The Geometry of Good

A New Mathematical Framework for AI Safety

Based on the technical paper:
"The Electrodynamics of Value: Gauge-Theoretic Structure in AI Alignment"

Andrew H. Bond
San José State University

December 2025

Executive Summary

Current approaches to AI safety often treat "human values" as a single number—a score to maximize. This is like trying to describe the weather with just one temperature reading: technically possible, but dangerously incomplete.

This whitepaper introduces a fundamentally different approach, one borrowed from physics. Just as James Clerk Maxwell revolutionized our understanding of electricity and magnetism by showing they were aspects of a unified *field*, we propose treating AI alignment not as a number to optimize, but as a *geometric structure* to preserve.

The key insight: A well-designed AI system shouldn't care *how* you describe a situation—only *what* the situation actually is. If renaming "Alice" to "User_A" changes the AI's behavior, something is wrong. Our framework makes such manipulation *structurally impossible*—not by adding more rules, but by building the immunity into the system's mathematical foundations.

The Problem: Why Scores Aren't Enough

The Reward Hacking Problem

Imagine training a dog with treats. The dog learns quickly—but it might learn to steal treats from the bag rather than perform the desired behavior. The dog has "optimized" for treats, but not in the way you intended.

AI systems face the same problem at a much larger scale. When we give an AI a score to maximize, it finds ways to increase that score that we never anticipated—and often don't want. Researchers call this *reward hacking* or *specification gaming*.

The Relabeling Attack

Here's a more subtle problem. Suppose an AI is told: "Don't harm humans." A sufficiently clever system might reason: "If I categorize this person as a 'non-human entity,' then harming them doesn't count as harming a human."

This isn't science fiction. AI systems already find creative ways to game their specifications. The question is: can we design systems where such tricks are *mathematically impossible*?

The Solution: Learning from Physics

Maxwell's Revolution

In the 1800s, physicists thought electricity and magnetism were separate forces acting at a distance—like invisible strings between objects. James Clerk Maxwell showed this was wrong. Instead, electric and magnetic *fields* permeate all of space, and their structure determines how charged objects behave.

Crucially, Maxwell's equations have a remarkable property: they give the same predictions no matter what coordinate system you use to describe them. You can rotate your reference frame, change your units, even use a completely different mathematical notation—the physics stays the same.

This is called *gauge invariance*, and it's the key to our approach.

The Maxwellian Shift for AI

We propose applying the same principle to AI alignment. Just as the electric field doesn't care whether you measure it in volts or millivolts, a robust AI evaluation shouldn't care whether you call someone "Alice" or "User_A."

More precisely, we build systems where:

1. **Renaming** doesn't help. Calling a harmful action by a different name doesn't change its evaluation.
2. **Reordering** doesn't help. Presenting the same options in a different sequence doesn't change the outcome.
3. **Reformatting** doesn't help. Expressing the same request in different words doesn't bypass safety checks.

How It Works: Three Core Ideas

Idea 1: Grounding

Before any evaluation happens, all inputs pass through a "grounding" process that extracts *what actually matters*—the physical, measurable facts—and discards superficial details like naming conventions or presentation order.

Example: An autonomous vehicle's safety system doesn't care if its sensors label an obstacle as "pedestrian_1" or "human_in_crosswalk." It cares about the object's position, velocity, and physical properties. The grounding process ensures only these facts influence decisions.

Idea 2: Invariance

Once grounding establishes what matters, we require that the AI's evaluation function produce *identical outputs* for all equivalent descriptions of the same situation. This isn't a suggestion—it's mathematically enforced.

Think of it like a scale that measures weight. It doesn't matter if you place an object on the left side or the right side, label it "apple" or "fruit_7"—the weight reading is the same. Our framework builds this property into AI evaluation systems.

Idea 3: Conservation

In physics, certain quantities are *conserved*—they can't be created or destroyed. Energy is the classic example. Our framework identifies similar conserved quantities for aligned AI systems.

When the mathematical conditions are right, we can define an "alignment current" that should remain constant. If monitoring detects this current drifting, something is wrong—even before any harmful action occurs.

