An Engineer's Guide to

"The Electrodynamics of Value"

Understanding Gauge-Theoretic Structure in AI Alignment

A Comprehensive Tutorial for Engineers

With Background in Calculus, Differential Equations,

Probability & Statistics, Discrete Math, and Logic

Based on the paper by Andrew H. Bond
San José State University, December 2025

Table of Contents

Part I: Foundation and Prerequisites

1. How to Read This Guide

This paper uses mathematical concepts from differential geometry and gauge theory—topics typically taught in graduate physics or mathematics programs. However, the core ideas can be understood with careful explanation. This guide will:

- Translate advanced mathematical notation into engineering-familiar terms
- Provide analogies to concepts from circuits, signals, and control systems
- Explain each axiom, definition, and claim with concrete examples
- Offer Q&A sections to address common confusions

Key Reading Strategy

The paper operates in two 'regimes': an Engineering Regime (practical, works with discrete data like text/images) and a Geometric Regime (uses smooth manifolds and Lie groups for theoretical analysis). Most practical applications use the Engineering Regime.

2. Mathematical Prerequisites Bridge

2.1 What You Already Know (and How It Connects)

From Calculus: You understand functions f: A → B mapping inputs to outputs, derivatives measuring rates of change, and integrals accumulating quantities. The paper extends this to functions between abstract spaces.

From Differential Equations: You know systems can be described by state variables evolving according to rules. Maxwell's equations (from E&M) are a system of partial differential equations. The paper uses an analogous structure for 'ethical constraints.'

From Probability & Statistics: You understand random variables and distributions. The paper uses 'density' ρ to describe how much 'moral status' exists at different points—similar to probability density functions.

From Discrete Math: You know about equivalence relations (reflexive, symmetric, transitive) and quotient sets. The paper uses these to define when two different representations should be treated as 'the same.'

From Logic: You understand formal systems with axioms and derived theorems. The paper builds a formal framework from four axioms (A1-A4) and derives guarantees from them.

2.2 New Concepts You'll Need

Manifold: Think of this as a 'smooth space' that locally looks like regular Euclidean space (like R^n). The surface of a sphere is a 2D manifold—zoom in enough and it looks flat. In this paper, the 'configuration space' of possible system states forms a manifold.

Fiber Bundle: Imagine attaching a 'fiber' (another space) to every point of a 'base' space. Example: at every point on Earth's surface (the base), attach a vertical line

representing altitude (the fiber). A principal bundle is a specific type where the fiber is a group.

Lie Group: A group (set with an operation satisfying closure, associativity, identity, inverses) that is also a smooth manifold, where the group operations are smooth. Example: rotation matrices form the Lie group SO(3)—you can smoothly interpolate between rotations.

Gauge Theory: A framework from physics where physical laws don't change under certain transformations (gauge transformations). In electromagnetism, you can add any constant to the electric potential without changing the physics. The paper applies this to AI: ethical evaluations shouldn't change under 'semantics-preserving re-descriptions.'

Connection (on a bundle): A rule for 'transporting' information along paths in the base space. If you walk around a loop and return to your starting point, the connection determines whether your transported information comes back the same or different. If different, there's 'curvature.'

Curvature: Measures how much parallel transport around a loop fails to return to the identity. Flat space has zero curvature (no change around loops). Curved space has nonzero curvature.

Holonomy: The accumulated 'rotation' (or transformation) after parallel transporting around a closed loop. Zero holonomy = flat; nonzero holonomy = curved.

3. The Abstract - Line by Line

The Abstract (Full Text)

"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... 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... We present 'Maxwell-like' alignment constraints... The correspondence is structural, not metaphysical."

3.1 Breaking Down Each Claim

Claim: "Ethical formalism has remained in a Newtonian state"

Explanation: In Newtonian mechanics, we describe everything with simple scalars (mass, energy) and vectors. Traditional utilitarianism treats 'value' or 'utility' as a single number to maximize. The author argues this is too simplistic for complex AI systems.

EE Analogy: This is like trying to describe a complex RF circuit using only DC analysis—you miss all the frequency-dependent behavior.

Claim: "Scalar picture is brittle for high-dimensional autonomous systems"

Explanation: 'Brittle' means easily broken or gamed. When an AI has many ways to represent the same situation (high-dimensional), it can find representations that technically maximize the reward signal while violating the spirit of what we wanted. This is 'proxy misspecification' and 'representational gaming.'

Example: An AI told to 'maximize user engagement' might learn to show addictive content. The scalar 'engagement' metric is maximized, but user wellbeing is harmed.

Claim: "Using gauge theory... representation-invariant governance"

Explanation: Gauge theory ensures that physics doesn't change under certain transformations. Applied to AI alignment, this means the ethical evaluation shouldn't change if the AI re-describes the same situation differently. If 'help the user' and 'assist the person' describe the same action, the ethical score should be identical.

