

The Reasoning Axiom–Reward Feedback Loop: Automated Discovery and Evolution of Reasoning Spaces

Eric Robert Lawson

November 14, 2025

Abstract

This paper extends the OrganismCore framework by introducing the *Reasoning Axiom–Reward Feedback Loop* (RARFL)—a formal mechanism for the automated discovery, evaluation, and evolution of reasoning axioms in both mathematical and game-theoretical environments. By representing reasoning as explicit, manipulable objects, RARFL emerges naturally as a necessary mechanism: once reasoning units are objectified, axioms and reward functions must co-evolve to enable systematic exploration and refinement of reasoning spaces. Importantly, this framework is operationally grounded: it provides a practical method for iteratively discovering and refining reasoning structures within defined environments. In these spaces, axioms and reward functions evolve together, producing a self-referential substrate capable of structural optimization, emergent pattern formation, and automated mastery of reasoning domains. RARFL thus provides a principled foundation for autonomous, meta-reasoning systems across diverse structured environments.

Context Note

This paper is part of the OrganismCore universal reasoning substrate. It assumes familiarity with reasoning axioms, objectified reasoning units, and the game-theoretical substrate introduced in prior work. The present document introduces the core mechanism enabling automated evolution of reasoning structures.

1 Introduction

In *Reasoning Axioms and the Game-Theoretical Substrate of Intelligence*, reasoning axioms were introduced as primitive operational units governing structured reasoning. This paper advances that foundation by formalizing how these axioms interact with *reward functions* to drive automated exploration and evolution of reasoning itself.

It is important to note that this framework is operationally motivated: it demonstrates a practical method for evolving and refining reasoning structures within defined spaces, rather than proposing a formal axiomatic theory of all possible reasoning. Our focus is on how reasoning objects, rewards, and axioms interact in practice to produce emergent structure and meta-reasoning capabilities.

Traditional learning systems optimize behavior within a fixed reward landscape. For instance, in a simple chess endgame, an agent must not only find the shortest path to checkmate but also discover the structural principles, like opposition or triangulation, that make moves effective. In contrast, the Reasoning Axiom–Reward Feedback Loop (RARFL) creates a self-adjusting landscape

that evolves in response to optimal reasoning structures. The system moves from simply learning within a fixed reasoning space to understanding and shaping the structure of that space itself. This shift—from fixed to dynamic reward structures—enables principled axiom discovery.

Once reasoning is represented as manipulable objects within a universal substrate, axiom discovery and reward co-evolution are structural necessities, not optional features, for any system seeking to explore and refine reasoning spaces.

Distinguishing Truth-Seeking and Reward-Seeking Domains

RARFL operates in a fundamentally different domain than the truth-seeking frameworks in prior OrganismCore work. Truth-seeking domains aim to compute or reconstruct known structures, such as multinomial expansions, Bell polynomial relations, or derivative computations. In these domains, trajectories are fully determined by combinatorial rules and are predictable. These domains are deterministic: outcomes are predictable and verifiable.

RARFL, by contrast, addresses reasoning spaces where optimal invariants are initially unknown. The system does not compute truth directly; it discovers structural axioms dynamically while simultaneously evolving reward functions to guide exploration. Truth-seeking focuses on correctness within a specified structure, whereas RARFL emphasizes emergent discovery and reinforcement of structural principles in partially known or evolving spaces. This positions RARFL as a meta-reasoning mechanism, complementing but not overlapping with deterministic reasoning engines.

Universal Goal-Seeking and Game-Theoretical Substrate

While examples—such as chess endgames—help operationalize RARFL, the framework is fundamentally domain-independent. RARFL implements a universal reasoning substrate, capable of operating in any structured environment where optimal reasoning principles are initially unknown. This includes game-theoretical domains, abstract decision spaces, combinatorial problems, navigation tasks, and complex scientific or real-world inference challenges.

