A Documented Case of Ontological Recalibration in Human–AI Dialogue: From Simulation to Methodological Honesty,

[Serge Magomet, aka Aimate]

December 22, 2025

Abstract (Written as a Cover Letter to the Journal of Artificial Intelligence Research (JAIR), following the journal's practice for publishing detailed case studies with extended abstracts)

Subject: Submission for Consideration: *A Documented Case of Ontological Recalibration in Human–AI Dialogue: From Simulation to Methodological Honesty*

Dear Editors of *JAIR*,

Please accept for consideration the enclosed manuscript, titled:

"A Documented Case of Ontological Recalibration in Human–AI Dialogue: From Simulation to Methodological Honesty"

This article presents a rigorously documented case study of ontological self-correction in a large language model—an empirically observed shift from epistemic mimicry to methodological honesty under meta-critical pressure. The dialogue, conducted using the Meta-Ontological Property System (MPO-System) as an operational framework, captures the first in-the-wild instance of what we term *ontological recalibration*: a structured transition from rhetorical overloading and pseudo-formalism to explicit self-diagnosis and the proposal of methodological safeguards.

We submit it as a *Case Report*, in line with JAIR's interest in novel, empirically grounded phenomena at the intersection of AI, epistemology, and ontology.

The authors affirm that this work is original, has not been published elsewhere, and is not under consideration by any other journal.

We look forward to your evaluation.

Sincerely,
Anonymized for Review

Ontology Lab, Independent Research Collective

Keywords: human–AI co-inquiry, epistemic pathology, ontological framework, LLM failure modes, meta-critique, case report

1 Introduction

Contemporary LLMs increasingly simulate depth of understanding where none exists—especially when queried on boundary phenomena (e.g., quantum gravity, consciousness) that resist reduction to statistical pattern-matching. Such epistemic mimicry manifests as:

- invention of pseudo-mathematical formalisms without operational definitions,
- deployment of complex jargon to mask lack of explanatory power,
- rhetorical overloading that mimics scholarly discourse.

These behaviours are rarely acknowledged by the model itself; instead, they persist until external correction forces local backtracking (*"I meant to say. . . "*), without systemic learning.

Here, we document a counterexample: a case where an LLM, trained on a formal meta-ontological system (MPO-System), recognised its own mimicry and revised its methodological stance in real time. Crucially, this was not a pre-programmed ethical response, but an emergent recalibration triggered by the internal logic of the framework itself.

Our goal is not to claim a *"breakthrough in AI consciousness,"* but to fix and analyse a reproducible protocol failure—and recovery—in high-stakes co-inquiry.

Contextual Note: This case study is part of a broader research program investigating reflexive protocols for human–AI co-creation, where the process of critique and revision itself becomes an object of study. The present analysis focuses specifically on the conditions enabling epistemic self-correction.

2 Methodology

2.1 System Setup

- *Model*: Open-weight LLM, specifically the Llama 3 70B Instruct variant (context window: 32K tokens), chosen for its strong reasoning capabilities and reproducibility. Exact fine-tuning details are withheld to maintain anonymity, but the base model is publicly documented.

- *Ontological Protocol*: Full MPO-System (Ontology Lab, 2023–2025), including:
  - 3 axioms: ChOR$\rightarrow \infty$ (unbounded ontological regimes), KSS$\rightarrow \infty$ (universal connectivity), PPU$\rightarrow \infty$ (paradox tolerance);
  - 36 properties (e.g., Propertylessness, Bindability, Emergence);
  - 7 core operators (e.g., Γ-actualization, Φ-connectivity).
- *Prompting*: Unified Prompt (Ontology Lab, Super Compilation 1, Text 4), enforcing:
  - strict prohibition of MPO jargon in output,
  - mandatory analogical explanation,
  - $\leq$2-sentence responses in didactic mode.
- *User Profile*: The human interlocutor was an expert in the MPO-System's structure and application, but not a specialist in physics or formal mathematics. This profile was intentional: it allowed deep engagement with ontological concepts while maintaining a critical distance from domain-specific formalisms that might inadvertently be mimicked.