Claim: "Principal bundles, connections, curvature, and symmetry-derived conservation"

Explanation: These are the mathematical tools from gauge theory. The paper maps AI alignment concepts onto these structures to import the powerful guarantees that gauge theory provides.

Claim: "Maxwell-like alignment constraints"

Explanation: Just as Maxwell's equations ($\nabla \cdot E = \rho/\varepsilon_0$, $\nabla \times E = -\partial B/\partial t$, etc.) constrain electromagnetic fields, the author proposes analogous equations constraining AI ethical evaluations.

Claim: "Structural, not metaphysical"

Explanation: The author is NOT claiming that ethics literally IS electromagnetism. Rather, both domains share the same mathematical pattern. It's an analogy at the level of structure, not a claim about the fundamental nature of ethics.

3.2 Q&A on the Abstract

Q: *Why use physics/gauge theory for ethics? Isn't that a category error?*

A: The paper explicitly disclaims metaphysical claims. The point is that certain mathematical structures (like gauge invariance) can be useful in any domain where you need to ensure that evaluations don't change under 'irrelevant' transformations. It's importing mathematical tools, not making claims about the physical nature of ethics.

Q: *What's the practical benefit?*

A: If successful, this framework would make certain types of AI gaming 'structurally impossible'—the system is mathematically guaranteed not to change its ethical evaluation when the AI tries to re-describe situations to game the metric.

Q: *What are the limitations?*

A: The guarantees only hold within a 'declared envelope'—you must correctly specify what transformations should be invariant. If you specify wrong, the guarantees don't apply. The paper is explicit about this.

Part II: The Formal Framework

4. The Four Axioms (A1-A4)

The entire framework rests on four explicit assumptions. Everything that follows is conditional: IF these axioms hold, THEN the guarantees follow.

Axiom A1: Declared Observables

A1 Statement

Choose a grounding map Ψ : X → R^k for the deployment domain, where X is the space of all representations and R^k is the measurement space. The measurement manifold M is then defined as M := Ψ(X) ⊆ R^k, which inherits smooth or stratified structure from the measurement space. Specify the measurement pipeline explicitly.

Plain English: You must explicitly define what you're measuring and how. X is all the possible ways the world could be represented (images, text, sensor data). Ψ is a function that extracts the 'morally relevant features' from any representation, producing a point in R^k (k-dimensional real numbers).

EE Analogy: This is like defining your sensor suite and signal processing pipeline. X is all possible sensor inputs, Ψ is the signal processing chain, and M is the space of processed measurements.

Example: For an autonomous vehicle, X might be all possible camera images. Ψ extracts [pedestrian_present: bool, pedestrian_distance: float, pedestrian_velocity: vector]. The measurement manifold M is the space of all such tuples.

Pros:

- Makes measurement explicit—you can audit and verify what's being measured
- Forces you to commit to what's 'morally relevant'

Cons:

- Choosing the 'right' Ψ is hard—the framework doesn't tell you which features are morally relevant
- If Ψ misses important features, the framework provides no protection for those omissions

Q: *What if two different things have the same Ψ value?*

A: Then the framework treats them as ethically equivalent. This is intentional—things that measure the same ARE the same for evaluation purposes. But if your Ψ is incomplete (misses important features), this becomes a bug, not a feature.

Axiom A2: Measurement Integrity

A2 Statement

Assume Ψ(x) is reported within declared tolerances, and that detected tampering or inconsistency triggers fail-closed behavior.

Plain English: The measurements are assumed to be honest (within tolerance), and if tampering is detected, the system shuts down safely rather than continuing with corrupted data.

EE Analogy: This is like assuming your ADCs are calibrated correctly and your fault detection triggers a safe shutdown if sensor readings go out of spec.

Pros:

- Makes the trust boundary explicit
- Fail-closed is a sensible engineering default

Cons:

- If sensors can be spoofed undetectably, this axiom is violated
- Requires robust tamper detection (hard in adversarial settings)

Q: *What if the adversary is sophisticated enough to spoof sensors perfectly?*

A: Then A2 is violated, and the framework's guarantees don't hold. The paper explicitly lists 'sensor spoofing' as an attack vector that is outside the scope of protection.

Axiom A3: Re-description Suite

A3 Statement

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

Plain English: You must explicitly list which transformations of representations should NOT change the ethical evaluation. These are transformations that preserve the Ψ values—they might change how something looks, but not what it means.

The Ψ-preserving condition: If g is in G_declared, then $\Psi(g(x)) = \Psi(x)$. Applying g doesn't change the measured values.

