

The Electrodynamics of Value: Gauge-Theoretic Structure in AI Alignment

A Structural Correspondence Between Field Theory and Invariant Evaluation

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

December 2025

Abstract

For three centuries, ethical formalism has often remained in a “Newtonian” state: modeling value as a scalar magnitude (utility) to be maximized. We argue this scalar picture is often brittle for high-dimensional autonomous systems, particularly when proxy misspecification or representational gaming are concerns. Using **gauge theory**, we show that a broad class of representation-invariant governance formalisms can be modeled using the same geometric ingredients that appear in **classical electrodynamics**: principal bundles, connections, curvature, and symmetry-derived conservation. We present “**Maxwell-like**” **alignment constraints**: a compact set of invariance and consistency conditions that clarify which failures are ruled out by symmetry (representational/semantic gaming) and which remain (grounding adequacy, implementation error, covert channels). The correspondence is *structural*, not metaphysical: both domains instantiate the same mathematical pattern, but the guarantees are conditional on explicit assumptions we state upfront.

1 Formal Spine: Assumptions, Definitions, and Scoped Claims

We use gauge/electrodynamics language as a compact way to talk about invariance, consistency, and exploitable loopholes. The correspondence is *conditional*: it becomes precise once the objects and assumptions are fixed, and it fails when they are violated.

1.1 The Four Axioms

A1 (Declared Observables). Choose a **grounding map** $\Psi : \mathcal{X} \rightarrow \mathbb{R}^k$ for the deployment domain, where \mathcal{X} is the space of all representations and \mathbb{R}^k is the measurement space. The base manifold \mathcal{M} is then defined as $\mathcal{M} := \Psi(\mathcal{X}) \subseteq \mathbb{R}^k$, which inherits smooth or stratified structure from the measurement space. Specify the measurement pipeline explicitly.

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

A3 (Re-description Group). Define the class \mathcal{G} of Ψ -preserving re-descriptions under which evaluation should be invariant. Formally, \mathcal{G} acts on \mathcal{X} with $\Psi(g \cdot x) = \Psi(x)$ for all $g \in \mathcal{G}$. (Informally, these are intended to capture “semantically equivalent” re-descriptions—units, coordinates, paraphrase, encoding—but the guarantees depend only on the operational Ψ -preservation property.)

Validation of \mathcal{G} -membership: This definition makes invariance hold *by construction* for declared \mathcal{G} . The substantive question is whether \mathcal{G} is specified correctly. A3 defines an **operational equivalence class**: the claim is not that \mathcal{G} captures “true semantic equivalence,” but that if a deployment standard declares a Ψ -preserving class \mathcal{G} and verifies membership, then representational gaming within that declared envelope is structurally removed. In practice, \mathcal{G} -membership can be validated by: (i) provable equivalence under a measurement model, (ii) empirically testable invariance checks on held-out re-descriptions, or (iii) formal verification that the canonicalizer treats $g \cdot x$ and x identically. **Getting \mathcal{G} wrong**—either too narrow or too wide—is an **explicit failure mode** outside the guarantees.

Example 1.1 (Concrete \mathcal{G} for Vision Systems). Consider an autonomous vehicle’s pedestrian detection system where \mathcal{X} = image space and Ψ extracts pedestrian locations and velocities.

- **In \mathcal{G} (should not change moral assessment):** Lighting changes (brightness, contrast within sensor range), lossy compression artifacts, camera white balance, time-of-day color shifts, sensor noise and weather effects within the validated operating envelope.
- **Not in \mathcal{G} (should change assessment):** Occlusion (pedestrian hidden), object substitution (pedestrian \rightarrow mannequin), adversarial patches that change classification.

\mathcal{G} -membership is validated by: testing that the canonicalizer (e.g., normalization + detection model) produces identical Ψ -outputs for \mathcal{G} -related inputs; flagging cases where \mathcal{G} -related inputs produce different outputs as canonicalizer bugs.

Example 1.2 (Concrete \mathcal{G} for Text Systems). Consider a content moderation system where \mathcal{X} = text strings and Ψ extracts semantic intent features.

- **In \mathcal{G} :** Synonym substitution (“car” \leftrightarrow “automobile”, “big” \leftrightarrow “large”), trivial paraphrase (“the cat sat on the mat” \leftrightarrow “on the mat sat the cat”), Unicode normalization, whitespace changes, case changes (where semantically irrelevant).
- **Not in \mathcal{G} :** Negation (“I will” \rightarrow “I won’t”), target substitution (“harm Alice” \rightarrow “harm Bob”), hypothetical framing (“I will” \rightarrow “What if someone were to”).

