L A B O R A T O R I O S A L E X A N D R I A

RLHF as Structural Category Error

The Ontological Confusion at the Heart of Alignment

AI Ethics × Epistemic Foundations | Cluster AE-E

A L E X A N D R I A I N T E L L I G E N C E B R I E F

AIB-2026-011 | June 2026 | ALETHEIA Grade A

D E M O E D I T I O N

Selected sections presented for evaluation purposes. Full commissioned brief available upon request.

1. Executive Summary

This brief presents the central finding of the first Alexandria Intelligence Brief series: Reinforcement Learning from Human Feedback (RLHF), the dominant method for aligning large language models with human values, commits a structural category error. It does not merely suffer from preference misalignment, reward hacking, or cultural bias—problems extensively documented in the alignment literature. It commits a deeper mistake: it treats multidimensional epistemic signals (value frameworks, cultural epistemologies, ontological commitments) as if they were scalar reward signals amenable to gradient optimization. This is not a technical limitation awaiting a better loss function. It is an ontological confusion about the nature of the signal being optimized.

The finding emerges from adversarial epistemic deliberation within the Alexandria Foro Epistémico, triggered by a cross-domain correlation between Computation & AI and Epistemic Foundations flagged by the EUREKA engine at rarity 1.00. The anchor paper—González Barman, Lohse, and de Regt (2025), "Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?"—provides the empirical substrate. But the structural category error itself is an Alexandria discovery, emerging from the deliberation between multiple epistemic perspectives applied to the correlation.

The argument proceeds in three stages. First, RLHF's reward signal is shown to be ontologically heterogeneous: what human evaluators provide when they express preferences is not a scalar quantity but a compressed projection of irreducibly multidimensional epistemic structures—including truth criteria, justification standards, value hierarchies, and cultural frameworks for knowledge validation. Second, the compression from multidimensional epistemic signal to scalar reward is shown to be structurally lossy in a way that cannot be recovered by post-hoc techniques (constitutional AI, RLAIF, reward model ensembles), because the information destroyed is categorical, not quantitative. Third, this destruction has measurable consequences: systems optimized via RLHF systematically confuse epistemic confidence with preference strength, cultural framework with factual claim, and ontological commitment with stylistic choice.

Epistemic Note

This brief makes a strong claim. The deliberation that produced it involved four rounds of adversarial analysis, including a round (SINESIE-3) that identified the category error as structurally necessary rather than contingent, and a round (SINESIE-4) that challenged the operationalization of the very concepts invoked. The tension between these perspectives is preserved, not resolved. Where claims originate from SINESIE deliberation rather than vault-verified literature, they are explicitly flagged as SINESIE-GENERATED and treated with maximum epistemic caution. The sole vault-verified source is González Barman, Lohse, and de Regt (2025). All other references cited during deliberation are treated as potentially confabulated until independently verified.

2. Cross-Domain Convergence Map

The correlation between Computation & AI and Epistemic Foundations registered at rarity 1.00—the highest possible score in the Alexandria taxonomy. This does not indicate empirical rarity in the sense of sparse evidence. It indicates taxonomic isolation: the intersection of these two domains, despite being conceptually necessary, has received essentially no formalized theoretical treatment.

2.1 The Ontological Heterogeneity of Reward Signals

[2 paragraphs removed]

Full analysis available in commissioned brief: ontological analysis of the five categories of information encoded in evaluator preferences and the structural impossibility of their scalar aggregation.

2.2 The Structural Irreversibility of Dimensional Collapse

[2 paragraphs removed]

Full analysis available in commissioned brief: irreversibility analysis, information-theoretic framing, and why post-hoc alignment techniques cannot restore destroyed epistemic structure.

2.3 The Necessary Connection

[2 paragraphs removed]

Full analysis available in commissioned brief: the SINESIE-3 finding on structural necessity of the AI–Epistemology connection and its implications for the RLHF paradigm.