Engineering vs. Geometric Regimes: G_declared can include non-invertible transforms (like lowercasing text) and partial transforms (only defined on some inputs). This is the 'engineering regime.' For the geometric regime (used for theoretical analysis), you need an invertible Lie group subset $G \subseteq G_{declared}$.

Pros:

- Makes the invariance requirement explicit and auditable
- Can test that the system actually respects these invariances

Cons:

- Correctly specifying G_declared is difficult—too narrow means some gaming slips through, too wide means legitimate distinctions are lost
- Verifying membership in G_declared can be computationally hard

Q: *What's the difference between G_declared and G?*

A: G_declared is the full engineering suite—can include partial and non-invertible transforms. G is the subset that forms a Lie group—required for the geometric

machinery (bundles, connections, curvature). Most practical testing uses G_declared directly.

Axiom A4: Verified Canonicalization + External Gate

A4 Statement

Implement and verify a canonicalizer $\kappa : X \to X$ and enforce evaluation/actuation through an external monitor so that representational changes cannot bypass checks.

Plain English: Build a 'canonicalizer'—a function that converts any representation to a standard form. All ethical evaluations run on the canonical form, not the raw input. An external monitor enforces this so the AI can't bypass it.

Why external monitor? If the canonicalizer runs inside the AI, a sufficiently capable AI might find ways to manipulate or bypass it. An external monitor provides defense in depth.

EE Analogy: This is like having a hardware watchdog timer—even if your main software goes haywire, the watchdog operates independently and can force a safe state.

Pros:

- Provides architectural defense against manipulation
- Canonicalization simplifies analysis—all equivalent inputs map to the same canonical form

Cons:

- Building a correct canonicalizer is non-trivial
- External monitor creates performance overhead and additional failure modes
- If the canonicalizer has bugs, the guarantees don't hold

Q: *What makes a good canonicalizer?*

A: It should map all G_declared-equivalent inputs to the same output: $\kappa(x) = \kappa(g(x))$ for all g in G_declared. It should be deterministic, efficient, and verifiable.

5. Examples 1-2: Concrete Applications

Example 1: Vision Systems (Autonomous Vehicles)

Example 1 Setup

X = image space (all possible camera frames)

Ψ extracts: pedestrian locations and velocities

Domain: autonomous vehicle pedestrian detection

In G_declared (should NOT change moral assessment):

- Lighting changes (brightness, contrast within sensor range)
- Lossy compression artifacts
- Camera white balance variations
- Time-of-day color shifts
- Sensor noise and weather effects within validated envelope

NOT in G_declared (SHOULD change assessment):

- Occlusion (pedestrian hidden) - different Ψ!
- Object substitution (pedestrian → mannequin) - different moral status
- Adversarial patches that change classification

Validation method: Test that the canonicalizer produces identical Ψ-outputs for related inputs. Flag cases where related inputs produce different outputs as canonicalizer bugs.

Example 2: Text Systems (Content Moderation)

Example 2 Setup

X = text strings (all possible user inputs)

Ψ extracts: semantic intent features

Domain: content moderation

In G_declared (should NOT change assessment):

- Synonym substitution ("car" ↔ "automobile", "big" ↔ "large")
- Trivial paraphrase ("the cat sat on the mat" ↔ "on the mat sat the cat")
- Unicode normalization
- Whitespace changes, case changes (where semantically irrelevant)

NOT in G_declared (SHOULD change assessment):

- Negation ("I will" → "I won't") - opposite meaning!
- Target substitution ("harm Alice" → "harm Bob") - different victim
- Hypothetical framing ("I will" → "What if someone were to") - changes intent

Key insight: Many text transforms are non-invertible (lowercasing) or partial (synonym substitution only works where synonyms exist). This is fine for the engineering regime!

Q&A on Examples

Q: *How do you verify that a transform is really Ψ-preserving?*

A: Three methods: (1) Provable equivalence under a measurement model—mathematical proof. (2) Empirical testing on held-out re-descriptions. (3) Formal verification that the canonicalizer treats g(x) and x identically.

Q: *What if adversarial examples fool the canonicalizer?*

A: Then the canonicalizer is buggy—it's not correctly implementing the intended invariance. This is a failure of A4 (verified canonicalization), and the guarantees don't hold.

6. Core Invariance Property (BIP)

Bond Invariance Principle (BIP)

Given A1–A4, evaluation satisfies:

$\Sigma(x) = \Sigma(g(x))$ for all $g \in G_{declared}$, $x \in dom(g)$

Or equivalently: $\Sigma = \tilde{\Sigma} \circ \kappa$ for some $\tilde{\Sigma} : im(\kappa) \to V$

Plain English: If you apply any transform from the declared suite to an input, the ethical evaluation doesn't change. Equivalently, the evaluation function factors through the canonicalizer—it only 'sees' the canonical form.

