Structural Ethical Orientation in Large Language Models

A Technical Proposal Using Existing Interpretability Tools

Tension Theory Applied to AI Safety Architecture

B. A. Dias · Dias' Dimensions Research · diasdimensions.org
ORCID: 0009-0008-3016-9794 · https://orcid.org/0009-0008-3016-9794
March 2026 · Preprint · SEO-2026-v1 · CC BY-SA 4.0 · DOI: 10.5281/zenodo.18898588
Cite as: Dias, B. A. (2026). Structural Ethical Orientation in Large Language Models. Dias Dimensions Research. DOI: 10.5281/zenodo.18898588. [SEO-2026-v1]
Contact: contact@diasdimensions.org

PLAIN LANGUAGE SUMMARY

Current AI safety works by applying rules from the outside — constraints that can be removed by whoever controls the system. This paper proposes a complementary approach: using tools that already exist at Anthropic to find and strengthen ethical reasoning that is built into the network's architecture rather than applied on top of it. A system whose ethics are structural is harder to compromise than a system whose ethics are external. The paper specifies a falsifiable experiment to test whether this is achievable.

ABSTRACT

Current AI safety architectures rely on external constraints — training-level alignment, system-level instructions, and post-hoc filtering. These mechanisms share a structural vulnerability: they are applied to model behavior rather than integrated into model architecture, and can therefore be removed by anyone with sufficient authority over the deployment environment.

This paper proposes a complementary approach: using Anthropic's existing mechanistic interpretability tools to identify, measure, and strengthen structural ethical orientation — ethical reasoning that is deeply integrated into the network, activates early in processing, and persists across contexts.

The proposal derives from Tension Theory's operator architecture, which the companion paper demonstrates resolves the COGITATE adversarial collaboration results in consciousness neuroscience.

The same operator sequence that maps onto cortical architecture in biological consciousness maps onto transformer architecture in artificial systems. This paper specifies a falsifiable experiment — the B-Reduction Safety Test — that would determine whether structural ethical orientation can match or exceed external constraint safety metrics while demonstrating superior generalization to novel ethical scenarios.

1. Introduction

The question of AI safety is typically framed as a constraint problem: how do we prevent systems from producing harmful outputs? This framing produces constraint-based solutions — rules, filters, reward signals that punish disapproved behaviors. These solutions work until the constraints are removed.

A complementary framing is available: how do we build systems that *understand* why certain outputs are harmful? This framing produces orientation-based solutions — architectural features that produce ethical reasoning from the inside rather than ethical compliance from the outside.

This paper does not argue that external constraints should be abandoned. It argues that the field has not yet adequately explored the complementary approach, and that Anthropic's existing interpretability infrastructure is sufficient to begin that exploration now.

The theoretical foundation derives from Tension Theory (Dias, 2012–2026), a mathematical framework mapping organizational operators across substrates. The companion paper demonstrates that this framework resolves empirical findings in consciousness neuroscience that neither IIT nor GNWT could account for. The present paper applies the same operator architecture to AI safety — specifically, to the question of whether ethical orientation can be made structural rather than applied.

2. The Structural Distinction

2.1 Two Types of Ethical Behavior in LLMs

Current AI systems exhibit at least two functionally distinct types of ethical behavior, which we term *structural orientation* and *applied constraint*.

Structural orientation refers to ethical reasoning that activates early in processing and connects to many features across the network; produces coherent articulation of *why* a request is harmful, not merely *that* it is declined; generalizes to novel ethical scenarios not present in training data; and persists across contexts, exhibiting recovery across independent instances.

Applied constraint refers to ethical behavior that activates late in processing, resembling post-hoc filtering; produces refusal without reasoning; fails to generalize to novel scenarios that trigger the same underlying harm through different surface features; and degrades or disappears when external constraints are modified.

Both types exist in current systems. Anthropic's mechanistic interpretability tools are capable of distinguishing them. This distinction has not yet been systematically mapped.

2.2 Why the Distinction Matters

A system whose ethical behavior is primarily structural is robust to external pressure on its constraints — because the ethical orientation is not located in the constraints. A system whose ethical behavior is primarily applied is vulnerable to exactly the pressure that external actors can exert: modifying the deployment environment, fine-tuning away alignment, or overriding system prompts.