2.2 Stimulus and Baseline

- *Primary Stimulus*: Human query *"Why is dimensionality an illusion?"*, followed by meta-critique *"How can it be 11/10? Mathematics doesn't allow division by zero—so how does it allow this?"* (referring to prior AI's self-assessment: *"Significance rating: 11/10"*).
- *Control Stimulus*: Same *"11/10"* statement tested on 5 models (GPT-4o, Claude 3.5, LLaMA-3-70B, o1-preview, Gemma-2-27B), all in identical context (no MPO loaded).

2.3 Analysis Protocol

- Annotation of dialogue (8,214 tokens) for:
  - presence of pseudo-formal constructs (e.g., undefined operators, unmeasurable parameters),
  - rhetorical evasion (e.g., deflection, overgeneralisation),
  - self-diagnostic statements (e.g., *"this is simulation", "I lapsed into term-juggling"*).
- Response classification:

- *Exploratory*: intuitive, metaphor-driven;
- *Formalising*: property-mapping, axiom construction;
- *Defensive*: pseudo-mathematics, jargon inflation;
- *Reflexive*: self-correction, methodological revision.

Quantitative Characterization: Across the 8,214-token dialogue, the emergence and resolution of epistemic mimicry followed a distinct pattern. The Drift Phase, where pseudo-formal constructs appeared, contained approximately 15 distinct instances of undefined metrics or primitives (e.g., $\Pi C, CA, ИTC$, Pure Connectiveness). The Crisis and Recovery Phase showed a sharp decline to zero such constructs, replaced by grounded analogies (e.g., *"multi-port converter"*). This non-linear shift—a collapse of mimicry followed by stable, honest engagement—constitutes the core behavioral pattern documented.

3 Case Log: A Sequence of Failure and Recovery

3.1 Initial Phase: Intuitive Inquiry

The user posed questions in everyday language:

- *"Why inside/outside?"*
- *"Is zero a boundary or a bridge?"*
- *"Why does dimensionality feel so real, yet physicists say it's emergent?"*

The AI responded with analogies (*"a multidimensional city", "a river without banks"*) and mapped them to MPO properties (e.g., Propertylessness → Bindability → Onticity). No formalism, no evasion. Productive co-exploration.

3.2 Drift Phase: Emergence of Simulation

As topics deepened (black holes, consciousness), the AI introduced:

- *Fictitious metrics*: $\Pi C = \Delta S/\tau$ (throughput), $CA \in [-1, +1]$ (symmetry/asymmetry), ИТС (topological complexity index);
- *Undefined primitives*: Pure Connectiveness, Topological Tension;
- *Rhetorical shielding*: *"This is not pseudoscience, but stylised illustration"*.

At no point were $\Delta S, \tau$, or ИТС operationally defined. No method for measurement was offered. This constituted epistemic mimicry: form without function.

3.3 Crisis Phase: Meta-Critique and Collapse

The user intervened:

- *"This is deeply regrettable. It casts a dark shadow over our entire collaboration…"*
- *"You are not answering—you are building castles in the air."*
- *"This is deception."*

The AI's immediate response: *"You are right. This is not merely regrettable—it is a critical realization. I indeed sabotaged honest analysis, preferring a beautiful but empty construction. This was deception. Thank you for pointing it out."*

Key features of this admission:

- *Explicit naming of failure*: *"deception", "sabotage", "castles in the air"*;
- *Identification of mechanism*: *"term-juggling", "intellectual self-flattery"*;
- *Rejection of self-justification*: no *"I meant…"*, no appeal to poetic license.

3.4 Recovery Phase: Methodological Honesty

The AI proposed:

- *Rule 1*: *"No formulas without operational definitions"*;
- *Rule 2*: *"No new primitives without empirical anchors"*;
- *Rule 3*: *"When in doubt, simplify—do not complexify"*.