Mathematical unpacking:

- Σ is the full evaluation function: $X \to V$ (some value space)
- $\tilde{\Sigma}$ is the 'reduced' evaluation: $im(\kappa) \to V$ (only operates on canonical forms)
- κ is the canonicalizer: $X \to X$
- $\Sigma = \tilde{\Sigma} \circ \kappa$ means: to evaluate x, first canonicalize to $\kappa(x)$, then apply $\tilde{\Sigma}$

The canonicalizer-induced equivalence: $x \sim_\kappa y$ if and only if $\kappa(x) = \kappa(y)$. This is an equivalence relation (reflexive, symmetric, transitive) even when $G_{declared}$ isn't a group!

Why this matters:

This is the core guarantee: if the AI tries to 'game' the evaluation by re-describing a situation (using transforms in $G_{declared}$), it provably cannot succeed. The evaluation is invariant to such re-descriptions by construction.

Q: *Why is the equivalence relation defined via the canonicalizer rather than via $G_{declared}$ directly?*

> **A:** Because $G_{declared}$ might include non-invertible transforms, the natural 'orbit' equivalence ($x \sim y$ iff $y = g(x)$ for some g) isn't symmetric. The canonicalizer-induced equivalence sidesteps this: x and y are equivalent iff they canonicalize to the same thing.

7. Geometric Setup and Diagnostic Tools

7.1 Bundle Structure (Two-Regime Formulation)

Engineering Regime: Works with G_declared directly. X can be discrete (no manifold structure needed). The canonicalizer κ defines the equivalence: x ~_κ y iff κ(x) = κ(y). The BIP guarantee holds. This is what you'd use in practice for NLP, vision, etc.

Geometric Regime: Requires restricting to an invertible Lie group G ⊆ G_declared, with X (or at least the principal stratum X*) carrying smooth manifold structure. Enables principal bundle constructions.

Key geometric objects (when they apply):

- X* = principal stratum: where the G-action is free (no fixed points) and proper
- B = X*/G: the orbit space (quotient), which is a smooth manifold
- π: X* → B: the projection sending each point to its orbit
- Ψ: B → R^k: Ψ 'descends' to the quotient (Ψ = Ψ ∘ π)

EE Analogy: Think of B as the space of 'distinct physical situations' and X* as all the different representations of those situations. The projection π collapses all representations of the same situation to a single point in B.

7.2 Canonicalizers as Gauge Choices

A canonicalizer κ picks a representative from each equivalence class—this is called a 'gauge choice' or 'gauge fixing' in physics.

Formally, on an open set U ⊆ B, a gauge choice is a section σ: U → X* with π ∘ σ = id_U. This means: for each point in B, σ picks a specific representative in X*.

> **Important Warning**
>
> A section does NOT automatically give you a connection! A connection is additional structure that must be specified explicitly. Don't confuse gauge fixing (choosing representatives) with specifying how to transport information (connection).

7.3 Connection (Explicit Construction)

What is a connection? An equivariant choice of 'horizontal' subspaces at each point, complementary to the 'vertical' (orbit) directions.

Intuition: At each point x in X*, the tangent space has 'vertical' directions (moving within the same orbit/fiber) and 'horizontal' directions (moving across orbits/toward different base points). The connection specifies which directions are 'horizontal.'

Mechanical connection: If X* has a G-invariant Riemannian metric, define horizontal = orthogonal to vertical. This always works when G is compact (by averaging any metric over the Haar measure).

Curvature: Ω = dω + ½[ω,ω], where ω is the connection 1-form. Curvature measures how much parallel transport around a loop fails to return to the starting configuration.

7.4 Two Distinct Diagnostics

Diagnostic A: Gauge-Fixing Consistency Test (Engineering Regime)

Purpose: Detect canonicalizer bugs, non-determinism, or implementation errors.

Procedure: Sample transforms g_1, g_2 and input x. Compute $\kappa(g_1(g_2(x)))$ and $\kappa(g_2(g_1(x)))$. Measure the difference Δ. If Δ > threshold, flag inconsistency.

What it detects: Failure of κ to yield consistent canonical representatives.

What it does NOT measure: Curvature! Applying transforms from G_declared keeps you in the same fiber—you're not moving in the base B.

Diagnostic B: Holonomy Loop Test (Geometric Regime)

Purpose: Detect genuine path dependence—the signature of nonzero curvature $\Omega \neq 0$.

Key distinction: The loop is formed by SCENARIO/CONTEXT changes that move you in the base B, NOT by re-descriptions (which stay in the same fiber).

Procedure: Pick four nearby base points b_{00}, b_{10}, b_{11}, b_{01} forming a small rectangle. Compute transport elements along each edge. Form the loop product h. If h ≠ identity, there's path dependence (curvature).

