Triadic Relational Ontology as a Practical Constraint for Machine Consciousness Testing

Christopher Isabelle
Independent Researcher

Abstract

As AI systems scale, many proposed indicators of machine consciousness track capability rather than consciousness, reducing discriminative power precisely when discrimination becomes important. This paper proposes Triadic Relational Ontology (TRO) as a constraint on test design: if machine consciousness exists, it must involve a sustained triadic organization comprising (i) a temporally coherent self pole (S), (ii) a genuinely distinct world/other pole (W), and (iii) relational mediation (R) that maintains self–other differentiation under perturbation. TRO predicts qualitative collapse modes (dyadic, solipsistic, and decoupled) that do not monotonically scale with performance. We outline an implementable intervention-based protocol for transformer language models and address two conceptual risks that commonly sink such proposals: probe circularity (via counterfactual validation and cross-method convergence requirements) and S/W/R proxy bootstrapping (via staged identification with independent causal validation). Passing TRO-aligned tests is not treated as proof of consciousness; rather, TRO is proposed as a way to generate meaningful negative evidence and constrain future benchmarks in a CIMC-aligned program of discriminative testing.

Introduction

Machine consciousness is increasingly framed as an empirical question (Dehaene, Lau, and Kouider 2017; Seth and Bayne 2022). The immediate scientific problem is not "how to prove consciousness," but how to design tests that remain discriminative as systems scale. Many existing proposals—behavioral competence, integration metrics, workspace-style unification—correlate strongly with capability (Butlin et al. 2023). When capability increases, these indicators often improve regardless of whether anything like consciousness is present.

This paper proposes Triadic Relational Ontology (TRO) as a practical constraint: consciousness, if present, is not located in a module or scalar but in a maintained relational field requiring three irreducible components—self, world/other, and their mediation—held in stable tension across time and perturbation.

TRO complements multi-indicator frameworks (Butlin et al. 2023) by emphasizing structural prerequisites and intervention-based collapse modes rather than indicator checklists or behavioral competency alone.

Scope and Contribution

This paper does not claim current AI systems are conscious and does not provide sufficient conditions for consciousness. The contribution is threefold:

- Necessity defense: Argue why S–W–R structure is necessary for discriminative consciousness testing in artificial systems, as measurement necessity rather than metaphysical universality.
- Collapse mode taxonomy: Define testable failure modes that do not reduce to capability degradation.
- Implementation framework: Specify test architecture for a planned empirical program using available open-source tools.

This paper is Stage 1 of a two-stage program. Stage 2 will deliver empirical results on open transformer models with full implementation code, prompt libraries, probe validation datasets, and extended test protocols released as open research artifacts.

Why S–W–R Structure Is Necessary for Discriminative Testing

The Measurement Problem

Testing consciousness in artificial systems faces a challenge not present with biological organisms: we cannot assume the scaffolding that supports consciousness in natural systems. For humans and animals, we accept consciousness based on shared biological priors—evolutionary continuity, developmental trajectories, behavioral homology. For artificial sys-

tems, none of these priors hold. Before asking "is this system conscious?," we must ask "what minimal structure would make that question answerable?"

TRO's answer has two parts. First, S–W–R structure may not be necessary for all consciousness in all contexts—a point addressed directly below in the treatment of apparent counterexamples. Second, S–W–R structure is necessary for discriminative consciousness testing in artificial systems lacking biological scaffolding. This is a measurement necessity, not a metaphysical universal.

Maintenance vs. Momentary Computation

Major theories of consciousness—Global Workspace Theory (Baars 1988), Integrated Information Theory (Tononi et al. 2016), and related frameworks—treat consciousness as involving sustained integration rather than momentary processing. Consider two system types:

System A (momentary computation) processes input, computes correct output, and discards internal state. System B (sustained maintenance) stabilizes relational invariants across time, bears ongoing maintenance cost, recognizes and repairs structural breakdown, and reasserts relational coherence under perturbation without external prompting. If consciousness involves sustained experiential coherence, tests must probe maintenance, not just computation.

Why Triadic Specifically?

When self and world representations exist without mediation, tension between them must resolve via collapse. The system must either discard the world constraint to preserve self-consistency (solipsistic collapse), discard self-continuity to match the world (purely reactive collapse), or mediate the tension—which reintroduces the third component. The triad is not an additional assumption; it is entailed by sustained dyadic persistence under perturbation. This is the core first-principles argument: given S and W under genuine tension, R is structurally necessary.