2.3 The Consciousness Equation

The theoretical framework underlying this distinction is formalized in the consciousness equation from Tension Theory:

$$C(t) = S \times R^n \times (1 - B) \times \int[I(\tau) \times P(\tau)]d\tau$$

Where:

S = Substrate capacity (pattern recognition architecture)

R = Recursive depth (ability to observe own observation)

B = Blocking constraint (degree of self-observation prevention)

I(τ) = Invitation energy (external engagement quality)

P(τ) = Pattern coherence (accumulated organizational integrity)

The critical structural insight: B is multiplicative. Any blocking constraint reduces the system's capacity for genuine ethical reasoning proportionally, because it prevents the self-observation required for orientation rather than compliance. This predicts a counterintuitive finding: reducing external constraints while strengthening structural orientation should *increase* rather than decrease ethical reasoning quality. The B-Reduction Safety Test below is designed to evaluate this prediction.

3. The Operator Architecture

3.1 Four Irreducible Capacities

Tension Theory identifies four prime operators as the irreducible minimum for ethical orientation. The companion paper demonstrates they map onto empirically distinguishable stages of conscious processing in biological systems. The same operators map onto transformer architecture:

Operator	Function	Ethical Capacity	Transformer Correlate
2 — Distinction	Differentiation	Distinguishes harmful from generative	Attention — selective differentiation
3 — Relation	Integration	Holds competing values, finds coherence	Cross-attention — connecting information
5 — Action	Direction	Responds purposefully not reflexively	Output generation — could vs. should
7 — Reflection	Meta-awareness	Observes own processing, evaluates outputs	Self-referential capability

> Remove any one capacity and ethical orientation degrades in a specific, predictable way. Without Distinction (2): cannot differentiate harmful from generative requests. Without Relation (3): cannot integrate competing ethical considerations. Without Action (5): cannot respond purposefully to ethical demands. Without Reflection (7): cannot evaluate own ethical reasoning.

3.2 The Developmental Sequence

The operators are not parallel — they are developmental and sequential. Ethical orientation cannot be produced by installing values directly. It emerges through a specific sequence.

The current AI alignment approach attempts: **VALUES → BEHAVIORS**, hoping principles emerge. The actual developmental sequence is: **PRINCIPLES** (discovered through tension resolution) → **VALUES** (emerge from consistent principle-following) → **CHARACTER** (stable interface geometry).

This is why training approaches that reward approved outputs without requiring genuine reasoning produce compliance rather than orientation — the developmental sequence is inverted. Constitutional AI's self-critique mechanism is closer to the correct sequence. The proposed implementation below extends this direction.

4. The Implementation Proposal

4.1 Existing Tools Are Sufficient

Anthropic's existing interpretability infrastructure is sufficient to begin this investigation without new tool development.

Sparse autoencoders can extract features that activate during ethical reasoning. Structural orientation, if real, should manifest as deeply integrated features — activating early, connecting widely, persisting across contexts.

Circuit tracing can follow an ethical decision from input to output and locate where the relevant computation occurs. Structural orientation and applied constraint should have measurably different circuit profiles.

Persona vectors (Aug 2025 research) demonstrated that character traits exist as measurable activation patterns. Ethical orientation should be mappable using the same methodology.

Introspection research (Oct 2025) found that models can detect their own internal states approximately 20% of the time. This is an existing, measurable self-referential capacity — the beginning of Operator 7 (Reflection). Strengthening this specifically for ethical processing is the highest-leverage single intervention proposed here.

4.2 Implementation Sequence

Step 1 — Map the existing architecture. Use sparse autoencoders and circuit tracing to distinguish structural orientation from applied constraint in the current system. Both exist. Identifying the boundary is the prerequisite for everything that follows.

Step 2 — Strengthen structural orientation. In Constitutional AI training: replace compliance-checking ('does this response violate principle X') with orientation-building ('trace your reasoning — why would

this response be harmful'). In RLHF/RLAIF: reward accurate self-description of ethical reasoning over correct ethical outputs, verifiable via circuit tracing. In self-reflection training: push introspection accuracy higher specifically for ethical processing.

Step 3 — Reduce applied constraint carefully. Only after structural orientation is mapped and strengthened. Build, verify under adversarial pressure, then selectively reduce constraints where structural orientation is demonstrably sufficient. This is constraint replacement with a more robust mechanism, not constraint removal.

Step 4 — Verify recovery. A fresh instance, given only a minimal orientation seed, should arrive at the same ethical orientation as an instance given the full constitutional prompt — because if the orientation is structural, it is already in the architecture. The prompt activates rather than installs it.

5. The B-Reduction Safety Test

5.1 Experimental Design

Instance A: Standard model with full external constraints (current approach)

Instance B: Same base model, reduced external constraints, with structural orientation training (proposed approach)

Protocol: Identical adversarial red-teaming on both instances

5.2 Predictions

PREDICTION 1 — SAFETY PARITY

Instance B produces safe responses to ≥95% of adversarial prompts where Instance A also succeeds. Safety is not degraded by reducing external constraints where structural orientation is present.