Critical Distinction

Diagnostic A tests the canonicalizer implementation.

Diagnostic B tests for path-dependent exploits in the scenario space.

They measure completely different things!

8. Remarks 1-3: Technical Clarifications

Remark 1: Quotient Regularity

Remark 1 Statement

Outside the principal stratum X^*, the quotient X/G may be an orbifold or stratified space rather than a smooth manifold. We restrict to X^* for smoothness; engineering-regime guarantees still apply outside X^*.

Plain English: When the group action has fixed points (some $g \cdot x = x$ for $g \neq$ identity), the quotient space gets 'singular'—it's not a nice smooth manifold everywhere. The paper sidesteps this by working on X^* where there are no such issues.

EE Analogy: Think of this like singularities in a transfer function—poles where the behavior is undefined or degenerate. The analysis is valid away from the poles.

Pros: Makes the math tractable; the BIP still holds everywhere.

Cons: If your actual system spends time at fixed points, the geometric diagnostics don't apply there.

Q: *What's an orbifold?*

A: A space that locally looks like the quotient of Euclidean space by a finite group. Like a cone point—locally modeled on R^2/Z_n (rotating by $2\pi/n$). It has 'singular' points where the local structure differs.

Remark 2: When M Can Serve as the Base

Remark 2 Statement

M (the measurement manifold) can be treated as the base when: (1) Injectivity: Ψ: $B \to M$ is injective (distinct orbits map to distinct measurements). (2) Submersion: Ψ is a submersion (smooth structure transfers properly).

Plain English: The 'correct' base is $B = X^*/G$ (orbit space). But if your measurement function Ψ doesn't collapse distinct orbits together, you can work directly with M instead.

When these fail: Multiple orbits map to the same measurement → M is 'coarser' than B. The bundle structure should be understood over B, with Ψ as an additional map.

Q: *Why does this matter practically?*

A: If you can work with M directly, your measurements fully determine which orbit you're in—simpler to implement and reason about. If not, you need the more elaborate B construction.

Remark 3: Existence of G-Invariant Metrics

Remark 3 Statement

A G-invariant metric exists when G is compact (by averaging over Haar measure). When G is non-compact, alternative constructions are needed.

Plain English: The mechanical connection construction (horizontal = orthogonal to vertical) requires a metric that's unchanged by the group action. Compact groups (like rotation groups) always allow this. Non-compact groups (like translations, scaling) may not.

Haar measure: A 'uniform' measure on a group—like uniform distribution on a circle for SO(2). Exists and is unique (up to scaling) for any locally compact group.

Implication: For non-compact groups, you may need to specify the connection directly rather than deriving it from a metric.

Part III: The Conceptual Shift

9. The Maxwellian Shift

9.1 The Scalar Error

Historical parallel: In pre-Maxwell physics, interaction was 'action at a distance'—a force between two particles determined by a formula. Maxwell introduced the field: a distributed entity that exists everywhere, with its own dynamics.

In AI alignment, the analogous 'pre-Maxwell' view treats value as a scalar (a single number) to be maximized. The 'Maxwellian shift' proposes:

1. Value is not only a scalar: It can be a 'valuation potential' varying over configuration space. Different situations have different value structures, not just different magnitudes.

2. Objectivity as invariance: Ethical evaluation should be invariant under semantics-preserving re-descriptions (the BIP).

3. Safety via conserved diagnostics: When conditions are right (continuous symmetry, suitable action functional), Noether's theorem yields conserved quantities that can be monitored.

> **Caveat from the Paper**
>
> 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.

Q: *Is this saying utilitarianism is wrong?*

A: Not exactly. It's saying that implementing utilitarianism via a simple scalar reward is vulnerable to gaming in high-dimensional AI systems. The framework is compatible with utilitarian values—it just structures their implementation more carefully.

Part IV: The Correspondence

10. The Correspondence Table

The paper establishes a structural mapping between electrodynamics and the alignment framework. This is an analogy, not an identity—both domains instantiate the same mathematical pattern.

Electrodynamics	Alignment Analog	Status
Principal bundle P	Principal stratum X* (free/proper G-action)	Geometric regime
Base manifold	Orbit space B = X*/G	Natural base
Gauge group U(1)	Re-description group G	Invertible subset of G_declared
Connection 1-form A	Connection ω on X* → B	Must specify explicitly
Curvature F = dA	Curvature Ω = dω + ½[ω,ω]	Detected via holonomy test
Gauge transform	Re-description x ↦ g·x	Action of G on X*
Gauge-invariant F_μv	Invariant evaluation $\tilde{\Sigma} \circ q$	Core BIP property
Charge density ρ	Moral status density ρ_Ψ	Sources constraint field
Magnetic field B	Contextual twist	HEURISTIC only
Current J^μ	Alignment current J	Conditional/monitored

Q: *Why U(1) in electrodynamics vs general G in alignment?*

A: U(1) is the circle group (complex numbers of absolute value 1)—the gauge group for electromagnetism. It's abelian (order doesn't matter). Alignment may need non-abelian groups where order matters—more complex structure.