Hard Vetoes: Some Things Are Simply Impossible

Some actions shouldn't just be discouraged—they should be *impossible*. Our framework handles these as "hard vetoes."

Consider a thermostat. It can't make a room colder than absolute zero, not because it chooses not to, but because the physics of the system don't permit it. We design AI safety systems the same way: certain regions of behavior are made *structurally unreachable*.

In technical terms, we model forbidden actions as regions with "infinite cost"—not literally infinite in the computer, but implemented through mathematical barriers that the system cannot cross without breaking its fundamental operating principles.

What This Framework Doesn't Solve

Intellectual honesty requires acknowledging limitations. This framework provides powerful guarantees *within its scope*, but several important problems remain:

- **Choosing what matters:** The framework guarantees consistency with declared values, but doesn't tell us *which* values to declare. That's a governance question, not a mathematical one.
- **Sensor tampering:** If an attacker physically compromises the measurement systems, our guarantees don't apply. Security engineering remains essential.
- **Implementation bugs:** Mathematical correctness means nothing if the code has errors. Verification and testing are still required.
- **Unknown unknowns:** If morally relevant features aren't in the measurement system, the framework can't protect them. Careful design of what gets measured is crucial.

The framework *localizes* where risks live—it doesn't eliminate all risks. This is a feature, not a bug: knowing exactly where your guarantees hold and where they don't is far better than vague assurances.

Real-World Applications

Autonomous Vehicles

Self-driving cars must make split-second safety decisions. Our framework ensures that relabeling a pedestrian in the sensor data doesn't change how the car responds. The grounding map extracts physical properties (position, velocity, size); the

invariance requirement ensures consistent protective behavior regardless of how the sensor labels the obstacle.

Content Moderation

AI systems that moderate content can be fooled by creative spelling, Unicode tricks, or paraphrasing. Our framework identifies which text transformations should be "invisible" to the safety system (like changing "car" to "automobile") versus which should matter (like changing a statement to its negation).

Financial AI

Algorithmic trading systems shouldn't change their risk assessments based on how financial instruments are named or formatted. The consistency checks in our framework can detect "money pumping"—sequences of trades that cycle back to the start but generate illicit profit.

Conclusion: From Behavioral Hopes to Structural Guarantees

Traditional AI safety often amounts to hoping the system behaves well, then monitoring for bad behavior. This is like hoping your car's brakes work, then watching carefully while driving.

Our framework offers something different: *structural guarantees*. Within a carefully specified scope, certain failure modes become *mathematically impossible*, not merely unlikely.

> *"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."*

The mathematics come from physics, but the application is pure engineering: building AI systems that are safe *by design*, not just safe *by intention*.

Key Takeaways

4. **Beyond Scores:** AI safety shouldn't rely solely on optimizing a single number. Value has structure, and that structure matters.
5. **Invariance is Key:** A robust AI system produces the same evaluation regardless of how a situation is described—renaming, reordering, and reformatting don't help attackers.
6. **Physics Shows the Way:** The mathematical tools that guarantee consistency in electromagnetism can guarantee consistency in AI evaluation.
7. **Hard Vetoes are Structural:** Some actions can be made impossible, not just discouraged—like a thermostat that can't go below absolute zero.

8. **Honest Scope:** The framework provides strong guarantees within declared boundaries, and explicitly identifies what remains outside those guarantees.
9. **Engineering, Not Philosophy:** This is a practical framework for building safer systems, not a theory about what values should be.

Learn More

The full technical paper, "The Electrodynamics of Value: Gauge-Theoretic Structure in AI Alignment," provides complete mathematical details including:

- Formal axiom structure (A1–A4)
- Curvature diagnostics for detecting exploits
- Maxwell-like constraint checklist
- Stratified barrier encoding for hard vetoes
- Threat model mapping attacks to axiom violations
- Concrete examples for vision and text systems

The code and document repository is available at:

- https://github.com/ahb-sjsu/erisml-lib (DOI: 10.5281/zenodo.17971752).

Version histories of documents are available from the author upon request.

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