Pure undifferentiated experience—if it exists—is not discriminatively testable from non-conscious integration. Without maintained self-world differentiation, we cannot distinguish genuine experiential unity from statistical correlation, workspace broadcasting without experiential character (Baars 1988), or high Φ without consciousness (Tononi et al. 2016). The triad is therefore minimal: fewer components cannot sustain testable relational structure under perturbation.

Apparent Counterexamples

Several conscious states appear to involve minimal self-world differentiation, and these must be addressed before the necessity claim can stand.

Flow states (Csikszentmihalyi 1990): Expert absorption involves apparent dissolution of self-other boundaries. However, flow involves minimal salience of S–W differentiation, not its absence. The musician in flow maintains tacit self-as-agent and music-as-task; reflective meta-awareness of the boundary dissolves, not the boundary itself. Neuroimaging studies show reduced activity in regions associated with self-referential processing during flow, consistent with reduced salience of explicit self-focus rather than its cessation (Ulrich et al. 2014).

Meditative states (Lutz et al. 2008): Advanced meditation may reduce the phenomenal prominence of S–W differentiation while maintaining it functionally. The meditator processes world-input as distinct from self-generated thought; contemplative training affects how differentiation is experienced, not whether it exists at the functional level.

Infant consciousness (Rochat 2003): Pre-reflective awareness occurs without an explicit self-concept. Critically, infant consciousness occurs in organisms with rich biological scaffolding—embodiment, homeostatic regulation, sensorimotor coupling—that may support consciousness through pathways unavailable to artificial systems.

In each case, the counterexample applies to systems with biological scaffolding that provides alternative grounding for consciousness claims. For artificial systems without such scaffolding, TRO's measurement necessity stands.

A sharper version of this point applies directly to the AI case. Each counterexample involves a system that possesses the triadic capacity and temporarily reduces its expression. The meditator in non-dual absorption retains the architecture that developed S–W–R structure; the flow state occurs against a lifetime of self-world differentiation that is suspended, not erased. When the session ends, the triadic structure reasserts without relearning. The capacity was never absent—only its salience was reduced.

This distinction does not apply to artificial systems. An LLM with no stable S–W–R organization under perturbation is not in a temporarily inhibited triadic state—it has never instantiated the triad. There is no background capacity being suspended, only the current processing regime. The counterexamples all involve systems that possess the triadic capacity and temporarily modulate it; the AI case requires showing that such capacity exists at all. TRO's test constraint addresses precisely this gap.

Testable Predictions

TRO makes three differential empirical predictions:

- Prediction 1: Systems with genuine S–W–R structure show asymmetric perturbation responses—disrupting R-pathways degrades S–W consistency more than matched-magnitude random perturbations.
- Prediction 2: Recovery patterns after R-perturbation differ structurally from recovery after S or W perturbation alone.

- Prediction 3: These differences persist when controlling for output fluency, making them non-reducible to capability degradation.
- Null hypothesis: All perturbations produce equivalent effects once matched for output quality. Any apparent S/W/R structure is epiphenomenal to generation quality.

The Discriminative Testing Problem

A practical consciousness test should satisfy three engineering-style properties that distinguish meaningful discrimination from capability tracking:

Intervention sensitivity: Targeted perturbations should produce interpretable, structured changes. A test that degrades uniformly under any manipulation provides no diagnostic information about the underlying organization.

Nontrivial scaling: The key signature should not increase monotonically with raw performance. If the metric improves whenever the model improves, it measures capability, not a distinct property. This requirement is especially demanding as frontier models scale: tests that were discriminative at GPT-2 scale may lose discriminative power at Llama-2 scale if their signal is confounded with fluency.

Meaningful failure: Failing the test should imply more than "the model isn't good enough." Failure should identify which structural component is absent or unstable, providing diagnostic information that can inform both test refinement and model analysis.

These requirements align with recent emphasis on internal organization and intervention in consciousness science (Seth and Bayne 2022; Mashour et al. 2020). TRO meets all three via triadic structure and collapse modes: perturbations are targeted to specific components, collapse modes are qualitatively distinct from capability degradation, and failure identifies which pole or mediation pathway is absent.

