Internal Epistemic Invariance

Toward Verifiable Reasoning Structure in Neural Systems

Research Proposal (Draft)

Executive Summary

We propose extending the Epistemic Invariance Principle (EIP) from behavioral input-output constraints to internal computational structure. Where EIP asks 'does the judgment change under structure-preserving transformations?', Internal EIP (I-EIP) asks 'does the *reasoning pathway* remain coherent under such transformations?' This requires tracking invariance through the tensor operations that constitute inference—activations, attention patterns, and weight configurations—not merely at the decision boundary.

If successful, I-EIP provides:

- A formal criterion for distinguishing genuine reasoning from brittle pattern-matching *at the mechanistic level*
- Interpretability tools grounded in invariance rather than post-hoc saliency
- Training objectives that enforce structural coherence, not just behavioral mimicry
- Audit infrastructure for certifying that AI systems reason about structure rather than exploit syntax

1. Motivation: The Limits of Behavioral Invariance

EIP (Paper 2) establishes that epistemically well-posed judgments must be invariant under declared structure-preserving transformations. This is testable via transformation suites applied to inputs and outputs.

But behavioral invariance has a gap: a system can satisfy EIP *by accident*—through compensating errors, memorization of equivalence classes, or superficial shortcuts that happen to produce stable outputs without tracking underlying structure.

The problem: Behavioral tests detect *that* invariance holds or fails, but not *why*. Two systems with identical EIP compliance may differ radically in robustness, generalization, and alignment—one reasons about structure, the other has learned a brittle mapping that happens to be stable on the test distribution.

The solution: Extend invariance requirements to internal representations. If a system genuinely tracks structure, its intermediate computations should transform coherently under Γ—not just its final outputs.

2. Core Concept: Internal Γ-Equivariance

2.1 Setup

Let f: X → Y be a neural model with L layers producing intermediate representations:

$$x \to h_1 \to h_2 \to \ldots \to h_l \to y$$

Let Γ be a declared group of structure-preserving transformations on inputs.

2.2 The Internal EIP Requirement

Definition (I-EIP). A model f satisfies Internal Epistemic Invariance with respect to Γ if:

1. **Layerwise equivariance.** For each layer i, there exists a representation $\rho_i: \Gamma \to GL(H_i)$ such that $h_i(g\cdot x) = \rho_i(g)\cdot h_i(x)$ for all $g \in \Gamma$, $x \in X$
2. **Coherence.** The family $\{\rho_i\}$ forms a coherent system: transformations compose consistently across layers.
3. **Grounding.** The output satisfies behavioral EIP: $f(g\cdot x) \approx f(x)$ (or equivariant output transformation where appropriate).
4. **Non-degeneracy.** Representations discriminate when structure changes: if x and x' are not Γ-equivalent, then $h_i(x) \neq h_i(x')$ for relevant layers.

2.3 Interpretation

- Condition 1 says each layer 'knows' how to transform its representations when the input transforms—the transformation commutes through the computation.
- Condition 2 prevents incoherent layer-by-layer transformations that happen to cancel out.
- Condition 3 connects internal structure to behavioral guarantees.
- Condition 4 prevents trivial solutions (constant representations).

3. Theoretical Framework

3.1 Tensor Invariance Signatures

For a weight tensor W in layer i, define the **invariance signature**:

$$\sigma(W, \Gamma) = \{ (g, \varepsilon) : ||W\cdot\rho_{i-1}(g) - \rho_i(g)\cdot W|| < \varepsilon \}$$

This measures how well W commutes with the declared transformations. Perfect equivariance implies σ captures all of Γ with ε = 0.

Proposition (sketch). If all weight tensors have invariance signatures covering Γ with bounded ε, then the model satisfies approximate I-EIP with error accumulating at most linearly in depth.

3.2 Attention as Relational Invariance

For transformer architectures, attention patterns A(x) define which tokens relate to which. Under I-EIP:

$$A(g\cdot x) = \pi(g)\cdot A(x)\cdot\pi(g)^{-1}$$

where π(g) is the permutation/transformation induced on token positions. This says: attention should track the *relational structure*, not the *positional encoding*.

Hypothesis. Transformers that violate this—attending differently to semantically equivalent but positionally distinct tokens—will exhibit brittle generalization detectable via attention invariance probes.

3.3 Gradient Invariance

If a model genuinely represents structure-preserving equivalence, equivalent inputs should produce equivalent learning signals:

$$\nabla\theta\ L(f(x), y) \approx \nabla\theta\ L(f(g \cdot x), y) \quad \text{for } g \in \Gamma$$

Violations indicate the model is learning to distinguish inputs that should be indistinguishable—a signature of representation dependence being *encoded* rather than *resisted*.

4. Implementation: I-EIP Infrastructure

4.1 Invariance Probes

Diagnostic modules inserted at each layer to measure equivariance error, estimate the representation $\rho_i(g)$ that best maps h → h_transformed, and log invariance witnesses for audit.

4.2 Equivariance Regularization

Training objective augmented with invariance penalty:

$$L\_total = L\_task + \lambda \cdot L\_invariance$$

where L_invariance = Σ_i $\Sigma\_g \in \Gamma$ $||h_i(g \cdot x) - \rho_i(g) \cdot h_i(x)||^2$. The representations ρ_i can be fixed (known symmetries), learned (discovered symmetries), or constrained (e.g., orthogonal transformations).

4.3 Coherence Verification

Cross-layer consistency check verifying that $\rho_j(g) \circ f_{i \to j} \approx f_{i \to j} \circ \rho_i(g)$ where $f_{i \to j}$ is the composed transformation from layer i to j. Violations indicate that intermediate layers are 'forgetting' or 'corrupting' the structural invariance.