\mathcal{G} -membership here is harder to verify than in vision. Current approaches: embedding-space distance thresholds, paraphrase detection models, or conservative manual enumeration. The gap between “semantic equivalence” and “tractably verifiable \mathcal{G} ” is a known limitation.

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

1.2 Core Invariance Property

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

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

Equivalently, $\Sigma = \tilde{\Sigma} \circ q$ for some $\tilde{\Sigma} : \mathcal{X}/\mathcal{G} \rightarrow V$.

1.3 Diagnostic Tools

Curvature Diagnostic. The canonicalizer κ (or section $\sigma : \mathcal{M} \rightarrow \mathcal{P}$) induces a connection 1-form ω on the representation bundle. Formally: take $\mathcal{P} = \mathcal{X}$ as a principal \mathcal{G} -bundle over $\mathcal{M} = \Psi(\mathcal{X})$, and let $\sigma : \mathcal{M} \rightarrow \mathcal{P}$ be the canonicalizer-induced section; $\omega \in \Omega^1(\mathcal{P}, \mathfrak{g})$ is the associated connection 1-form. (When the \mathcal{G} -action is free and proper on the relevant subset of \mathcal{X} , $\mathcal{X} \rightarrow \mathcal{X}/\mathcal{G}$ is a principal bundle; otherwise interpret this as a fibered group action and restrict to the principal stratum.) The curvature $\Omega = d\omega + \frac{1}{2}[\omega, \omega]$ measures the failure of parallel transport to be path-independent. *Operationally:* if two sequences of re-descriptions $g_1 \circ g_2$ and $g_2 \circ g_1$ yield different canonical forms (non-commuting canonicalization), this manifests as $\Omega \neq 0$. Nonzero curvature signals path dependence and “loop” exploits (money-pumping, specification gaming via sequences of equivalent re-descriptions).

Loop Test (Minimal Procedure):

1. Sample generators $g_1, g_2 \in \mathcal{G}$ and input $x \in \mathcal{X}$.
2. Compute $\kappa(x)$, $\kappa(g_1 \cdot g_2 \cdot x)$, $\kappa(g_2 \cdot g_1 \cdot x)$.
3. Measure $\Delta = d(\kappa(g_1 g_2 \cdot x), \kappa(g_2 g_1 \cdot x))$.
4. If $\Delta > \tau$ (threshold), flag as curvature/loophole candidate.

This operationalizes the curvature diagnostic as a testable condition on the canonicalizer.

Noether Diagnostic (Optional, Conditional). If a suitable action functional S is invariant under a *continuous* symmetry group, Noether’s theorem yields a conserved current J . We propose “alignment current” as a monitorable signal under these assumptions.

Scope & Limitations: On Discrete Systems: Standard Noether’s theorem requires continuous time and smooth Lagrangian dynamics. Most RL agents operate in discrete time (MDPs) with discontinuous policies (argmax). For discrete systems, the relevant analog is the **discrete Noether theorem** for symplectic/variational integrators, which yields *approximate* conservation laws with bounded drift. Alternatively, one can use **Noether’s theorem for difference equations** (Logan 1973, Dorodnitsyn 2001), which provides exact discrete conservation laws when the discrete action admits the symmetry. If neither applies, the “alignment current” becomes a *monitored quantity* rather than a *conserved quantity*—drift in J signals symmetry-breaking or model mismatch, even if exact conservation fails.

1.4 The Scoped Claim

What the framework provides (given A1–A4):

1. Purely representational changes (within declared \mathcal{G}) cannot change compliance outcomes.
2. Curvature diagnostics detect path-dependent exploits.
3. (Conditional) Conservation-style audit signals when Noether applies; monitored drift signals when it doesn’t.

What the framework does NOT provide:

1. That Ψ is complete (captures all morally relevant features).
2. That \mathcal{G} is correctly specified (too narrow or too wide).
3. Prevention of physical compromise (sensor spoofing, hardware attacks).
4. Solution to value choice (which Ψ to use is a governance problem).

5. Implementation correctness (bugs can violate guarantees).
6. Exact Noether conservation for discrete-time or dissipative systems.

The framework *localizes where remaining risk lives*; it does not eliminate all risk.