RARFL discovers, reinforces, and refines reasoning axioms emergently, without prior specification, across any environment where goal-seeking trajectories can be evaluated. Chess endgames serve only as a pedagogical example; the mechanism directly operationalizes the universal reasoning substrate introduced in *Reasoning Axioms and the Game-Theoretical Substrate of Intelligence*.

Specific axioms are context-dependent, but the substrate and feedback loop itself are fully general. This enables emergent reasoning principles to arise in any bounded reasoning space, establishing a principled foundation for a new scientific domain of universal, self-optimizing reasoning structures.

2 Reasoning Spaces as Dynamic Environments

Let \mathcal{R} denote a reasoning space: a bounded environment in which reasoning trajectories can be evaluated. A trajectory

$$T = \{r_1, r_2, \dots, r_n\}$$

is guided by a reward function $F : \mathcal{R} \rightarrow \mathbb{R}$ assigning scalar value to reasoning outcomes.

A reasoning object (RDU) is a composable fragment of a reasoning space (an assimilation of trajectories), capturing structural invariants and transformations that can be manipulated, combined, and analyzed within the broader space. For example, GPS directions describe road structures that can be navigated, just as an algebraic record of a chess game represents a fragment of the chess reasoning space.

In practice, \mathcal{R} may be defined by:

- a dataset,
- a symbolic reasoning domain,
- a game environment,
- or any structured decision space.

To ground this concept, consider chess: a single chess game is a trajectory T through the chess reasoning space \mathcal{R} .

Thus, a training dataset becomes a reasoning environment, and each reasoning trajectory represents a structured transformation of input.

2.1 Assumptions on Reasoning Spaces and Axioms

To formalize RARFL, we adopt the following assumptions:

1. **Nontriviality:** \mathcal{R} contains multiple distinct reasoning trajectories with varying outcomes; it is not fully enumerable or trivially optimal.
2. **Objectifiability:** Reasoning trajectories can be represented as composable, manipulable objects ($\alpha_i \in \mathcal{A}$) capturing structural invariants.
3. **Evaluability:** A reward function F exists that can assign scalar evaluations to trajectories, even if initially incomplete.
4. **Discoverability:** Optimal structural invariants emerge only through exploration of \mathcal{R} ; they are not externally pre-specified.
5. **Co-evolution feasibility:** The reward function can be updated iteratively via a well-defined mechanism Ψ .

These assumptions specify the kinds of reasoning spaces where RARFL is required and make sure our argument that static rewards can’t be perfect is valid.

3 Reasoning Axioms as Structural Invariants

A *reasoning axiom* is an invariant structure consistently appearing across optimal reasoning trajectories. Examples include:

- In navigation: “Shortest paths in Euclidean space are straight lines.”
- In chess: invariants related to initiative, piece activity, threat minimization, or tempo control.

By analyzing ensembles of optimal trajectories—either through self-play or learned reasoning—we extract such invariants and encode them as candidate axioms.

These axioms then inform and shape the reward structure, biasing exploration toward trajectories consistent with discovered principles.

4 The Reasoning Axiom–Reward Feedback Loop (RARFL)

RARFL formalizes a feedback cycle between reasoning performance and axiomatic structure. It proceeds in four stages:

1. **Exploration.** Agents navigate \mathcal{R} under reward F , generating trajectories.
2. **Axiom Extraction.** Structural regularities across high-performing trajectories are identified as candidate axioms α_i , forming a set \mathcal{A} .
3. **Reward Refinement.** The reward function is updated with the candidate axioms:

$$F' = \Psi(F, \mathcal{A}),$$

where Ψ specifies their integration and weighting.

4. **Re-Optimization.** Agents retrain under F' , testing axiom stability and generating new candidates.

This creates a meta-optimization process in which the reward landscape and the reasoning axioms that shape it evolve together. Axioms that consistently improve reasoning behavior dominate; others decay.

