Draft abstract (paper #2)

Autonomous AI systems increasingly issue judgments that mix factual inference, causal attribution, and normative choice. Yet many failures that look like “bias,” “misalignment,” or “hallucination” share a common root: **representation dependence**—the system’s output changes under redescriptions that should preserve meaning (renaming variables, reordering options, schema migrations, paraphrases, unit changes, coordinate transforms). We propose an operational account of **AI epistemology**: the rules by which an AI determines when two inputs describe the *same situation*, what counts as admissible evidence, how uncertainty modulates conclusions, and what justifications must accompany claims.

We introduce the **Epistemic Invariance Principle (EIP)**: a judgment procedure is epistemically well-posed for a domain D only if it is invariant (up to declared equivalence) under all transformations that preserve the domain’s task-relevant structure. To avoid trivial invariance (e.g., constant outputs), we pair EIP with **non-degeneracy on the quotient space** (the system must discriminate when underlying structure changes) and **uncertainty stability** (either robust conclusions over bounded uncertainty sets or explicit abstention/escalation). We show how normative invariance (e.g., BIP in ethics) is a specialization of EIP, and we present an implementation blueprint that treats invariance as infrastructure: canonicalization/quotienting, witness-producing invariance tests in CI, declared “lens” artifacts (policy profiles), and machine-checkable audit records binding decisions to equivalence declarations, evidence provenance, and transformation trials.

We evaluate the framework using transformation-based test suites across mathematical, semantic, causal, and normative tasks, demonstrating that invariance violations predict brittle generalization and that enforcing EIP reduces representation-driven failures while preserving performance on structure-changing counterfactuals. EIP reframes “objectivity” as an auditable constraint rather than a metaphysical claim, providing a practical epistemic contract for trustworthy AI judgment.

Detailed outline

1. Introduction

- Problem: AI judgments change under redescriptions that preserve meaning.
- Why this is epistemology (operational): “same situation,” admissible evidence, uncertainty handling, justification obligations.
- Contributions (bullet list):

1. Formalize EIP for general AI judgment (not just ethics).
2. Add **non-degeneracy** + **uncertainty stability** to make EIP substantive.
3. Provide infrastructure blueprint: canonicalization/quotients, invariance CI, audit artifacts, declared lenses.
4. Empirical transformation-suite evaluation across domains.

2. Motivating failures as representation dependence

- Short vignettes (1–2 paragraphs each, not long):
  - Math: variable renaming / reorder premises flips answer.
  - NLP: paraphrase flips stance or factual claim.
  - Planning: option order changes selection (presentation bug).
  - Causal: equivalent graph encodings yield different interventions.
- Unifying diagnosis: system tracks syntax, not structure.

3. Operational AI epistemology

- Definition: an "epistemic contract" specifying:
  - Equivalence relation on inputs (what counts as same-case)
  - Evidence/provenance requirements
  - Uncertainty model and decision policy under uncertainty
  - Justification format + trace obligations
- Distinguish:
  - **Epistemic layer** (well-posedness, invariance, uncertainty)
  - **Task content layer** (the actual facts/norms learned or encoded)

4. Formal framework: situations, transformations, and judgments

- 4.1 Situations and representations
  - X: representation space (strings, graphs, tables, sensor states)
  - S: semantic/task-relevant state space (often implicit)
  - $e: X \to S$: semantics map (not assumed known perfectly)

- 4.2 Structure-preserving transformations
  - Define a transformation family G acting on X.
  - "Structure-preserving" means: $e(x) = e(g \cdot x)$ for intended g.
  - Practical reality: e is approximated → we use declared equivalences and audits.
- 4.3 Judgment function
  - $J: X \to Y$, where Y can be:
    - discrete decisions, distributions, structured outputs, proofs, plans
  - Decompose into: extraction → inference → aggregation → decision.

