The End of Armchair Ethics

How Invariance Testing Makes Normative Systems Falsifiable

Andrew H. Bond

andrew.bond@sjsu.edu

Department of Computer Engineering,

San José State University December 2025

Abstract

For twenty-five centuries, ethical claims have been unfalsifiable. You cannot run an experiment to determine whether utilitarianism is correct. No observation settles whether Kant was right. This paper argues that the question was malformed. The answerable question is not whether an ethical theory is true, but whether an ethical judgment system is consistent, non-gameable, and accountable. These are empirical properties with pass/fail criteria. We present a framework—Philosophy Engineering—that provides the first falsifiability conditions for normative systems: declared invariances, operational tests, minimal witnesses for failures, and machine-checkable audit artifacts. The result is not a new ethical theory. It is the infrastructure that makes ethical engineering subject to the same discipline as any other engineering domain.

1. The Twenty-Five Century Stalemate

Consider the state of ethical discourse. Utilitarians argue with deontologists. Virtue ethicists critique both. Contractarians propose alternatives. The debate continues, century after century, because there is no empirical arbiter. No experiment can be run. No observation is decisive. The arguments are sophisticated, the intuitions are marshaled, but the stalemate is permanent.

This is not a failure of ethics. It is a category confusion. We have been asking a question that may have no answer—"Which ethical theory is true?"—while ignoring questions that demonstrably do.

Compare: "Is Euclidean geometry true?" is not quite the right question. The better questions are: Is Euclidean geometry internally consistent? Does it apply to this domain? Do its predictions match observation in the relevant regime? These are answerable. The first is provable. The second and third are empirical.

Ethics has lacked this separation. We conflated the (possibly unanswerable) question of ethical truth with the (demonstrably answerable) questions of ethical system integrity. This paper concerns only the latter.

2. The Testable Questions

An ethical judgment system—whether implemented in a human institution or an AI—takes representations of situations as input and produces evaluations as output. About such systems, we can ask:

> **Consistency:** Does the system produce the same judgment for the same situation, regardless of how that situation is described?
>
> **Non-gameability:** Can the system be exploited by redescribing situations in semantically equivalent but superficially different ways?
>
> **Accountability:** When judgments differ, can the system attribute the difference to a change in the situation, a change in the evaluative commitments, or uncertainty?
>
> **Non-triviality:** Does the system actually distinguish between genuinely different situations, or does it collapse everything to a constant output?

These are not metaphysical questions. They are engineering questions. They have operational definitions. They produce yes/no answers. They generate witnesses when the answer is no.

3. The Core Principle

The central construct is the Epistemic Invariance Principle (EIP):

> *A judgment procedure is epistemically well-posed only if it is invariant—up to declared output equivalence—under a declared set of meaning-preserving transformations.*

Unpacked: if you declare that certain transformations of input (renaming variables, permuting option order, paraphrasing, changing units) should not change the judgment, then the system must actually be invariant under those transformations. If it is not, you have a witness: a specific transformation that flips the output. This witness is a reproducible, minimal counterexample. It is evidence of a defect.

This is falsifiability. Not for ethical truth—that may remain forever beyond reach—but for ethical system integrity. And system integrity is what we can engineer.

4. The Geometry of Consistency

There is a deeper structure here. The requirement that evaluation be invariant under a group of meaning-preserving transformations is precisely the structure studied in gauge theory—the mathematics of symmetry and invariance that underlies modern physics.

This is not metaphor. The formal apparatus is identical:

> Representations live in a space. Transformations act on that space. Invariance means the evaluation function factors through the quotient—it depends only on the equivalence class, not the representative.

Curvature measures the failure of invariance to be path-independent. Non-zero curvature means there exist sequences of transformations that return to the "same" representation but produce different evaluations. This is an exploit. It is detectable.

Conservation laws (when applicable) provide monitored quantities that should remain stable. Drift signals symmetry-breaking or model mismatch.

The physics vocabulary is not essential. What is essential is the operational content: invariance requirements can be tested, violations can be witnessed, and the geometric structure provides diagnostic tools for finding exploitable inconsistencies.

5. What This Is Not

This framework does not:

Tell you which values to hold. The choice of what matters is a governance problem, not an engineering problem. The framework tests whether your chosen commitments are applied consistently, not whether they are correct.

Prove ethical truths. We remain agnostic about whether there are ethical truths in any metaphysical sense. The framework operates at the level of systems, not reality.