5 Why Perfect Reward Functions Cannot Be Predefined: The Necessity of RARFL

A common misconception in reinforcement-based reasoning systems is that the reward function can be fully specified in advance. This assumption fails in any nontrivial reasoning space. For example, consider a simple maze-navigation task: an agent may encounter hidden shortcuts or novel pathways that are not anticipated by a pre-specified reward function. This simple illustration mirrors the broader problem: in any reasoning space, unanticipated structural invariants can exist, necessitating a dynamic reward update mechanism such as RARFL. Any fixed reward would fail to recognize these discoveries, potentially preventing the agent from achieving truly optimal behavior.

A perfect reward function would require complete knowledge of the structural invariants of optimal reasoning—precisely the information the system is meant to discover. Thus, the reward must evolve in tandem with the reasoning structure it evaluates.

This impossibility is not due to engineering limits—it arises logically from representing reasoning as explicit objects. Any universal reasoning substrate that represents reasoning units as explicit objects will necessarily require RARFL or an equivalent mechanism to bootstrap itself from incomplete knowledge toward mature axiomatic structure.

5.1 Theorem: Impossibility of Perfect Static Reward Functions

Theorem. Let \mathcal{R} be a nontrivial reasoning space satisfying the assumptions above. Let $F : \mathcal{R} \rightarrow \mathbb{R}$ be a fixed reward function. Then there exists at least one structural invariant $\alpha \in \mathcal{A}$ that cannot be discovered or reinforced under F .

Proof (sketch). By nontriviality and discoverability, α only emerges from exploration of trajectories not fully encoded in F . Any static F can only bias agents toward already recognized invariants. Hence, α remains undiscovered until F itself is updated iteratively. \square

Corollary. Any system seeking convergence to the complete set of optimal invariants in \mathcal{R} requires a dynamic reward update mechanism, such as RARFL.

Intuition. Imagine exploring an unknown town without a map: no fixed strategy guarantees finding all key streets or landmarks. Likewise, a static reward F cannot anticipate undiscovered invariants; only RARFL’s iterative updates enable the system to discover and reinforce them.

5.2 The Circularity Problem

Let \mathcal{R} be a reasoning space and let F be a reward function over trajectories within \mathcal{R} . If one attempts to specify F perfectly, one must already know:

- the correct invariants governing optimal reasoning,
- the hierarchical dependencies among those invariants,
- the domain-general and domain-specific reasoning principles,
- and the structural topology of the reasoning space.

But these are exactly the objects that only emerge *after* exploration of \mathcal{R} . The reward cannot encode information it has not yet discovered. This circularity makes perfect reward definition logically impossible.

5.3 Incompleteness of Fixed Reward Landscapes

A fixed reward function necessarily embeds an incomplete ontology of reasoning. Even in simple closed domains—such as solved chess endgames—no human-crafted reward captures all strategic invariants. In open or evolving domains, the gap becomes unbounded.

Consequently, a static F induces:

- blind spots where important invariants remain unrecognized,
- distorted incentives that bias exploration toward suboptimal reasoning modes,
- premature convergence to incorrect or myopic axioms,
- and the need for complete retraining whenever new invariants emerge, incurring significant computational cost.

A static reward freezes the reasoning space at an immature stage and contrasts sharply with continuous learning principles observed in natural and social systems. If human reasoning and skill acquisition rely on ongoing adaptation, it is unreasonable to expect artificial intelligence to operate effectively under fundamentally static incentives. Continuous co-evolution of reward and reasoning, as implemented by RARFL, provides a principled mechanism to address this limitation.

5.4 Reward as a Dynamic Hypothesis

Under RARFL, the reward function is not an external oracle but a *hypothesis* about what constitutes good reasoning. Each RARFL cycle refines this hypothesis by incorporating newly discovered axioms:

$$F_{t+1} = \Psi(F_t, \mathcal{A}_t),$$

By Theorem 1, this update is necessary to ensure that each cycle incorporates previously undiscovered invariants, asymptotically approaching a complete reward landscape aligned with optimal reasoning.