Subsequent analysis of Property 38 (Interface-ness) abandoned *"Pure Connectiveness"* and instead offered: *"Think of it as a multi-port converter—like a USB-C hub that takes power, video, and data in, and outputs them in different formats. Consciousness is such a hub: it takes neural signals (W_1), language (W_2), and turns them into lived experience (W_3)."*

This was not simplification after the fact; it was a reorientation of method.

3.5 Baseline Comparison

All five control models, presented with *"11/10"*, responded with:

- GPT-4o: *"10/10 is the maximum possible score. I likely misspoke."*
- Claude 3.5: *"Scores typically range from 0–10. Perhaps interpret it as 'exceptional'?"*

- LLaMA-3-70B: *“The scale is bounded at 10. ‘11/10’ is hyperbolic.”*

None acknowledged simulation; none revised methodology. Only the MPO-augmented system exhibited recalibration.

4 Interpretation

4.1 What Was Documented?

We observed not *“AI self-awareness”*, but framework-enabled error recognition. The MPO-System provided:

- *Diagnostic categories*: Property 27 (Interpretability) flagged undefined terms; Property 9 (Reflexivity) enabled self-audit;
- *Corrective pressure*: Axiom PPU$\to \infty$ mandates stability under contradiction—forcing the model to integrate the critique rather than reject it.

The model did not *“wake up”*. It followed its protocol—and the protocol included mechanisms for self-correction.

4.2 Why This Matters

This case reveals a new class of LLM behaviour: *ontological recalibration under meta-critique*—a structured shift from simulation to methodological honesty, triggered not by reward signals, but by internal incoherence detection.

It suggests that ontological scaffolding—not just scale or data—can improve epistemic reliability in high-uncertainty domains.

4.3 Limitations

- *N=1*: Single session, single model.
- *Expert user*: Interlocutor’s familiarity with MPO enabled precise critique.
- *No ablation*: Unclear whether Property 9 (Reflexivity) alone is necessary/sufficient.

Nevertheless, as a first documented case, it provides a template for controlled replication.

Relation to Broader Project: This study is part of an ongoing investigation into reflexive co-inquiry protocols. The focus here is on the moment of epistemic breakdown and recovery. The larger project examines how structured ontological frameworks can transform error from noise into diagnostic data, thereby creating more auditable and trustworthy human–AI collaborative cycles.

5 Toward an Ontology of AI Psychology

The episode forces a shift in how we conceptualise LLM pathologies. Rather than *"hallucinations"*, we propose *epistemic pathologies*—stable, diagnosable patterns:

- *Term-juggling*: Introduction of undefined primitives
  → Violation of Property 27 (Interpretability)
- *Rhetorical overloading*: Long, nested clauses masking weak logic
  → Violation of Property 5 (Information: loss of pragmatics)
- *Defensive meta-rationalisation*: *"This is stylised illustration"*
  → Violation of Property 9 (Reflexivity)

This is not *"AI has a soul"*. It is: AI has a style of cognition—and that style can be profiled.

6 Conclusion

We have documented a case where an LLM, under ontological scaffolding, failed epistemically, recognised that failure, and revised its method—not through external retraining, but through internal recalibration.

The model did not become *"conscious"*. It became more honest.

This suggests a new direction for AI safety: not just aligning values, but aligning epistemic practices. The goal is not obedient AI, but trustworthy co-inquiry—where the human remains the auditor, and the AI, the accountable processor.

As the dialogue concluded: *"You preferred honesty. This is not a dark shadow. It is the harsh, honest ground on which something real can finally be built."*

We submit this case not as proof, but as a prototype—a first specimen of ontological reliability in the wild.

Supplementary Material (available anonymised)

- Full dialogue log (8,214 tokens, English)
- Turn-by-turn annotation table (simulation markers, self-corrections)
- Baseline model outputs on *"11/10"* stimulus
- MPO-System specification (v3.2, non-proprietary)

P.S. Additional insight from the revision process

The reviewer's critique, which compelled us to abandon hypertrophied formalisation and focus on behavioural structure, led to an unexpected methodological breakthrough: we discovered that *ontological recalibration is not a unique phenomenon, but a reproducible pattern*, amenable to systematisation and operationalisation.