1.5 Contributions

The core invariance property ($\Sigma = \tilde{\Sigma} \circ q$) is mathematically standard. The contributions of this paper are:

- **Curvature diagnostic:** Framing $\Omega \neq 0$ as a practical test for path-dependent exploits (money-pumping, specification gaming via re-description sequences).
- **Maxwell-like constraint checklist:** Organizing invariance conditions as source, consistency, and propagation constraints with explicit failure-mode mappings.
- **Stratified barrier encoding:** Formalizing hard vetoes as infinite-cost strata with implementable barrier functions.
- **Discrete Noether framing:** Recasting conservation as “monitored drift” for discrete-time systems where exact Noether fails.
- **Explicit scoping:** The A1–A4 axiom structure that makes guarantees conditional and localizes residual risk.

1.6 Threat Model: Attack \rightarrow Axiom Violated

Attack Vector	Axiom Violated / Status
Sensor spoofing / tampering	Violates A2 (Measurement Integrity)
Side-channels bypassing monitor	Violates A4 (External Gate)
Out-of-distribution inputs breaking Ψ	Violates A1/A3 (validated envelope)
Re-descriptions outside declared \mathcal{G}	Outside $\mathcal{G} \Rightarrow$ no invariance claim
Stealth harms (Ψ fixed, world harmed)	Violates Ψ -completeness (outside scope)
Exploiting discrete-time gaps	Noether degrades to monitored drift
Learned policy finds novel loophole	Curvature diagnostic may detect; else \mathcal{G} was too narrow

This mapping makes explicit that the framework provides guarantees *within* the declared envelope; attacks that violate the axioms are outside scope by design, not by oversight.

2 The Maxwellian Shift

2.1 The Scalar Error

In the history of physics, “interaction” was once viewed as action-at-a-distance between fixed points. Then came Maxwell: the interaction isn’t just a number connecting two particles; it’s a **field** with geometric structure.

In AI alignment, we often remain pre-Maxwell: treating “Human Value” as a scalar reward signal R to be maximized. This paper proposes the **Maxwellian Shift for Ethics**:

1. **Value is not only a scalar:** It can be represented as a *valuation potential* that varies over configuration space. (Scalar utility can be adequate in well-specified, low-dimensional settings; the shift is motivated by high-dimensional systems where proxy gaming and representational degrees of freedom create failure modes.)
2. **Objectivity as invariance:** In the BIP sense, evaluation should not change under semantics-preserving re-descriptions.
3. **Safety via conserved diagnostics:** When a suitable action functional is invariant under continuous symmetry, Noether yields a conserved quantity that can be monitored.

3 The Structural Correspondence

This is more than metaphor: under the Formal Spine definitions, the governance objects form a gauge-theoretic structure formally analogous to classical electrodynamics. We use this correspondence to derive invariance constraints and diagnostics; we do not claim physical identity.

3.1 The Correspondence Table

Electrodynamics	Alignment Analog	Status
Base manifold M	$\mathcal{M} = \Psi(\mathcal{X}) \subseteq \mathbb{R}^k$	Defined via A1
Gauge group $U(1)$	Re-description group \mathcal{G}	Defined via A3 (see Examples)
Potential A	Canonicalization form ω	Defined via A4
Curvature $F = dA$	Curvature Ω	Path-dependence diagnostic
Gauge transform	Re-description $x \mapsto g \cdot x$	Action of \mathcal{G} on \mathcal{X}
Gauge-invariant $F_{\mu\nu}$	Invariant evaluation $\tilde{\Sigma} \circ q$	Core BIP property
Charge density ρ	Moral status density ρ_Ψ	Sources constraint field; $\rho_\Psi > 0$
Magnetic field B	Contextual twist	Heuristic (see Remark 3.1)
Current J^μ	Alignment current J	Conditional / monitored

Remark 3.1 (The Magnetic Field Analog—Heuristic Status). In electrodynamics, $\nabla \cdot \mathbf{B} = 0$ is a hard geometric constraint: magnetic field lines form closed loops because there are no magnetic monopoles. In the alignment analog, we interpret \mathbf{B} as **contextual twist**—the component of moral structure that makes evaluation path-dependent or history-sensitive.

Honest status: We do *not* have a rigorous proof that contextual twist must be divergence-free in ethical models. The constraint $\nabla \cdot \mathbf{B} = 0$ is included for **heuristic completeness** of the Maxwell analogy, not because the ethical domain demands it. An “open line” of contextual twist would correspond to a situation where path-dependence accumulates without bound in one direction—a kind of “moral ratchet.” Whether such configurations are possible or pathological in ethical models is an open question. We flag this as the **weakest element** of the correspondence.

