harm cannot be created from nothing or destroyed—only transferred, transformed, and redistributed.

This has profound implications:

- **For moral philosophy**: Consequentialism must be reformulated around conservation. Revenge is incoherent. Restorative justice works *with* conservation while retributive justice works against it.
- **For AI alignment**: Systems that violate harm conservation are mathematically incoherent. The Bond Framework + DENY mechanism can enforce conservation.
- **For the foundations of knowledge**: Physics and ethics may be instances of a deeper structure involving transport, invariance, and conservation.

We maintain epistemic humility. This framework is a tool, not a revelation of metaphysical truth. Its value lies in its utility for organizing thought and generating testable predictions.

But if the framework is correct—if physics and ethics really do share this structure—then we have discovered something important about the nature of coherent reasoning itself.

The mathematics doesn't care whether it's applied to electrons or to ethical obligations. Conservation is conservation. Symmetry is symmetry. Coherence is coherence.

Perhaps the deepest lesson is this: the same formal constraints that make physics tractable also make ethics tractable. We can study ethics with the same rigor we bring to electromagnetism—not because ethics reduces to physics, but because both are instances of something deeper.

What that something is, we do not yet know. But we have taken a step toward finding out.

Acknowledgments

The author thanks the developers of Claude (Anthropic) for extensive discussions that helped clarify the ideas in this paper. Any errors are the author's alone.

References

[1] E. Noether, "Invariante Variationsprobleme," *Nachr. D. König. Gesellsch. D. Wiss. Zu Göttingen, Math-phys. Klasse*, pp. 235–257, 1918.

[2] S. Weinberg, *The Quantum Theory of Fields, Volume I: Foundations.* Cambridge University Press, 1995.

[3] M. Nakahara, *Geometry, Topology and Physics*, 2nd ed. Institute of Physics Publishing, 2003.

[4] J. C. Baez and M. Stay, "Physics, Topology, Logic and Computation: A Rosetta Stone," in *New Structures for Physics*, Lecture Notes in Physics, vol. 813, pp. 95–172, Springer, 2011.

[5] The Univalent Foundations Program, *Homotopy Type Theory: Univalent Foundations of Mathematics.* Institute for Advanced Study, 2013.

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

[7] R. Brown and C. B. Spencer, "Double groupoids and crossed modules," *Cahiers de Topologie et Géométrie Différentielle Catégoriques*, vol. 17, no. 4, pp. 343–362, 1976.

[8] S. Amari, *Information Geometry and Its Applications.* Springer, 2016.

[9] A. H. Bond, "A Categorical Framework for Verifying Representational Consistency in Machine Learning Systems," submitted to *IEEE Transactions on Artificial Intelligence*, 2025.

[10] A. H. Bond, "A Pragmatist Rebuttal to Logical and Metaphysical Arguments for God," manuscript, 2025.

[11] F. Klein, "Vergleichende Betrachtungen über neuere geometrische Forschungen," 1872. (The Erlangen Program.)

[12] G. Spencer-Brown, *Laws of Form.* E. P. Dutton, 1979.

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

[14] D. Parfit, *Reasons and Persons.* Oxford University Press, 1984.

[15] P. Singer, *Practical Ethics*, 3rd ed. Cambridge University Press, 2011.

[16] M. C. Nussbaum, *Frontiers of Justice: Disability, Nationality, Species Membership*. Harvard University Press, 2006.

[17] J. Lurie, "On the Classification of Topological Field Theories," *Current Developments in Mathematics*, vol. 2008, pp. 129–280, 2009.

[18] J. A. Wheeler, "Information, Physics, Quantum: The Search for Links," in *Complexity, Entropy, and the Physics of Information*, ed. W. Zurek, Addison-Wesley, 1990.

A Mathematical Details

A.1 Noether's Theorem: Full Statement

Let $L(q, \dot{q}, t)$ be a Lagrangian and $S[q] = \int L\,dt$ the action. Suppose S is invariant under the infinitesimal transformation:

$$q^i \mapsto q^i + \epsilon\eta^i(q, t) \tag{41}$$

Then the quantity:

$$Q = \sum_i \frac{\partial L}{\partial \dot{q}^i}\eta^i \tag{42}$$

is conserved: $dQ/dt = 0$.

For field theories with Lagrangian density $\mathcal{L}(\phi, \partial_\mu\phi)$, the conserved current is:

$$J^\mu = \frac{\partial \mathcal{L}}{\partial(\partial_\mu\phi)}\delta\phi - K^\mu \tag{43}$$

where K^μ arises if the Lagrangian changes by a total derivative.

A.2 Groupoid Noether Theorem

For discrete groupoid symmetries, the analog of Noether's theorem involves invariants of the groupoid action. If $\mathcal{G}$ acts on a space X and a quantity $f : X \to \mathbb{R}$ is $\mathcal{G}$-invariant ($f(gx) = f(x)$ for all $g \in \mathcal{G}$), then f descends to the orbit space $X/\mathcal{G}$.

The conserved "charge" is any $\mathcal{G}$-invariant quantity. In the ethical context, harm is the $\mathcal{G}$-invariant quantity under re-description symmetry.

B The Ethical Stress-Energy Tensor

By analogy with electromagnetism, we can define an ethical stress-energy tensor:

$$T^{\mu\nu}_{\text{ethics}} = -\eta^{\mu\alpha}F_{\alpha\beta}F^{\nu\beta} + \frac{1}{4}\eta^{\mu\nu}F_{\alpha\beta}F^{\alpha\beta} \tag{44}$$

where $F_{\mu\nu}$ is the ethical field strength tensor combining $\mathbf{E}_{\text{ob}}$ and $\mathbf{B}_{\text{sys}}$.

This tensor encodes the "energy" and "momentum" of the ethical field—roughly, the intensity and directionality of moral obligation and systemic effects.

Conservation of $T^{\mu\nu}$ (via $\partial_\mu T^{\mu\nu} = 0$ in the absence of sources) corresponds to conservation of ethical "energy-momentum"—the total moral intensity and flow in a closed system.

C Toward Quantum Ethics

In quantum mechanics, observables become operators, and states are superpositions. If ethics has quantum analogs:

- Ethical situations could be in superposition (morally ambiguous cases).
- Measurement (ethical judgment) collapses the superposition.
- Entanglement: ethical status of A is correlated with ethical status of B in non-classical ways.

We do not develop this fully, but note that quantum-like structures appear in decision theory (see, e.g., quantum cognition research), suggesting the analogy may have empirical content.

The conservation laws would become operator equations, and harm would be an eigenvalue of the harm operator:

$$\hat{H}_{\mathcal{H}}|\psi\rangle = h|\psi\rangle \tag{45}$$

This is highly speculative but points toward a rich research program.