Eliminate all failure modes. Sensor spoofing, implementation bugs, incomplete grounding, adversarial inputs outside the declared envelope—these remain. The framework localizes where risk lives; it does not eliminate risk.

Replace human judgment. It provides discipline for judgment systems, human or artificial. The discipline is: declare your invariances, test them, produce witnesses when they fail, audit everything.

6. The Paradigm Shift

The shift can be summarized in one table:

Traditional Ethics	Philosophy Engineering
Is this action wrong?	Does this judgment flip under meaning-preserving transforms?
Is utilitarianism correct?	Is this utilitarian system internally consistent and non-gameable?
Argue from intuitions	Produce minimal witnesses for failures
Debate endlessly	Run the test suite
Status: unfalsifiable	**Status: pass/fail**

The right column does not answer the questions in the left column. It replaces them with answerable questions. This is not a retreat. It is a clarification of what can be known.

7. The Pragmatist Foundation

This framework rests on a pragmatist epistemology: formal systems—logic, mathematics, modal semantics—are tools, not mirrors of metaphysical structure. They are judged by their utility, coherence, and predictive power, not by their alleged correspondence to a mind-independent reality.

This is not skepticism. It is epistemic discipline. The null hypothesis is: do not posit metaphysical structures until evidence forces you. Evidence would be evidence. Logical arguments from modal systems are not evidence. Intuitions about necessity are not evidence. The arguments may be interesting, but they do not meet the burden.

Under this view, the gauge-theoretic structure of invariance requirements is not a discovery that "ethics has shape" in some metaphysical sense. It is the observation that if you want representation-invariance, gauge-theoretic tools are the appropriate formalism—just as arithmetic is appropriate if you want to count. The tool fits the problem. That is all.

8. Implementation

The framework is implementable. The key components are:

> **Transformation registries:** Explicit, versioned, hashed declarations of which input transformations are meaning-preserving.
>
> **Test suites:** Generators that produce transformed inputs spanning the declared invariances.
>
> **Canonicalizers:** Functions that map inputs to normal forms within equivalence classes, enforcing invariance by construction.
>
> **Witness reducers:** Algorithms that minimize failing transformations to produce compact, reproducible counterexamples.
>
> **Audit artifacts:** Machine-checkable records of what invariances were declared, what tests were run, what witnesses were produced, and what lens (evaluative commitments) was applied.

These are not aspirational. They are specified in sufficient detail to be built. The accompanying technical whitepaper provides JSON schemas, theorem statements, and worked examples.

9. The Forest and the Trees

It is easy to miss what is happening here. The formalism looks like mathematics. The vocabulary borrows from physics. The implementation details are software engineering. Each piece, in isolation, is familiar.

The whole is not familiar. For twenty-five centuries, normative claims have been debated without any discipline of falsification. Ethical systems have been proposed, critiqued, refined, and abandoned based on argumentative force, intuitive

plausibility, and rhetorical skill. There was no test suite. There were no witnesses. There was no audit.

This framework provides those things. Not for ethical truth—that may be beyond us—but for ethical systems. And in an era of autonomous AI systems that make normative judgments at scale, the integrity of those systems is what matters practically.

The question is no longer "Is this AI's ethics correct?" That question may be unanswerable. The question is: "Does this AI's judgment system flip under semantically equivalent redescriptions? Can it be gamed? Can its failures be witnessed and audited?" These questions have answers. We now have the infrastructure to find them.

10. Conclusion

Philosophy has often been accused of making no progress. In ethics, the accusation has force: the debates of antiquity are recognizably continuous with the debates of today. This is not because ethicists lack intelligence or rigor. It is because the questions they ask—questions about ethical truth—may not admit of empirical resolution.

The contribution of this work is to separate the unanswerable from the answerable. We do not solve the problem of ethical truth. We dissolve the conflation that made ethical system integrity seem equally intractable. It is not.

Ethical systems can be tested for consistency, non-gameability, accountability, and non-triviality. These tests produce pass/fail verdicts. Failures produce witnesses. Witnesses enable debugging. Debugging enables improvement. Improvement is progress.

This is what it looks like when philosophy becomes engineering.

References

Bond, A.H. (2025). "The Electrodynamics of Value: Gauge-Theoretic Structure in AI Alignment." Technical report, San José State University.

Bond, A.H. (2025). "Philosophy Engineering: A Technical Whitepaper on the Epistemic Invariance Principle." Technical report, San José State University.

Bond, A.H. (2025). "A Pragmatist Rebuttal to Logical and Metaphysical Arguments for God." Technical report, San José State University.

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