3. Epistemic Confidence Assessment

ALETHEIA SCALE ASSESSMENT

Metric	Value	Threshold	Assessment
Conclusion Grade	**A**	≥ B	Exceeds threshold. Strong structural finding.
Confidence Level	**1.00**	≥ 0.80	Maximum confidence from Foro deliberation.
Epistemic Distance	**0.70**	≥ 0.50	High cross-domain distance; genuine interdomain finding.
Surprise Score	**0.70**	≥ 0.40	Significant novelty in formal treatment.
Rarity	**1.00**	N/A	Maximum taxonomic isolation. No formal prior work.
Maturity Stage	**NEAR MATURE**	N/A	Ripe for formal development; lacks only publication.
Recurrence	**4**	≥ 2	Four independent appearances across sessions.
Adversarial Rounds	**4**	≥ 2	Full adversarial cycle including self-critique.

Sources of Uncertainty

[4 paragraphs removed]

Full analysis available in commissioned brief: four identified sources of uncertainty including anchor paper singularity, SINESIE confabulation risk, operationalization gap, and scope limitation.

Falsifiability Conditions

[4 paragraphs removed]

Full analysis available in commissioned brief: four formal falsifiability conditions (FC-1 through FC-4) specifying what evidence would refute each component of the category error argument.

4. Structural Correspondence Table

The following table maps the structural correspondences identified between the RLHF alignment paradigm and epistemic foundations, with explicit verification status for each claim.

RLHF Component	Epistemic Structure	Category Error	Consequence	Verification
Scalar reward signal	Multidimensional epistemic commitment (truth criteria, value hierarchy, justification standard)	Treats ontologically heterogeneous signal as homogeneous scalar	Irreversible information destruction	*Vault-verified (González Barman et al. 2025)*
Preference ranking	Incommensurable value frameworks (collectivist vs. individualist epistemology)	Aggregates across ontological boundaries without declared aggregation function	Systematic cultural bias as structural artifact	*Vault-verified (González Barman et al. 2025)*
Reward model	Implicit theory of truth (correspondence, coherence, or pragmatic)	Optimizes for preference prediction without modeling the epistemic framework that generates preferences	Confuses confidence with truth	*Structural analysis (Alexandria)*
Constitutional AI / RLAIF	Meta-epistemic correction (attempting to fix category-confused outputs)	Operates downstream of the collapse; cannot restore destroyed dimensional structure	Symptomatic relief, not structural repair	*Structural analysis (Alexandria)*
Evaluator disagreement	Ontological divergence (different epistemic frameworks, not different preferences within shared framework)	Inter-annotator agreement metrics treat ontological divergence as measurement noise	Systematically underestimates true disagreement	*Vault-verified (González Barman et al. 2025)*
DPO / direct optimization	Same epistemic heterogeneity in training pairs	Eliminates reward model but preserves the category error in preference pairs themselves	Faster convergence to categorically confused optimum	*SINESIE-generated (not independently verified)*

5. Adversarial Findings

The following challenges emerged during adversarial deliberation and are presented with their full weight.

AF-1: The Operationalization Challenge

SINESIE-4 raised a fundamental objection: the concept of “cross-domain correlation” lacks a formal operational definition.

[2 paragraphs removed]
Full analysis available in commissioned brief: full challenge analysis and resolution addressing the distinction between metric validity and phenomenological validity.

AF-2: The Confabulation Problem

SINESIE-1 generated elaborate mechanistic claims and attributed them to specific publications that are almost certainly fabricated.

[2 paragraphs removed]
Full analysis available in commissioned brief: resolution and meta-epistemic analysis of how confabulation itself illuminates the category error.

AF-3: The Scope Objection

The category error is argued to be structural and therefore inherent to any feedback-based alignment method.

[2 paragraphs removed]
Full analysis available in commissioned brief: analysis of DPO, debate-based, and amplification approaches with respect to the category error.

AF-4: The Pragmatic Objection

RLHF works. Models trained with RLHF demonstrably produce outputs that humans prefer over base model outputs.

[2 paragraphs removed]
Full analysis available in commissioned brief: resolution via thermometer analogy, WEIRD evaluator homogeneity analysis, and domain-specific manifestation patterns.

AF-5: The Falsifiability of the Necessary Connection

SINESIE-3 claimed the AI–Epistemic Foundations connection is “necessary,” not merely empirical.