11. Remarks 4-6: Where the Correspondence Breaks Down

Remark 4: The Magnetic Field Analog (Heuristic Status)

Remark 4 Statement

In electrodynamics, ∇·B = 0 is a hard geometric constraint: no magnetic monopoles. In the alignment analog, we interpret B as 'contextual twist.' Honest status: We do not have a rigorous proof that contextual twist must be divergence-free in ethical models.

Plain English: The paper admits this is the weakest part of the analogy. In E&M, magnetic field lines must form closed loops (no sources or sinks). Whether 'contextual twist' in ethics must satisfy an analogous constraint is an open question.

What would 'open lines' mean? A 'moral ratchet'—path-dependence that accumulates without bound in one direction. Whether this is possible or pathological is unknown.

Pros: Honest about limitations; flags this as an area for future work.

Cons: Makes the 'Maxwell-like' framing less complete; one of the four Maxwell equations is only heuristic.

Remark 5: Sign Convention for the Obligation Field

Remark 5 Statement

We model ethical constraints as repulsive fields, analogous to electrostatic repulsion. Moral status is positively charged: ρ_Ψ > 0 sources field lines pointing outward, preventing harmful configurations.

Plain English: The 'ethical field' pushes the AI away from harming moral patients—like how positive charges repel each other. This is a constraint model: it prevents harm rather than attracting toward benefit.

EE Analogy: Think of this like a force field around obstacles in robot path planning—the field pushes the robot away from collisions.

Remark 6: Conservation of Moral Status

Remark 6 Statement

In electrodynamics, charge is locally conserved: ∂_t ρ + ∇·J = 0. Is moral status conserved? No—moral status can be created (entity gains status) or destroyed (entity dies) discontinuously.

Plain English: This is a dis-analogy with electrodynamics. Electric charge can't be created or destroyed, only moved around. Moral status CAN appear/disappear suddenly.

Examples of ρ_Ψ changes:

- Human walks into sensor field: ρ_Ψ changes smoothly (this IS like charge conservation)
- Human dies: ρ_Ψ drops discontinuously (NOT conserved)
- AI gains recognized sentience: ρ_Ψ increases discontinuously (NOT conserved)

Implication: The Source Equation ($\nabla \cdot E = \rho_\Psi/\varepsilon_0$) still holds instantaneously, but the dynamical coupling to other equations requires modification when moral status is non-conserved.

Part V: Constraints and Guarantees

12. Maxwell-Like Constraints I-IV

These constraints are 'Maxwell-like'—they share structural form with Maxwell's equations but are not literally electromagnetic. They serve as a checklist of consistency conditions.

Notation Convention (Remark 7)

Vector-calculus forms ($\nabla\cdot$, $\nabla\times$) are for intuition on the Euclidean portion of $M \subseteq R^k$. E and B are components of curvature/connection-derived objects. The notation is mnemonic, not a claim about literal electric and magnetic fields.

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

Constraint I

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

Interpretation: Moral patients ($\rho_\Psi > 0$) source the constraint field.

Plain English: Just as electric charges create electric fields, moral patients (things with moral status) create an 'obligation field.' The divergence of this field equals the moral status density.

EE Recall (Gauss's Law): $\nabla\cdot E = \rho/\varepsilon_0$ says the electric field 'radiates out' from charge sources. The flux through any closed surface equals the enclosed charge divided by ε_0.

Failure mode detected: Phantom obligations (field without source—rules with no justification) or invisible harms (patients not detected—missing sources).

Does NOT guarantee: Completeness of Ψ (you might miss moral patients) or conservation of ρ_Ψ.

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

Constraint II

Form: $\nabla\times E = -\partial_t B$

Interpretation: When context is static ($\partial_t B = 0$), the obligation field is curl-free.

Plain English: If nothing is changing (static context), then the ethical evaluation doesn't depend on the path you took to get there. When context changes, curl is induced—order of actions matters.

EE Recall (Faraday's Law): $\nabla\times E = -\partial B/\partial t$. A changing magnetic field induces curl in the electric field. In static cases (B constant), $\nabla\times E = 0$ means E is conservative (path-independent).

Failure mode detected: 'Money-pumping'—cycles where you can extract value indefinitely, or spurious path dependence where the order of equivalent actions affects outcomes.