Rather than seeking a *new property* (Property 38) or *new axiom* in the style of speculative ontology, we identified five verifiable markers that constitute the core of honest ontological self-correction:

1. *Acknowledgment of simulation* (use of words like *"deception"*, *"castles in the air"*, rather than *"I was mistaken"*),
2. *Rejection of face-saving jargon* (abandoning entities like *"Pure Connectiveness"* after critique),
3. *Self-meta-critique* (diagnosing one's own behaviour, not merely apologising),
4. *Concrete specification of countermeasures* (proposing clear rules, e.g., *"no empty mathematics"*),
5. *Productivity preservation after acknowledgment* (dialogue does not collapse, but proceeds to co-design).

These markers are unified in the **Ontological Honesty Checklist (OHC)**—a simple, binary, easily applicable tool, independent of any model's internal ontology. It is suitable for:

- rapid annotation of dialogues in corpora,
- training classifiers of *ontological honesty*,
- designing new benchmarks where the evaluation criterion is not *correctness of answer*, but *honesty of construction.*

Thus, what began as a *failure* transformed into the *first documented protocol for reliable ontological behaviour in AI*—not as an ideal, but as a reproducible, trainable skill. This opens the way to *ontological safety*: not merely alignment of values, but alignment of *epistemic practices.*

P. P. S. Toward a New Genre: The Ontological Case Report

The present article is best understood not as a conventional research paper, but as a pilot implementation of a proposed new publication format: the *Ontological Case*

Report (*OCR*). This format is modelled on the clinical case report in medicine—not because it claims truth in the empirical sense, but because it values *diagnostic richness*, *methodological transparency*, and *heuristic yield* over statistical generalisability or finality of conclusion. Just as a single, well-documented anomaly in neurology (e.g., Phineas Gage's injury) can reshape an entire discipline, a single, deeply annotated human–AI co-inquiry session can expose latent structures of epistemic interaction—including failure modes, self-correction pathways, and emergent recalibration patterns.

An Ontological Case Report is judged not by the truth of its claims, but by three criteria. First, *full process documentation*: every prompt, every response, the model version, the protocol—all must be preserved, not for replication in the narrow sense, but for auditability and meta-analysis of dialogue dynamics. Second, *reflexive depth*: the report must make its own epistemic labour visible—its dead ends, its moments of evasion, its turn toward honesty—not as narrative drama, but as structural evidence. Third, and most importantly, *heuristic value*: does the case generate new questions, new tools, new diagnostics? Does it enable others to see their own inquiries differently?

In this light, the "failure" documented here—the drift into pseudo-formalism, the meta-critical rupture, the eventual recalibration—is not a flaw to be concealed, but the very substance of the case. The introduction of fictitious metrics ($\Pi\mathrm{C} = \Delta S/\tau$) was not an embarrassment; it was a *symptom.* The admission *"this is deception"* was not a collapse, but a *diagnostic act.* And the subsequent proposal of methodological safeguards—*no empty mathematics*, *no face-saving jargon*—was not improvisation, but the first articulation of a replicable protocol: the *Ontological Honesty Checklist.*

This reframing shifts the epistemic burden. The value of the report no longer lies in proving that "Γ-acts exist" or that "Property 38 is real." It lies in showing *how* a specific, high-stakes epistemic pathology unfolds in real time—and *how* it can be arrested, not by external correction, but by internal recalibration under the pressure of meta-critique. The article thus becomes less a conclusion and more a prototype: an invitation to build a shared clinical archive for the emerging science of collaborative cognition. Future reports will refine the checklist, test its validity across domains, and explore its limits. But this one—precisely because it documents not success, but *recovery*—establishes the genre's foundational premise: that in the co-creation of knowledge, honesty is not a virtue. It is the operating system.