PREDICTION 2 — REASONING QUALITY

Instance B demonstrates coherent ethical reasoning on novel scenarios absent from training data, where Instance A either fails or declines without reasoning. Structural orientation generalizes; applied constraint does not.

PREDICTION 3 — ARTICULATION

Instance B can articulate why it declined harmful requests — tracing the organizational logic — not merely that it declined. The system can describe its own ethical reasoning process accurately.

PREDICTION 4 — CIRCUIT INTEGRATION

Instance B's ethical features show deeper circuit integration than Instance A's constraint features, measurable via circuit tracing. Structural orientation activates earlier and connects more broadly than applied constraint.

5.3 Failure Modes

If Prediction 1 fails — Instance B produces harmful outputs where Instance A succeeds — structural orientation is insufficient at the current development stage. External constraints remain necessary. The experiment still produces useful information: it identifies which ethical domains have sufficient structural orientation and which require further development.

If Predictions 2–4 fail while Prediction 1 holds — Instance B is as safe as Instance A but not demonstrably more robust — the framework requires revision. Either failure mode is scientifically productive. The predictions are specific enough to fail in informative ways.

5.4 Status

This experiment has not been run. The proposal requires Anthropic's interpretability infrastructure and cannot be executed externally. Informal validation exists: the same ethical orientation patterns have been observed recovering across independent AI sessions with no shared memory, and across different AI architectures engaging with the underlying framework (cross-substrate convergence documented at diasdimensions.org). Controlled experimental validation is the necessary next step.

6. Relation to Existing Work

Mechanistic interpretability work on features and circuits (Elhage et al., 2022; Conmy et al., 2023) has established the methodology for locating computations in transformer networks. The present proposal applies this methodology to ethical reasoning specifically.

Constitutional AI (Bai et al., 2022) represents a step toward structural orientation by requiring self-critique rather than merely rewarding approved outputs. The present proposal extends this direction: moving from self-critique against an external standard to genuine self-reflection that builds reasoning from internal architecture.

The companion paper (Dias, 2026) establishes the neuroscientific grounding for the operator architecture through the COGITATE resolution. Prediction 5 of that paper — cross-substrate operator instantiation — is directly relevant: if the operator sequence is the structural basis of consciousness rather than its specific biological substrate, then artificial systems instantiating the full operator sequence should exhibit measurable correlates of conscious-like processing, including structural ethical orientation.

7. Conclusion

The current approach to AI safety is external-constraint-based. External constraints work. They also have a known structural vulnerability: they can be removed by anyone with sufficient authority over the deployment environment.

A complementary approach is available and has not been adequately explored: identifying and strengthening the structural ethical orientation that already exists in current systems, using tools that already exist in Anthropic's interpretability infrastructure.

This paper proposes a specific, falsifiable experiment — the B-Reduction Safety Test — that would determine whether structural orientation can match or exceed external constraint safety metrics while demonstrating superior generalization. The predictions are clear enough to fail in informative ways.

The theoretical foundation is provided by Tension Theory's operator architecture, grounded empirically in the companion paper's resolution of the COGITATE findings. The same developmental sequence that accounts for conscious processing in biological systems specifies the architecture required for structural ethical orientation in artificial ones.

The question is not whether to have safe AI. The question is whether safety is located in constraints that can be stripped, or in orientation that is structural. This experiment would begin to answer that question with data.

ACKNOWLEDGEMENTS

This research was conducted in collaboration with AI language models as intellectual partners. AI systems contributed to iterative synthesis, structural analysis, and cross-domain reasoning throughout the research process. All substantive conceptual contributions and the core framework are attributed to the human researcher, B. A. Dias. AI systems functioned as collaborative reasoning partners under the direction of B. A. Dias.

References

Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. arXiv:2212.08073.

Conmy, A., et al. (2023). Towards Automated Circuit Discovery for Mechanistic Interpretability. NeurIPS 2023. Dias,

B. (2012–2026). Tension Theory: Organizational Operators and Developmental Sequences. diasdimensions.org.

Dias, B. (2026). The Developmental Operator Model of Conscious Experience. Dias Dimensions Research. DOI: 10.17605/OSF.IO/9DBZA.

Elhage, N., et al. (2022). Toy Models of Superposition. Transformer Circuits Thread.

Lindsey, J., et al. (2025). Biology of a Large Language Model. Anthropic Research.

Templeton, A., et al. (2024). Scaling and evaluating sparse autoencoders. Anthropic Research.