4.4 Audit Artifacts

Extend EIP audit schema to include internal traces: layer errors, coherence scores, attention invariance, gradient alignment, and failure localization (which layers/heads break invariance). This enables diagnosis of *where* in the network invariance breaks down.

5. Experimental Program

5.1 Probing Existing Models

Experiment 1: Transformer Equivariance Audit

- Models: GPT-2, LLaMA, Claude (if accessible)
- Transformations: token permutation, paraphrase pairs, variable renaming in code
- Metrics: layer-wise equivariance error, attention pattern invariance, representation similarity (CKA)
- Hypothesis: larger models exhibit better internal invariance; instruction-tuned better than base

Experiment 2: Attention Invariance and Generalization

- Task: mathematical reasoning (variable renaming should be invisible)
- Probe: do attention patterns change when variables are renamed?
- Hypothesis: models with higher attention invariance generalize better to novel variable names

5.2 Training with I-EIP Constraints

Experiment 3: Equivariance Regularization

- Task: logical reasoning, causal inference, or normative judgment
- Training: standard vs. I-EIP regularized (λ > 0)
- Hypothesis: I-EIP regularization improves behavioral invariance, generalization, and robustness

Experiment 4: Architectural Invariance

- Compare: standard transformer vs. explicitly equivariant architecture
- Hypothesis: architectural invariance provides stronger guarantees than regularization alone

5.3 Normative Reasoning Case Study

Experiment 5: BIP Internalized

- Task: resource allocation under fairness constraints (connects to SGE)
- Transformations: option reordering, participant ID relabeling, unit rescaling
- Probe: do internal representations of 'fairness' remain stable under bond-preserving transformations?
- Hypothesis: models trained with I-EIP produce more consistent, auditable normative judgments

6. Open Directions: Learning Γ

The deepest extension: can systems *discover* which transformations should be structure-preserving?

- **Symmetry discovery.** Recent work learns equivariances from data. Could be extended to semantic/normative domains.
- **Causal structure.** Transformations that preserve causal structure might be learnable from observational + interventional data.
- **Democratic discovery.** In the DEME framework, stakeholder deliberation surfaces which equivalences matter. Could hybrid systems propose candidate Γ for human ratification?

Conjecture. A system exhibits *reflective* understanding if it can: (1) identify which transformations leave its judgments invariant (self-model), (2) evaluate whether those transformations *should* leave judgments invariant (meta-reasoning), and (3) update its Γ declarations accordingly (learning). This approaches something like epistemic autonomy.

7. Significance and Positioning

7.1 Why This Matters

Problem	I-EIP Contribution
Interpretability	Invariance-grounded probes localize failures to specific layers/heads
Alignment	Internal coherence ensures the system reasons about intent, not syntax
Robustness	Equivariant representations resist adversarial perturbations that preserve structure
Auditability	Full-stack invariance traces from input through computation to output
Trust	Mathematical guarantees, not behavioral spot-checks

7.2 Relationship to Prior Work

- **Geometric deep learning** (Bronstein et al.): provides mathematical language for equivariance; we extend to semantic/normative domains
- **Mechanistic interpretability** (Anthropic, Neel Nanda et al.): probes internal structure; we add invariance as an organizing principle
- **Causal representation learning** (Schölkopf et al.): learns invariant representations for causal reasoning; we generalize to epistemic/normative invariance
- **SGE/BIP** (Paper 1): establishes normative invariance; we extend to internal verification
- **EIP** (Paper 2): establishes behavioral epistemic invariance; we extend to computational structure

7.3 The Research Arc

The four-paper progression:

1. **SGE (Theory):** Mathematical foundations for normative invariance
2. **EIP (Epistemology):** Generalization to epistemic judgment; behavioral tests

3. **I-EIP (Mechanism):** Extension to internal representations; tensor-level verification
4. **Learned Γ (Autonomy?):** Systems that discover and refine their own invariances

8. Timeline and Resources

Phase	Duration	Outputs
Formalization	3 months	I-EIP definitions, theoretical results
Infrastructure	2 months	Invariance probe toolkit
Probing expts	3 months	Empirical characterization of existing models
Training expts	4 months	I-EIP regularization, normative case study
Writing	2 months	Full paper draft
Total	**~14 months**	**Submission-ready paper + toolkit**

Compute requirements: Moderate—probing experiments use existing models; training experiments are medium-scale.

Collaborations sought: Mechanistic interpretability researchers, geometric deep learning experts, alignment researchers.

9. Conclusion

Internal EIP extends the invariance principle from behavioral testing to mechanistic verification. If a system genuinely reasons about structure rather than syntax, this should be visible in its internal computations—not just its outputs.

The proposal is ambitious but tractable: theoretical foundations exist (geometric deep learning, representation theory), experimental methodology is established (probing, regularization, architectural constraints), and the application domain is clear (connecting to SGE/BIP for normative reasoning).

What's new is the synthesis: invariance as a *unifying principle* across behavioral, internal, and learned levels—with full auditability from stakeholder deliberation through tensor operations to final judgment.

We're not just asking 'did the system get the right answer?' We're asking 'did it get there the right way?'

That's what distinguishes reasoning from pattern-matching.

References (Selected)

- Bronstein, M. et al. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges.
- Cohen, T. & Welling, M. (2016). Group Equivariant Convolutional Networks.
- Schölkopf, B. et al. (2021). Toward Causal Representation Learning.
- Zhou, A. et al. (2020). Meta-Learning Symmetries by Reparameterization.
- [SGE Paper] Stratified Geometric Ethics: Mathematical Foundations for Verifiable Moral Reasoning.

- [EIP Paper] The Epistemic Invariance Principle (in preparation).