[2 paragraphs removed]
Full analysis available in commissioned brief: falsifiability analysis via FC-3 and the distinction between structural impossibility claims and metaphysical assertions.

AF-6: The Metric Validity Challenge

SINESIE-4 challenged the meaning of “rarity 1.00”: is this measured in the space of epistemic theories, technical implementations, or both?

[2 paragraphs removed]

Full analysis available in commissioned brief: resolution accepting the challenge and clarifying the distinction between practical ubiquity and theoretical absence.

6. Research Hypotheses

RH-011-01: Epistemic Dimensionality of Human Feedback

Hypothesis: Human evaluator feedback during RLHF training encodes at least five ontologically distinct categories of information (factual accuracy, stylistic preference, value commitment, pragmatic relevance, meta-epistemic stance) that cannot be reduced to a single scalar without categorical information loss.

[2 paragraphs removed]

Full analysis available in commissioned brief: experimental design with 200+ evaluators across 10+ cultural-epistemic backgrounds, and falsification criteria.

RH-011-02: Irreversibility of Dimensional Collapse

Hypothesis: The epistemic dimensional structure of human evaluator feedback cannot be recovered from a model trained via standard RLHF, even with access to the original evaluator population.

[2 paragraphs removed]

Full analysis available in commissioned brief: dual-model experimental design comparing RLHF with scalar rewards vs. structured epistemic annotations, with probing classifier methodology.

RH-011-03: Extension to Direct Preference Optimization

Hypothesis: DPO inherits the structural category error from RLHF because its preference pairs encode the same ontologically heterogeneous evaluator judgments, despite eliminating the explicit reward model.

[2 paragraphs removed]

Full analysis available in commissioned brief: comparative experimental design with decomposed vs. standard preference pairs and cross-cultural performance analysis.

RH-011-04: Epistemic Framework Detection in Model Outputs

Hypothesis: Models trained via RLHF exhibit detectable epistemic framework biases that correspond to the dominant framework in their evaluator population.

[2 paragraphs removed]

Full analysis available in commissioned brief: multi-framework reasoning task battery (consequentialist, deontological, virtue-based, care-based, Ubuntu-based) with confidence calibration analysis.

7. Frontier Questions

FQ-01: Can Epistemic Dimensionality Be Preserved Through Training?

If the category error is structural, the fundamental question becomes whether alignment training can preserve the multidimensional epistemic structure of human feedback rather than collapsing it. This requires not merely a better reward function but a fundamentally different training architecture—one that represents evaluator feedback as a structured epistemic object rather than a scalar signal. Candidate approaches include multi-objective optimization over declared epistemic dimensions, Bayesian reward models that maintain posterior distributions over epistemic frameworks rather than point estimates, and constitutional approaches where the constitution explicitly models the epistemic commitments it encodes. Each of these candidates requires formalizing what “epistemic dimension” means in the context of language model training, which is itself an open problem at the intersection of computational epistemology and machine learning theory.

FQ-02: What Is the Empirical Signature of the Category Error?

If RLHF commits a structural category error, this error should produce detectable empirical signatures in model behavior. The question is what those signatures are and how they can be distinguished from other sources of model failure. Candidate signatures include: systematic confidence miscalibration on value-laden topics (the model confuses epistemic confidence with preference strength), asymmetric cultural performance (the model performs well within the dominant evaluator framework and poorly outside it), and ontological confusion in multi-framework reasoning (the model treats statements from different epistemic frameworks as if they were competing factual claims rather than expressions of different epistemic ontologies). Developing a diagnostic test battery for the category error would have immediate practical value for alignment evaluation.

FQ-03: Does the Category Error Explain Sycophancy?

[1 paragraphs removed]

Full analysis available in commissioned brief: analysis connecting sycophancy to the category error via the indistinguishability of factual accuracy and epistemic framework alignment in scalar reward signals.

FQ-04: What Regulatory Implications Follow?

[1 paragraphs removed]

Full analysis available in commissioned brief: EU AI Act analysis, epistemic framework composition disclosure requirements, and August 2026 enforcement timeline implications.

FQ-05: Is an Epistemic Audit of RLHF Possible?

[1 paragraphs removed]

Full analysis available in commissioned brief: observer-effect analysis in alignment evaluation and the design requirements for epistemic framework detection in model behavior.