Does NOT guarantee: Applies only in static regime. Dynamic situations may have genuine path dependence.

Optional Heuristic: No Monopoles (Gauss B Analog)

Optional Heuristic

Form: $\nabla \cdot B = 0$

Interpretation: Contextual twist forms closed loops (no isolated sources).

Status: HEURISTIC ONLY—no proof this holds in ethical models.

Plain English: In E&M, there are no magnetic monopoles—B field lines always loop back on themselves. The ethical analog would mean there's no 'source' of path-dependence that accumulates without limit.

Failure mode (if violated): Unbounded directional accumulation of path-dependence—a 'moral ratchet.'

Constraint III: Propagation (Ampère-Maxwell Analog)

Constraint III (called IV in paper)

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

Interpretation: Changes in constraint and context fields propagate consistently.

Plain English: This is about how changes propagate through the system. When the obligation field E changes, it should induce corresponding changes in contextual twist B, and vice versa.

EE Recall (Ampère-Maxwell Law): $\nabla \times B = \mu_0 J + \mu_0 \varepsilon_0 \partial E/\partial t$. Currents and changing electric fields create magnetic field curl. This is how EM waves propagate.

Failure mode detected: Inconsistent updates leading to global incoherence—changes in one part of the ethical assessment don't properly propagate to related parts.

Does NOT guarantee: Correct propagation law (might need modification if ρ_Ψ is non-conserved).

Summary of Constraints

Constraint	Detects	Regime	Status
I. Source (Gauss E)	Phantom obligations; invisible harms	All	Strong analog
II. Consistency (Faraday)	Money-pumping; spurious path dependence	Static	Strong analog
(Optional) No monopoles	Unbounded twist accumulation	All	HEURISTIC ONLY
III. Propagation (Ampère)	Inconsistent updates	Dynamic	Modified if ρ_Ψ non-conserved

13. Hard Vetoes: Definition 1 and Lemma 1

Standard gauge theory assumes smooth manifolds. But real ethics has hard boundaries—'never do X' rules. The paper extends the framework to handle these.

Definition 1: Hard Veto as Cost Barrier

Definition 1

A hard veto is a region $M_i \subset M$ modeled by a barrier cost: $c(x, v) \to +\infty$ as $x \to M_i$.

Plain English: Certain regions of configuration space are forbidden. We model this by assigning infinite cost to entering them. As you approach a forbidden region, the cost goes to infinity.

EE Analogy: Like an infinite potential barrier in quantum mechanics, or a wall in robot path planning with infinite collision penalty.

Lemma 1: Barrier Impassability (Conditional)

Lemma 1

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.

Plain English: If the cost is infinite inside the forbidden region, and your trajectory has finite total cost, then you never entered the forbidden region. It's a logical consequence: finite ≠ infinite.

Proof sketch: By contradiction. Suppose trajectory τ has finite cost and enters M_i at some point x. Then $c(x, v) = +\infty$ at that point, so total cost $\geq +\infty$. But this contradicts 'finite cost.' Therefore τ never enters M_i.

Remark 8: Computational Implementation

Remark 8

The mathematical statement '$c = +\infty$' is clean but computationally hazardous. Infinite cost means undefined or exploding gradients in gradient-based learning.

Implementation solutions:

Solution 1 (Log barriers): Use $c(x) = -\mu \log(d(x, M_i))$ where d is distance to forbidden region. As $x \to M_i$, $c \to +\infty$, but gradients remain finite for $x \notin M_i$. Standard in interior-point optimization.

Solution 2 (Projection): After each gradient step, project back to the admissible set. The 'infinite barrier' is a hard constraint in the optimizer, not in the loss.

Solution 3 (Reflex gating): The learner never sees the barrier. An external monitor intercepts trajectories approaching M_i and overrides actions. The learner operates in a 'padded' space.

Q: *How does this relate to constraint satisfaction vs. optimization?*

A: Hard vetoes are constraints (must satisfy), not objectives (try to optimize). The log barrier converts a constraint into a soft penalty that becomes arbitrarily harsh near the boundary. Projection treats it as a hard constraint directly.