5. The Epistemic Invariance Principle

- 5.1 EIP (core)
  - For declared $G: J(x) \sim J(g \cdot x)$ for all $g \in G$.
  - "$\sim$" allows canonicalization / equivalence on outputs.
- 5.2 Canonicalization and quotients (engineering form)
  - Canonicalizer $\kappa: X \to X$ such that $\kappa(x) = \kappa(g \cdot x)$.
  - Enforced invariant judgment: $J_\kappa(x) = J(\kappa(x))$.
  - When canonicalization fails, tests produce witnesses.
- 5.3 Relationship to ethics
  - Show BIP as normative specialization of EIP:
    - domain $D =$ normative choice under bond structure
    - equivalences = bond-preserving transformations
    - lens declaration = governance profile

6. Making EIP non-trivial

- 6.1 Non-degeneracy on the quotient
  - Problem: constant function satisfies invariance.
  - Add "separating power" condition:

        - There exist structure-changing pairs where outputs differ appropriately.
    - Define minimal discriminative adequacy metrics:
        - coverage of quotient classes
        - sensitivity to certified structure changes
- 6.2 Uncertainty stability (epistemic humility)
    - Model uncertainty set $U(x)$or distribution over S.
    - Requirement: either
        - robust decision under all $s \in U(x)$, or
        - abstain/escalate with bounded risk.
    - Connect to safety: refuse to claim objectivity when invariances can't be certified.

7. Infrastructure blueprint: EIP as a systems requirement

- 7.1 Declared equivalence registry
    - A "transform spec" describing allowed G:
        - reorder, relabel, unit scale, schema migrations, paraphrase classes, etc.
    - Versioned and auditable.
- 7.2 Declared lens artifacts
    - Policy/profile artifact (hash + signature)
    - Lens changes are allowed but must be explicit.
- 7.3 Witness-producing invariance CI
    - Test harness generates:
        - transformation trials
        - pass/fail with minimal witness (x, g)
    - Store witnesses and diffs for regression tracking.
- 7.4 Machine-checkable audit artifacts

- JSON schema:
    - tool/version, profile hash, extractor version
    - baseline decision + canonicalized decision
    - per-transform outcome + mapping metadata
    - bond/evidence signatures + provenance pointers
- Verification: third party can validate “this claim respects declared invariances.”

8. Methodology: transformation suites

- Define a standard evaluation protocol:
    - Sample cases x
    - Sample transformations $g$from each class
    - Compute invariance violation rate
    - Compute discriminative adequacy on structure-changing counterfactuals
    - Track abstention/uncertainty outcomes
- Report metrics:
    - Invariance PASS rate by transform type
    - Minimal witness complexity (how “small” the counterexample is)
    - Non-degeneracy score on certified structure changes
    - Accuracy/task utility (to show invariance isn’t achieved by trivialization)

9. Experiments

(Keep this realistic: 2–4 experiments, not a mega-benchmark.)

- 9.1 Math reasoning invariances
    - transforms: variable renaming, premise reordering, notation variants
    - outputs: proof/answer equivalence
- 9.2 Semantic invariance in NLP tasks

    - transforms: templated paraphrase, synonym swap, translation round-trip (declared subset)
    - outputs: entailment/QA/stability checks
- 9.3 Decision/planning invariance
    - transforms: option reorder, ID relabel, unit changes
    - outputs: selected action + ranked list
- 9.4 Normative case study (bridge to SGE/BIP)
    - show EIP → BIP specialization with a short example
    - highlight lens-change vs bond-change vs bond-preserving transforms

10. Results and analysis

- Which transforms are most failure-inducing (usually paraphrase + schema)
- How canonicalization reduces failures
- How non-degeneracy checks prevent "constant-output" gaming
- Uncertainty stability: when abstention is triggered and whether it's calibrated
- Case studies: show 1–2 minimal witnesses and how they guide fixes

11. Related work

- Fairness notions as invariance constraints (individual fairness, counterfactual fairness)
- Invariant/generalization literature (IRM, causal invariance)
- Formal verification / property-based testing
- Epistemology of ML / robustness vs semantics-preserving transforms
  (Keep it focused: 1–2 pages.)

12. Limitations and open problems

- Declaring G is itself a governance task; incomplete $G \Rightarrow$ false confidence.
- Semantic equivalence is hard; paraphrase invariance must be constrained and audited.
- Canonicalization can be expensive or lossy.

- Tradeoffs between invariance and utility; risk of over-normalization.
- Adversarial manipulation of the "equivalence" layer.

13. Conclusion

- Re-state: objectivity claims become **auditable invariance claims**.
- EIP as epistemic infrastructure: "same situation" must be declared, testable, and logged.
- Ethics is a key domain, but the principle generalizes to reasoning and decision-making broadly.

Appendices (high leverage)

- A. Formal definitions and proofs (EIP, non-degeneracy lemmas, uncertainty stability properties)
- B. Audit artifact JSON schema (with a small example)
- C. Transformation registry examples (YAML specs for transforms)
- D. Additional minimal witnesses (more examples without bloating main text)