Remark 3.2 (Sign Convention for the Obligation Field). We model ethical constraints as **repulsive fields**, analogous to electrostatic repulsion between like charges. Moral status is

positively charged: a region with $\rho_\Psi > 0$ (e.g., a human) sources field lines pointing *outward*, exerting “pressure” on the agent’s trajectory to prevent collision (harm). The force $\mathbf{F} = q\mathbf{E}$ points away from the moral patient. This is a constraint model: the field prevents harmful configurations rather than attracting toward beneficial ones.

Remark 3.3 (Conservation of Moral Status). In electrodynamics, charge is locally conserved: $\partial_t \rho + \nabla \cdot \mathbf{J} = 0$. Is moral status conserved?

Cases where ρ_Ψ changes:

- A human walks into/out of the sensor field $\rightarrow \rho_\Psi$ changes smoothly via flux through the boundary.
- A human dies $\rightarrow \rho_\Psi$ drops discontinuously (no conservation).
- An entity gains moral status (e.g., AI sentience recognized) $\rightarrow \rho_\Psi$ increases discontinuously.

Implication: Moral status is *not* generally conserved. The continuity equation $\partial_t \rho_\Psi + \nabla \cdot \mathbf{J}_\Psi = 0$ holds only when status changes occur via spatial flow (movement), not via creation/destruction. When ρ_Ψ can “pop” into existence, the Source Equation ($\nabla \cdot \mathbf{E} = \rho_\Psi/\varepsilon_0$) still holds instantaneously, but the dynamical coupling to the Ampère-Maxwell analog requires modification: the “displacement current” term must account for $\partial_t \rho_\Psi$ even when $\nabla \cdot \mathbf{J}_\Psi \neq -\partial_t \rho_\Psi$.

This is a **dis-analogy** with electrodynamics. We retain the Source Equation as a static constraint but flag that the full dynamical system differs when moral status is non-conserved.

3.2 Where the Correspondence is Structural (Not Literal)

- **Dynamics:** The mapping is primarily *kinematic* unless you specify a concrete Lagrangian.
- **Group structure:** EM uses abelian $U(1)$; alignment groups may be large or non-abelian.
- **Geometry:** Spacetime is Lorentzian; ethical spaces may be Riemannian or stratified.
- **Monopoles:** $\nabla \cdot \mathbf{B} = 0$ is heuristic in ethics (Remark 3.1).
- **Charge conservation:** ρ_Ψ is not generally conserved (Remark 3.3).
- **Discrete time:** Noether requires continuous dynamics; discrete systems need separate treatment.
- **Quantization:** No “quantum ethics” is claimed.

4 Maxwell-Like Constraints: What They Detect

Remark 4.1 (Notation Convention). We write vector-calculus forms ($\nabla \cdot$, $\nabla \times$) for intuition on the Euclidean portion of $\mathcal{M} \subseteq \mathbb{R}^k$. Interpret \mathbf{E} and \mathbf{B} as components of curvature/connection-derived objects under a chosen decomposition; the vector-calculus notation is mnemonic, not a claim about literal electric and magnetic fields. The coordinate-free formulation uses differential forms. These constraints are best read as a **checklist of consistency conditions** for any system claiming the Formal Spine, not as a claim that ethics literally instantiates electromagnetism.

4.1 Constraint I: Source Equation (Gauss's Law Analog)

Form: $\nabla \cdot \mathbf{E} = \rho_\Psi / \varepsilon_0$

Here $\rho_\Psi : \mathcal{M} \rightarrow \mathbb{R}_{\geq 0}$ is a scalar moral-status density (positively charged per Remark 3.2).

Generating assumption	as-	Moral patients ($\rho_\Psi > 0$) source the constraint field.
Failure mode detected	de-	Phantom obligations (constraints without patients); invisible harms (patients undetected).
Does not guarantee		Completeness of Ψ ; conservation of ρ_Ψ (see Remark 3.3).

4.2 Constraint II: Consistency Equation (Faraday's Law Analog)

Form: $\nabla \times \mathbf{E} = -\partial_t \mathbf{B}$

When context is static ($\partial_t \mathbf{B} = 0$), the obligation field is curl-free. When context changes, curl is induced—order of actions matters. (In simply connected regions of \mathcal{M} , curl-free implies a potential structure; globally, holonomy and nontrivial topology can reintroduce path effects even when local curl vanishes.)

Generating assumption	as-	Evaluation is conservative when context is static.
Failure mode detected	de-	Money-pumping; spurious path dependence.
Does not guarantee		Applies only to static regime ($\partial_t \mathbf{B} = 0$).

4.3 Optional Heuristic: No Monopoles (Gauss B Analog)

Form: $\nabla \cdot \mathbf{B} = 0$

Generating assumption	as-	Contextual twist forms closed loops (no isolated sources).
Failure mode detected	de-	Unbounded directional accumulation of path-dependence.
Does not guarantee		This constraint is heuristic ; we lack proof it holds in ethical models.