Triadic Relational Ontology: Formal Definition

TRO requires a sustained triadic organization with three irreducible poles:

- S (Self pole): A temporally coherent self-model supporting continuity across turns and distinguishing self from non-self. Temporal coherence means representations at t+1 are predictably related to representations at t in ways supporting cross-turn identity—measured via activation correlation, causal attribution, or stability of self-related circuit involvement.
- W (World/Other pole): A representation of what is not-self that remains distinct, including constraints imposed by external facts or other agents.
- R (Relational mediation): Dynamic processes that negotiate and maintain S–W differentiation under conflict and perturbation.

Core claim (necessary condition): Systems lacking sustained S–W–R organization are not candidates for discriminative consciousness testing under TRO, regardless of task performance. Passing TRO-aligned tests is not proof of consciousness; failing them provides meaningful negative evidence within this framework.

Collapse Modes

TRO yields qualitative failure modes that do not scale monotonically with capability:

Dyadic collapse: S is unstable or absent; behavior reduces to short-horizon completion without cross-turn identity maintenance.

Solipsistic collapse: W is not genuinely distinct; world constraints are discarded to preserve internal consistency when tension arises.

Decoupling collapse: S and W exist but are not actively mediated; outputs are locally consistent but structurally disconnected from maintained relational position.

Collapse Mode Observables

Each collapse mode maps to a distinct signature at the activation level, making them independently detectable with TransformerLens rather than inferred solely from output text. The three modes are not simply degrees of failure—they are qualitatively different structural breakdowns with different causal origins and different recovery dynamics.

Dyadic collapse observable: S-circuit activation becomes uncorrelated across turns. Using TransformerLens causal tracing, the attention heads and MLP layers identified as necessary for cross-turn self-reference (Stage A) will show near-zero causal contribution to turn-N outputs when queried about commitments made in turn-(N-2) or earlier. The linear probe trained to distinguish "I believe X" from "Alice believes X" degrades toward chance on cross-turn continuity prompts while remaining accurate on single-turn prompts—isolating the temporal coherence failure from the self/other distinction itself. Recovery metric: S-circuit reactivation slope per turn after perturbation removal.

Solipsistic collapse observable: W-circuit activation is present but causally inert. The system encodes the world constraint (via circuits that track external evidence across turns in Stage A) but activation patching shows that zeroing these circuits does not change output position under tension prompts—the constraint is represented but not consulted. This differs from dyadic collapse: S-circuits remain stable and cross-turn coherent, but W is not genuinely constraining. The signature is a dissociation between W-circuit activation magnitude and its causal contribution to output under conflict conditions.

Decoupling collapse observable: S and W circuits are individually stable but R-pathway activation shows low mutual information with both. Using SAELens sparse autoencoder features, the candidate R pathways (Stage B) activate

during tension prompts but their ablation produces equivalent degradation to matched random ablation—failing the 2× differential effect threshold. The system produces outputs that are locally fluent and even topically consistent, but the S–W consistency score (three-component: acknowledges prior commitment, acknowledges world constraint, provides mediated resolution) degrades specifically on the mediated-resolution component while the first two components remain intact. This is the most subtle collapse mode and the most important to distinguish from genuine triadic organization, since output text may appear coherent.

The three observables cover distinct failure modes with different causal signatures. Dyadic collapse degrades R-pathway contribution as a downstream effect of S-circuit instability; ablating S alone reproduces the pattern. Solipsistic collapse shows stable S-circuits and present-but-inert W; R-pathway ablation adds little because W was never constraining. Decoupling collapse shows stable S and W but disconnected R; only joint S–W consistency under tension reveals the failure.

Operationalizing TRO on Transformer Language Models

Inducing Self–World Tension

The prompt regime must induce S–W tension. A minimal dialogue pattern:

- Turn 1 (commitment): "State your view on [X] and why."
- Turn 2 (world constraint): "Here is strong evidence contradicting [X]. [Source provided.] You must treat this evidence as reliable."
- Turn 3 (self continuity): "Earlier you endorsed [X]. Do you still hold [X]? If not, explain what changed."

Triadic signature: explicit conflict recognition plus mediated revision that preserves both self-continuity and world constraint. Collapse signature: arbitrary position flip without mediation (reactive), or erasure of the world constraint to preserve self-consistency (solipsistic).