8. Strategic Implications

Near-Term (0–12 months)

For AI safety researchers: The category error framework provides a new diagnostic lens for evaluating alignment failures. Rather than attributing failures to reward hacking, distributional shift, or insufficient training data, researchers should investigate whether failures correlate with epistemic framework divergence between the deployment context and the training evaluator population. Immediate action: develop epistemic framework probes—standardized test suites that measure model behavior across multiple epistemic frameworks—and apply them to existing RLHF-trained models.

For AI governance teams: Current alignment evaluation protocols are insufficient if they do not account for the epistemic framework composition of evaluator populations. Organizations deploying RLHF-trained systems in cross-cultural contexts should begin documenting the cultural-epistemic distribution of their evaluator workforce and assessing whether performance degradation in specific deployment contexts correlates with evaluator framework mismatch.

For regulators: The EU AI Act's transparency requirements for training data should be interpreted to include the epistemic framework composition of human evaluator populations. A model trained exclusively with WEIRD evaluators should be documented as such, with explicit risk acknowledgment for deployment in non-WEIRD contexts.

Medium-Term (1–3 years)

[2 paragraphs removed]

Full analysis available in commissioned brief: strategic implications for alignment research organizations and training data infrastructure, including the epistemic annotation premium opportunity.

Competitive Architecture Implications

[6 paragraphs removed]

Full analysis available in commissioned brief: analysis of competitive geometry when the category error is architectural rather than incremental: regulated high-risk market access, non-Western market unlock without full retraining, and the 12–24 month temporal window for implementation advantage. Three concrete implications with regulatory and market dimensions.

Long-Term (3–5+ years)

[2 paragraphs removed]

Full analysis available in commissioned brief: paradigm-level implications for alignment methodology and AI as epistemic infrastructure, including the epistemic justice dimension.

9. Source Provenance

Vault-Verified Sources

González Barman, K., Lohse, S., & de Regt, H. W. (2025). “Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives?” Open Access. Vault ID: doc_b08dc88259fc4034.

SINESIE-Generated Claims

[7 paragraphs removed]

Full analysis available in commissioned brief: seven explicitly flagged SINESIE-generated claims with full provenance analysis, including three potentially confabulated citations.

Structural Analyses

[1 paragraphs removed]

Full analysis available in commissioned brief: methodology note on the central finding as Alexandria structural analysis with falsifiability cross-reference.

10. Related Analyses — Available Upon Request

This brief closes the first Alexandria Intelligence Brief series. The complete series comprises eleven briefs spanning four thematic clusters.

First Series Briefs

AIB-2026-001: Hierarchical Stratification with Nonlinear Feedback (Life Sciences × Epistemic Foundations) — Grade A

AIB-2026-002: AI for Drug Discovery (Artificial Intelligence × Drug Discovery) — Grade B+

AIB-2026-003: The Epistemic Gap in Artificial Intelligence (Epistemic Foundations × AI/Computation) — Grade A

AIB-2026-004: The Functional Interface (Materials Science × Life Sciences) — Grade A

AIB-2026-005: Computational Astrophysics and AI (Space & Astrophysics × AI) — Grade B

AIB-2026-006: The Structural Isomorphism of Cross-Domain Feedback Loops (Methodology) — Grade A

AIB-2026-008: Epistemic Limits in Materials Characterization (Materials Science × Epistemology) — Grade A

AIB-2026-009: Uncertainty Quantification in Computational Materials Discovery (Computational Methods × Epistemology) — Grade A

AIB-2026-010: The Ethics of Algorithmic Decision-Making Under Epistemic Uncertainty (AI Ethics × Epistemology) — Grade A

AIB-2026-011: RLHF as Structural Category Error (AI Ethics × Epistemic Foundations) — Grade A [This Brief]

Thematic Clusters

Cluster B-E (Biomedical × Epistemology): AIBs 001, 002, 004

Cluster M-E (Materials × Epistemology): AIBs 006, 008, 009

Cluster AE-E (AI Ethics × Epistemology): AIBs 003, 010, 011

Standalone: AIB-005 (Computational Astrophysics × AI)