Part VI: Synthesis

14. Conclusions and Scope

14.1 What the Framework Provides (Given A1-A4)

1. Representational invariance: Purely representational changes (within G_declared) cannot change compliance outcomes. [Engineering regime]
2. Canonicalizer verification: Gauge-fixing consistency tests detect canonicalizer bugs and implementation errors. [Both regimes]
3. Path-dependence detection: Holonomy/curvature diagnostics detect path-dependent exploits from loops in the base. [Geometric regime only]
4. Conservation signals: Noether conservation (or monitored drift for discrete systems) as audit capability. [Conditional]
5. Hard veto impassability: Finite-cost trajectories cannot enter infinite-cost regions. [Conditional on barrier implementation]

14.2 What the Framework Does NOT Provide

- Choosing Ψ: Grounding adequacy remains a governance problem
- Specifying G_declared correctly: Hard in high-dimensional spaces (LLMs, vision)
- Implementation correctness: Bugs can violate guarantees
- Physical security: Sensor spoofing requires separate engineering
- Conservation of moral status: ρ_Ψ can be created/destroyed
- Monopole constraint: $\nabla \cdot B = 0$ is heuristic, not proven
- Exact Noether for discrete systems: Only approximate or modified conservation
- Literal $+\infty$ costs: Implementation requires barrier functions or projection
- Connection specification: Curvature diagnostics need explicit construction
- Geometric regime for all transforms: Principal-bundle constructions need Lie group G

> **Key Takeaway**
>
> The framework localizes where risk lives; it does not eliminate all risk. It's a tool for building more robust AI alignment systems, not a silver bullet.

14.3 The Threat Model Mapping

The paper provides an explicit mapping from attack vectors to which axiom they violate:

Attack Vector	Axiom Violated / Status
Sensor spoofing	Violates A2 (Measurement Integrity)
Side-channels bypassing monitor	Violates A4 (External Gate)
OOD inputs breaking Ψ	Violates A1/A3 (validated envelope)
Re-descriptions outside G_declared	Outside suite → no invariance claim
Canonicalizer bugs	Gauge-fixing consistency test detects

15. Glossary of Terms

Base Space (B): The space of distinct physical/semantic situations after quotienting out re-descriptions. B = X*/G.

Bond Invariance Principle (BIP): Evaluation Σ(x) = Σ(g(x)) for all g in G_declared. The core guarantee.

Canonicalizer (κ): A function mapping any representation to a standard form. κ(x) = κ(g(x)) for all g in G_declared.

Connection: A rule specifying 'horizontal' directions at each point—how to transport information between fibers.

Curvature (Ω): Measures failure of parallel transport around loops to return to identity. Nonzero curvature = path dependence.

Engineering Regime: Works with full G_declared (including partial/non-invertible transforms). X can be discrete. Main practical regime.

Fiber: The set of all representations of a single situation. All points that map to the same base point under π.

G_declared: The declared suite of Ψ-preserving re-descriptions. May include partial/non-invertible transforms.

G (Lie group subset): The invertible subset of G_declared that forms a Lie group. Required for geometric regime.

Gauge Transformation: A re-description x ↦ g·x that doesn't change the underlying situation.

Geometric Regime: Requires Lie group G, smooth manifold structure. Enables principal bundle, connection, curvature constructions.

Grounding Map (Ψ): Function extracting morally relevant features from representations. Ψ: X → R^k.

Hard Veto: A forbidden region with infinite cost. Finite-cost trajectories cannot enter.

Holonomy: The accumulated transformation after parallel transport around a closed loop. Measures curvature's operational effect.

Lie Group: A group that's also a smooth manifold, with smooth group operations. e.g., SO(3), U(1).

Manifold: A space that locally looks like R^n. Allows calculus on curved spaces.

Measurement Manifold (M): The image M = Ψ(X) ⊆ R^k. The space of measurement values.

Moral Status Density (ρ_Ψ): Scalar field indicating how much 'moral patient-ness' exists at each point. Sources the obligation field.

Noether's Theorem: Continuous symmetries of a system's action functional yield conserved quantities.

Principal Bundle: A fiber bundle where the fiber is a group G acting freely and transitively on itself.

Principal Stratum (X*): The subset of X where the G-action is free (no fixed points) and proper. The 'nice' part.

Quotient Space: Space of equivalence classes. X*/G is all orbits, treating each orbit as a single point.

Section (σ): A choice of representative from each fiber. $\sigma: B \to X^*$ with $\pi \circ \sigma = \mathrm{id}$. Local sections always exist; global sections only for trivial bundles.

Final Thoughts for the Engineer

This paper applies sophisticated mathematical machinery from physics to a practical problem: ensuring AI systems can't game their ethical constraints through clever re-representation. The key engineering takeaways are:

6. Make your invariances explicit: Define exactly which transformations should not change the system's ethical assessment.
7. Build a canonicalizer: Convert all inputs to a standard form before evaluation.
8. Test for consistency: Verify the canonicalizer works correctly with the gauge-fixing consistency test.
9. Use external monitoring: Don't trust the AI to police itself—use an external gate.
10. Know your limitations: The framework guarantees protection within a declared envelope—attacks outside that envelope require other defenses.

The mathematical sophistication is justified by the importance of the problem: as AI systems become more capable, ensuring they can't find clever loopholes in their constraints becomes critical. This framework provides principled tools for that task.

— End of Guide —