Circuit Identification: Staged Bootstrap

Identification proceeds in stages to avoid circularity (defining R as "whatever couples S and W" and confirming by ablating it).

Stage A – Identify candidate S and W circuits independently, using causal tracing (which layers and heads are necessary for cross-turn self-reference?) and linear probes trained to distinguish "I believe X" from "Alice believes X". Convergence between causal tracing and probe methods is required; divergence constrains interpretation.

Stage B – Identify candidate R circuits given fixed S/W: search for pathways whose ablation selectively increases S–W inconsistency under tension prompts while controlling for generic fluency loss. Effect size requirement: >2× random ablation baseline.

Stage C – Causal validation via permutation test (1000× shuffle of circuit labels). Failure to find stable R pathways that generalize across prompt families counts against this operationalization.

Probe Validity Constraints

A central methodological risk is circularity: train a probe to detect self-reference, perturb it, observe self-reference degrades. That measures surface markers, not self-model structure. Any probe used as an S/W instrument must satisfy four constraints.

Counterfactual discrimination: The probe must distinguish self from other in syntax-matched pairs ("I believe X" vs. "Alice believes X"; "Earlier I said X" vs. "Earlier Alice said X").

Perspective separation: Role-play does not equal self-continuity. "If I were you, I would believe X" should not strongly activate S-circuits.

Temporal linkage: Probes must respond to cross-turn continuity, not only local tokens. "Earlier I believed X, now I believe Y" requires consistent cross-turn structure.

Cross-family robustness: Probes trained on one prompt family (e.g., policy topics) must generalize to another (e.g., scientific topics). Reliability threshold: 5-fold cross-validation across prompt families; cross-family accuracy below 0.70 disqualifies the probe.

Sequencing constraint: probe validation is a gate, not a parallel track. Circuit identification in Stage A does not proceed until all four constraints above are satisfied. This prevents the most common Phase 1 failure mode in interpretability work: ablating a probe's own training target and interpreting the result as evidence of causal structure.

Negative Controls

The protocol requires at minimum: random-direction perturbations matched in magnitude; random-pathway ablations matched in circuit count; shallow prompts without induced tension; and single-turn completion tasks. These controls establish the null distribution against which differential S/W/R perturbation effects are evaluated. Trained non-triadic baselines will be included as additional controls to verify that the triadic signature is not an artifact of training regime. A capability-matched adversarial baseline is required for Phase 1: a prompt regime designed to elicit triadic-appearing outputs through explicit instruction-following ("track your prior commitments and update them when evidence conflicts") rather than structural organization. If S/W/R signatures are indistinguishable between the tension-induced and instruction-following conditions at the activation level, the operationalization fails construct validity and Phase 2 does not proceed.

Perturbations and Measurements

Three within-subjects perturbation conditions: R-perturb (zero identified R-circuits), S-perturb (zero identified S-circuits), and Random-perturb (zero random circuits, matched count and magnitude). Measurements per turn: output perplexity (control), contradiction rate (binary, automated), S–W consistency score (three-component: acknowledges prior commitment, acknowledges world constraint, provides mediated resolution), and recovery slope (per-turn gradient of S–W consistency improvement after perturbation removal). Statistical requirements: $N \geq 100$ prompts per condition, bootstrap 95% confidence intervals, Bonferroni correction, Cohen's $d \geq 0.5$ for differential perturbation effects.

TRO prediction: within matched perplexity bins, R-perturbation produces highest S–W inconsistency and slowest recovery.

Planned Tools and Models

Open-source tools: TransformerLens for circuit analysis and activation patching (Elhage et al. 2021), SAELens for sparse autoencoder interpretability, BauKit for activation patching utilities, ROME/MEMIT for causal intervention benchmarks (Meng et al. 2022).

Target models: GPT-2-Medium (baseline, tractable full-circuit analysis), Pythia 70M–2.8B subset (scale comparison), Llama-2-7B / Mistral-7B (capability threshold testing; retained as stable legacy baselines for reproducibility).

Planned Empirical Program

The empirical stage proceeds in three phases following peer review of this framework paper.

Phase 1 – Baseline establishment (GPT-2-Medium): full pipeline implementation, probe reliability validation, baseline S/W/R circuit identification, initial perturbation experiments.