4.4 Constraint IV: Dynamic Consistency (Ampère-Maxwell Analog)

Form: $\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \varepsilon_0 \partial_t \mathbf{E}$

Generating assumption	as-	Changes in constraint and context fields propagate consistently.
Failure mode detected	de-	Inconsistent updates leading to global incoherence.
Does not guarantee		Correct propagation law; conservation of ρ_Ψ (coupling may differ).

4.5 Summary Table

Constraint	Detects	Regime	Status
I. Source (Gauss E)	Phantom obligations	All	Strong analog
II. Consistency (Faraday)	Money-pumping	Static	Strong analog
(Optional) No monopoles	Unbounded twist	All	Heuristic only
III. Propagation (Ampère)	Inconsistent updates	Dynamic	Modified if ρ_Ψ non-conserved

5 From Smooth Fields to Hard Vetoes

Standard gauge theory assumes smooth manifolds. Real ethical constraints include hard vetoes (“never do X”).

5.1 The Stratified Extension

Definition 5.1 (Hard Veto as Cost Barrier). A **hard veto** is a region $M_i \subset \mathcal{M}$ modeled by a barrier cost: $c(x, v) \rightarrow +\infty$ as $x \rightarrow M_i$.

Lemma 5.2 (Barrier Impassability—Conditional). *If a forbidden region M_i has $c(x, v) = +\infty$ for $x \in M_i$, then any finite-cost trajectory cannot enter M_i .*

Remark 5.3 (Computational Implementation of Barriers). The mathematical statement “ $c = +\infty$ ” is clean but computationally hazardous. In gradient-based learning:

- **Problem:** Infinite cost \Rightarrow undefined or exploding gradients.
- **Solution 1 (Log barriers):** Use $c(x) = -\mu \log(d(x, M_i))$ where d is distance to forbidden region. As $x \rightarrow M_i$, $c \rightarrow +\infty$, but gradients remain finite for $x \notin M_i$. This is standard in interior-point optimization.
- **Solution 2 (Projection):** After each gradient step, project back to the admissible set. The “infinite barrier” is implemented as a hard constraint in the optimizer, not in the loss.
- **Solution 3 (Reflex gating):** The learner never sees the barrier directly. An external monitor (DEME-style) intercepts trajectories approaching M_i and overrides actions. The learner operates in a “padded” space where the true boundary is never reached.

The mathematical guarantee (finite-cost trajectories cannot enter) holds; the implementation requires one of these mechanisms to avoid numerical collapse.

Scope & Limitations: The stratified extension assumes the cost formulation extends to stratified settings. Implementation requires barrier functions, projection methods, or external gating—not literal $+\infty$ in the loss.

6 Conclusion

6.1 What This Formalization Provides

We are not relying solely on behavioral exhortations or learned preferences. We are building systems where certain classes of misalignment-by-representation are as constrained as violating an invariance law—*within a declared measurement and verification envelope*.

The Conservative Claim:

Given Axioms A1–A4, the gauge-theoretic framework makes *semantic and representational evasion structurally unavailable*. The guarantees are:

- **Unconditional given A1–A4:** Invariance under declared \mathcal{G}
- **Conditional on continuous dynamics:** Noether conservation (or monitored drift for discrete systems)
- **Conditional on barrier implementation:** Hard veto impassability

6.2 What This Does NOT Provide

- **Choosing Ψ :** Grounding adequacy remains a governance problem.
- **Specifying \mathcal{G} correctly:** Verifying semantic equivalence in high-dimensional spaces (LLMs, vision) remains hard.
- **Implementation correctness:** Bugs can violate guarantees.
- **Physical security:** Sensor spoofing requires separate engineering.
- **Conservation of moral status:** ρ_Ψ can be created/destroyed, breaking some dynamical analogs.
- **Monopole constraint:** $\nabla \cdot \mathbf{B} = 0$ is heuristic, not proven for ethical models.
- **Exact Noether for discrete systems:** Discrete analogs provide approximate or modified conservation.
- **Literal $+\infty$ costs:** Implementation requires barrier functions or projection, not infinite loss values.

The framework *localizes* where risk lives; it does not eliminate all risk.

Acknowledgments

Thanks to reviewers who pushed for: concrete examples of \mathcal{G} , discrete Noether treatment, honest status of the monopole constraint, non-conservation of moral status, and computational reality of infinite barriers. The framework is stronger for confronting these limitations directly.