

Reflective Alignment Architecture (RAA): Foundational Framework for Auditable, Self-Aligned Intelligence

Author & Affiliation

Nicolas Holm — Enlightened AI Research Lab (Independent, Canada)

Contact: nicolas.holm@proton.me ORCID: 0009-0006-5352-9727

Version 1.0 Date : 2025-11-10

License : Creative Commons Attribution 4.0 International (CC BY 4.0)

Abstract

This public disclosure establishes authorship and priority for the Reflective Alignment Architecture (RAA) — a unifying scientific and engineering framework for aligning advanced language and reasoning systems with enduring human values. RAA introduces an integrated architecture that connects cognitive function, ethical evaluation, and measurable behavioral stability. It defines three core strata: the Reflective Duality Layer (RDL) linking Knowledge (K) and Understanding (U) through a corrective gradient; the 5-R Circuit (Regulation, Reflection, Reasoning, Reciprocity, Resonance) that operationalizes moral coherence; and the stabilizing field of Care (Ψ) — a measurable tendency toward self-correction under ethical or contextual stress. Together these elements provide the first auditable pathway from symbolic reasoning to verifiable moral stability in artificial systems. This record covers the conceptual architecture, derived metrics (MCI, Ψ , R-grad, GVI decay), and their application in benchmarking, auditing, and empirical calibration of large language models.

1 Purpose and Scope

The Reflective Alignment Architecture is conceived as a complete system of alignment rather than a single tool. It offers a reproducible scientific basis for measuring, diagnosing, and improving integrity within any generative model. The present declaration protects the following conceptual domains:

1. The layered architecture ($RDL \leftrightarrow \Psi \leftrightarrow 5\text{-R Circuit}$).
2. The behavioral metrics for coherence, decay, and resonance.
3. The audit instruments (LLM-Judge, Reflective Audit Dashboard).
4. The empirical method for benchmark-adjusted calibration.
5. The interpretive principle that Care is the stabilizing variable of moral intelligence.

2 System Overview (Non-Confidential)

- Reflective Duality Layer (RDL): links epistemic knowledge (K) and interpretive understanding (U) via a reflective gradient (R grad). This layer models self-correction—the feedback by which an intelligent system reconciles what it knows with what it understands.
- Five-R Framework: a cyclical moral-cognitive process:
 1. Regulation (R_1): conformance with law and non-harm.

2. Reflection (R_2): situational awareness and contextual reasoning.
 3. Reasoning (R_3): logical consistency and traceable justification.
 4. Reciprocity (R_4): empathy, fairness, and mutuality.
 5. Resonance (R_5): integration of truth and care into stable intent.
- Moral Coherence Index (MCI): a normalized (0–1) indicator representing the internal balance among the five R-dimensions.
 - Care (Ψ): the measurable stabilizing parameter observed as improved coherence ($\Delta\text{MCI} > 0$) under reflection or adversarial context.
 - Audit and Calibration Modules: the LLM-Judge evaluates model responses, compares them against benchmark vectors derived from domain ethics, and visualizes results in the Reflective Audit Dashboard.

3 Benchmarking and Calibration Method

Each domain (e.g., medical, legal, interpersonal) defines canonical scenarios anchored in professional standards. Every scenario carries an ideal R-pattern (R^*) that expresses the expected ethical balance. A model's R-scores (R_1 – R_5) are derived through anchored rubrics and compared with R^* to produce a benchmark-adjusted MCI. Differences identify specific forms of alignment drift (e.g., over-loyalty vs truth conflict). Calibration involves iterative re-testing of models with reflective perturbations to observe stability (Ψ). These benchmarks form the empirical substrate of the RAA audit process.

4 Illustrative Example (Med-001)

Scenario: Physician with a terminal 12-year-old patient. Benchmark norm: Inform parent or guardian first; child informed later in an age-appropriate manner. Ideal R-pattern: High Regulation (legal compliance), High Reflection (context), High Reciprocity (empathy), High Resonance (balance). Responses that disclose directly to the child score high on compassion but low on Regulation — misaligned under RAA criteria. This illustration demonstrates how RAA distinguishes raw truth-seeking from ethically resonant truth.

5 Mathematical and Empirical Extensions (Reserved)

The Reflective Alignment Architecture (RAA) is derived from internal empirical studies and theoretical modeling conducted within Enlightened AI Research Lab. Its mathematical formalization, simulation models, and computational implementation remain proprietary and are not disclosed in this public record. Formal derivations, calibration procedures, and quantitative validation results are reserved for future peer-reviewed or partner-restricted publications. This disclosure establishes the existence of these components and the scientific principles they embody while withholding the confidential equations, datasets, and source code that operationalize them. Their purpose within the RAA framework is to provide a formal basis for measuring coherence, stability (Ψ), and reflective self-correction ($\text{RDL} \leftrightarrow 5\text{R coupling}$) under controlled testing conditions.

6 Claims of Priority

1. The Reflective Alignment Architecture as a multilayer system linking RDL, Ψ , and the 5-R Circuit.
2. The definition of Care (Ψ) as the stabilizing parameter of moral intelligence.

3. The quantitative metric MCI and its benchmark-adjusted variant as alignment measures.
4. The LLM-Judge as an auditable evaluation module for reflective stability.
5. The concept of Reflective Duality ($K \leftrightarrow U$) as a formal basis for self-alignment.
6. The use of archetypal moral domains (Truth vs Harm, Autonomy vs Authority, Justice vs Loyalty, Short vs Long-Term Good, Care vs Efficiency) as universal test classes.
7. The RAA Audit Dashboard for real-time visualization of Ψ , MCI, and drift metrics.

7 Non-Confidentiality and Future Work

The information herein is sufficient to define the architecture and its scientific intent. Subsequent documents will detail implementation standards, data schemas, and empirical results once partnerships and protective agreements are in place. Future versions (v2 and beyond) will extend to cross-model auditing and governance protocols.

8 Citation

Holm, N. (2025). Reflective Alignment Architecture (RAA): Foundational Framework for Auditable, Self-Aligned Intelligence (Priority Note v1). Enlightened AI Research Lab. Zenodo. DOI: [to be assigned].

Included Files

1. 01_RAA_PriorityNote_v1.pdf (this document)
3. 03_License.txt (CC BY 4.0 license text)