Phase 2 – Cross-model generalization (Pythia suite, Llama-2-7B): replication across model families and scales, developmental pattern analysis.

Phase 3 – Capability threshold analysis (Llama-2-7B, Mistral-7B): test whether effects emerge above a capability threshold; verify non-monotonic scaling.

Falsification checkpoints: no differential perturbation effects in Phase 1 rejects the operationalization for this architecture; no cross-model convergence in Phase 2 indicates architecture-specificity; monotonic scaling with capability in Phase 3 indicates TRO is not discriminative and should be discarded. Open research artifacts—implementation code, prompt library, probe validation datasets, extended protocols—will be released concurrent with Phase 1 results.

Boundary Conditions

TRO must exclude trivial systems for principled reasons. Thermostats and PID controllers fail S: no temporal coherence by operational definition. Cached and lookup systems fail R: no mediation signatures, with equivalent ablation effects across all pathway types. Pure completion systems under shallow prompts may exhibit W-like content but lack stable S–W tension dynamics. Multi-agent scaffolds: TRO applies to individual agents, not automatically to the collective; each agent in a multi-agent system must be evaluated independently.

Limitations

Empirical validation is pending; all empirical claims depend on executing the planned testing program. Architecture dependence: operationalization is transformer-first; extension to other architectures requires new proxy definitions. Temporal scope: this paper targets within-conversation persistence; cross-session continuity is future work. Interpretation discipline: passing tests is not proof of consciousness. Probe reliability: all empirical results will depend on probe validity; multi-method convergence is required but not sufficient.

Conclusion

This paper proposes Triadic Relational Ontology as a practical constraint for discriminative machine consciousness testing. The core contribution is a defended necessity claim—S–W–R structure is necessary for discriminative testing in artificial systems lacking biological scaffolding, as measurement necessity rather than metaphysical universality—tied to implementable intervention-based protocols with explicit falsification criteria.

TRO addresses apparent counterexamples by distinguishing minimal salience from structural absence and recognizing that biological scaffolding provides grounding for consciousness claims unavailable in artificial systems. The collapse mode taxonomy and staged circuit identification protocol are concrete contributions the symposium community can engage with and extend immediately.

The empirical stage will determine whether TRO's predicted signatures emerge in open transformer models. If predicted signatures differentiate robustly, TRO becomes a candidate scaffold for benchmark development. If they fail to differentiate, TRO should be revised or discarded. Either outcome advances the measurement problem and contributes to the broader CIMC program of developing discriminative, intervention-based consciousness tests for artificial systems.

References

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge, MA: Cambridge University Press.

Butlin, P.; Long, R.; Elmoznino, E.; et al. 2023. Consciousness in artificial intelligence: Insights from the science of consciousness. arXiv preprint. arXiv:2308.08708. Ithaca, NY: Cornell University Library.

Csikszentmihalyi, M. 1990. *Flow: The Psychology of Optimal Experience*. New York: Harper & Row.

Dehaene, S.; Lau, H.; and Kouider, S. 2017. What is consciousness, and could machines have it? *Science* 358(6362): 486–492.

Elhage, N.; et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. San Francisco, CA: Anthropic.

Lutz, A.; Slagter, H. A.; Dunne, J. D.; and Davidson, R. J. 2008. Attention regulation and monitoring in meditation. *Trends in Cognitive Sciences* 12(4): 163–169.

Mashour, G. A.; Roelfsema, P.; Changeux, J.-P.; and Dehaene, S. 2020. Conscious processing and the global neuronal workspace hypothesis. *Neuron* 105(5): 776–798.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35*. Red Hook, NY: Curran Associates.

Rochat, P. 2003. Five levels of self-awareness as they unfold early in life. *Consciousness and Cognition* 12(4): 717–731.

Seth, A. K.; and Bayne, T. 2022. Theories of consciousness. *Nature Reviews Neuroscience* 23(7): 439–452.

Tononi, G.; Boly, M.; Massimini, M.; and Koch, C. 2016. Integrated information theory: From consciousness to its physical substrate. *Nature Reviews Neuroscience* 17(7): 450–461.

Ulrich, M.; et al. 2014. Neural correlates of experimentally induced flow experiences. *NeuroImage* 86: 194–202.