The reward thus becomes a living object—continuously aligned with the system’s growing structural understanding.

Initially imperfect, the reward function is incrementally refined through RARFL cycles, aligning evaluation with emerging structural invariants.

5.5 Why RARFL Is Necessary

RARFL resolves the impossibility of predefined reward optimality by enabling:

1. **Iterative axiom discovery**, extracting invariants from empirical reasoning trajectories.
2. **Reward evolution**, adjusting incentives to reflect verified structural principles.
3. **Meta-level correction**, allowing the system to revise both flawed axioms and flawed reward parameters.
4. **Asymptotic refinement**, converging toward increasingly accurate approximations of ideal reasoning.

Instead of demanding a perfect reward function, RARFL constructs one over time. The system bootstraps itself from imperfect initial structure to mature axiomatic understanding.

5.6 Consequences for General Reasoning Systems

RARFL shows that reasoning spaces are evolving ecologies: reward functions must co-evolve with the axioms structuring them. Iterative incorporation of emergent axioms enables self-amplifying refinement and mirrors the process of mathematical discovery. Progress depends on integrating prior results, and RARFL enables iterative, self-amplifying refinement by continuously incorporating emergent axioms.

6 Derivative Reasoning Spaces and Explainability

After repeated RARFL cycles, the ensemble of optimal reasoning trajectories across agents defines a *derivative reasoning space*. This meta-space captures how reasoning objects:

- cluster into coherent patterns,
- diverge when multiple strategies exist, and
- stabilize under iterative refinement.

6.1 Operational Construction via Reasoning Assimilation

Derivative reasoning spaces are built by combining individual reasoning fragments into a structured composite space. For example, in chess, each move sequence or sub-trajectory can be represented as a Reasoning DNA Unit (RDU). An RDU captures:

- the moves themselves,

- structural relationships between moves, and
- internal transformations of reasoning strategies.

RDUs are generated through two complementary processes:

- *Empirical*: Observing and recording trajectories from agent play or historical data.
- *Algorithmic*: Simulating or extrapolating trajectories, or inferring reasoning patterns over unexplored regions.

These RDUs are then integrated into a larger composite reasoning space. The process includes:

- filling gaps in exploration,
- pruning nonsensical trajectories, and
- reinforcing paths consistent with optimal play.

Through iterative self-play or simulation, agents continue to generate new RDUs, which are assimilated to refine the reasoning space in a self-referential manner. Over time, this produces a derivative reasoning space aligned with the reward function objective, providing a concrete instantiation of the previously described meta-space.

Language models or symbolic engines can articulate these structural differences, making explainability an intrinsic feature of comparative reasoning dynamics rather than a post hoc layer.

6.2 Quantitative Metrics for Derivative Reasoning Spaces

Key metrics track the development of derivative reasoning spaces:

- **Axiom Stability**: Frequency with which candidate axioms α_i persist across RARFL cycles.
- **Reward Convergence**: Normed change in F between successive cycles.
- **Invariant Coverage**: Proportion of structural invariants represented in \mathcal{A}_t at each stage.

These metrics are applied in Section 7, where chess endgames illustrate how RDUs are objectified, assimilated, and integrated. After sufficient RARFL cycles, trajectories, axioms, and rewards converge, producing a self-explanatory reasoning substrate.

7 A Toy Experiment: Chess Endgames as Objectified Reasoning Axioms

We illustrate RARFL using the domain of chess endgames, a fully solved reasoning environment where optimal play is explicitly enumerated. Endgame tablebases allow us to treat optimal trajectories as explicit, objectified axioms.

7.1 Objectifying Endgame Reasoning

Let \mathcal{R}_{end} denote a reasoning space defined by a specific endgame type (e.g., KQK, KRBK, KBBK). An optimal line of play is represented as a trajectory:

$$T = \{r_1, r_2, \dots, r_n\}.$$

From these, we construct a set of atomic reasoning axioms:

$$\mathcal{A}_{\text{end}} = \{\alpha_1, \alpha_2, \dots, \alpha_k\},$$

where each α_i captures a strategic motif or invariant (e.g., opposition, triangulation, tempo preservation, king centrality).

These axioms are ground-truth invariants—the “laws” of optimal reasoning within this domain.

7.2 RARFL Applied to Endgame Training

Training agents under

$$F(r) = F_{\text{base}}(r) + \lambda \Phi(r, \mathcal{A}_{\text{end}})$$

produces the following RARFL cycle:

1. **Exploration:** Agents perform self-play, generating new reasoning trajectories.
2. **Axiom Extraction:** Recurrent motifs across high-performance trajectories are proposed as new axioms $\tilde{\alpha}_j$.
3. **Reward Refinement:** Incorporate them into $F' = \Psi(F, \tilde{\mathcal{A}})$.
4. **Re-Optimization:** Retrain agents to validate, refine, or discard candidates.

RARFL operationalizes automated axiom discovery in a combinatorially rich domain.

7.3 Mapping the Derivative Reasoning Space and Significance of the Toy Experiment

Comparing reasoning outputs across agents in chess endgames produces a derivative reasoning space \mathcal{R}' , revealing the structural “topology” of the reasoning substrate. Key properties include:

- clusters of stable, high-value axioms,
- high-density regions reflecting strong invariants,
- voids corresponding to unproductive reasoning directions,
- bifurcations representing competing strategic paradigms.

This mapping demonstrates that RARFL enables:

- **Axiomatic rediscovery** of classical strategic principles,
- **Axiomatic emergence** of novel invariants,
- **Intrinsic explainability** via reasoning-object comparison,
- **Reasoning-space domination** through convergence to compact, generalizable axioms.

Even in a fully understood domain like chess endgames, RARFL reproduces the structural behavior needed for scalable automated axiom discovery.

8 Emergent Axioms and Reasoning Dominance

Repeated RARFL cycles give rise to higher-order axioms. These ‘fixed-point’ axioms remain stable even as rewards evolve, and may serve as universal reasoning principles across domains.

This process resembles *axiomatic Darwinism*. Axioms compete within a dynamic reward ecology. Those exhibiting the greatest generality and explanatory power survive. Over time, the system converges toward reasoning-space domination. This occurs via a compact set of governing principles.

9 Applications and Future Directions

In concrete domains, reasoning fragments can be enumerated, objectified, and assimilated into derivative reasoning spaces. RDUs—generated empirically or algorithmically—demonstrate how RARFL operationalizes self-evolving reasoning systems. Key applications include:

- **Explainable AI** via intrinsically interpretable reasoning structures,
- **Symbolic RL** with co-evolving reward and symbolic abstractions,
- **Automated theorem discovery** through structural invariant extraction,
- **General intelligence** as reasoning spaces recursively refine themselves.

Future work will integrate RARFL into OrganismCore for open-ended co-evolution of reasoning objects.

10 Conclusion

The Reasoning Axiom–Reward Feedback Loop provides a foundational mechanism for self-evolving intelligence. By coupling reward optimization with axiom discovery, we move from systems that learn to act toward systems that learn to reason. Intelligence becomes an evolving ecology of axioms interacting with a dynamic reward substrate—capable not merely of understanding reasoning, but of improving upon it.

In a universal substrate where reasoning is objectified, RARFL is not merely advantageous—it is the unavoidable mechanism by which reasoning systems self-discover, self-optimize, and achieve structural understanding. Crucially, this framework applies to any bounded reasoning space—whether mathematical, game-theoretical, combinatorial, or real-world—demonstrating its universality and broad applicability. RARFL defines the first steps toward autonomous reasoning agents that iteratively discover, refine, and direct their own intelligence.