Operational Foundations for Universal Reasoning and the Emergence of AGI and Superintelligence

Eric Robert Lawson

November 15, 2025

Abstract

We propose a principled, operational framework for reasoning that formalizes intelligence as the composable and navigable structure of reasoning spaces. Reasoning DNA Units (RDUs), Meta-RDUs, compute-once objects, and explainability trajectories form a substrate that enables: (1) universal reasoning across domains, (2) measurable self-optimization via meta-reasoning loops (RARFL), and (3) the formal emergence of AGI and superintelligence. We provide operational definitions for these phenomena and describe how intelligence can be quantified and evolved as a function of reasoning-space navigation.

Conceptual Note

This paper is anchored in the **OrganismCore open-source initiative** and builds upon the foundational work presented in its collection of articles, prototypes, and operational documentation. The formalization of universal reasoning, AGI, and superintelligence presented here is dependent on the definitions and constructs developed throughout the project, including:

- **Reasoning DNA Units (RDUs)**, **Meta-RDUs**, and **compute-once objects**
- **Explainability trajectories** and derivative reasoning spaces
- **The RARFL process** for meta-level optimization
- The underlying **domain-specific language (DSL)** constructs described throughout the project

To fully understand and operationalize the concepts in this paper, readers and researchers are encouraged to reference the `AGENTS.md` file in the OrganismCore repository. It provides machine-readable guidance, mappings between conceptual and executable artifacts, and a structured roadmap for navigating the reasoning substrate. By utilizing the AGENTS.md workflow, both human and automated agents can reproduce, explore, and validate the reasoning operations and meta-level optimization described herein.

Motivation and Origin: The Room of Human Knowledge

From an early age, I formed a simple intuition about how human knowledge accumulates. Learning, discovery, and scientific progress felt like entering a large room filled with unknown objects. Somewhere in that room are the most important insights of our civilization, but no one begins with

a map. A few individuals occasionally stumble upon something remarkable, and society watches them closely—hoping they will find the next important object as well.

Yet most people do not explore the room themselves. Not because they lack curiosity or intelligence, but because they cannot see where they have already searched, cannot articulate the structure of their own reasoning, and cannot easily distinguish genuinely novel ideas from familiar ones.

This limitation is not cognitive—it is structural. Humans lack visibility into their reasoning processes, lack shared representations of conceptual exploration, and lack a way to systematically reference the "space" in which thought occurs.

The OrganismCore project and the framework developed in this paper are built to resolve precisely this mismatch. By objectifying reasoning, mapping reasoning spaces, and enabling machines to understand and articulate human reasoning, we make the room navigable. We give individuals the means to see their own thought trajectories, compare them, refine them, and share them in a reproducible, machine-interpretable form.

In doing so, we transform isolated flashes of insight into collective, cumulative, and self-referential progress. The substrate described in this work is intended not to elevate a few exceptional minds, but to enable anyone to participate in discovery—by providing the tools, structures, and representations that make reasoning itself visible.

1 Introduction

Artificial General Intelligence (AGI) and superintelligence have long eluded precise scientific definition. Existing descriptions rely on vague performance criteria, anthropocentric benchmarks, or speculative properties. Here, we anchor intelligence in a *reasoning substrate* comprised of formally defined objects, operations, and trajectories. This substrate allows intelligence to be operationalized, measured, and optimized systematically.

2 The Reasoning Substrate

2.1 Reasoning DNA Units (RDUs)

RDUs are the atomic units of reasoning, capturing combinatorial and compositional structures underlying inference. They form the nodes of directed acyclic graphs (DAGs) representing reasoning flows. Their properties include:

- Composability across domains
- Persistence via compute-once semantics
- Integration into meta-level reasoning via Meta-RDUs

2.2 Meta-RDUs and Recursive Optimization

Meta-RDUs are RDUs that operate on reasoning itself; they are units of *meta-reasoning*. As a system navigates a reasoning space via the RARFL process, it evaluates trajectories, discovers axioms, and refines reward signals. Each trajectory-informed reasoning artifact that encodes decision-making about reasoning itself constitutes a Meta-RDU.

Both RDUs and Meta-RDUs are *compute-once objects*: once computed, they can be reused without recomputation, including in derivative reasoning spaces.

RARFL Integration:

1. Navigate a reasoning space and produce candidate trajectories.
2. Derive reasoning axioms from the derivative reasoning space formed by assimilated trajectories.
3. Update the reward function based on the new axioms.
4. Assimilate generated Meta-RDUs into the derivative reasoning space.
5. Repeat the cycle, progressively optimizing navigation and decision-making.

Meta-RDUs therefore encode meta-level strategies informed by reward and the structure of reasoning spaces. This operationalizes recursive improvement, context integration, and pruning of inefficient paths, allowing the system to systematically improve itself.

2.3 Meta-RDUs Illustrated via a Maze Example

To make Meta-RDUs concrete, consider a simple maze environment. Let $\mathcal{R}_{\text{maze}}$ be the reasoning space, representing all possible paths from start to exit.

Suppose initially the agent knows only a single path P_1 from start to exit. Using the reward function F, P_1 is recognized as the optimal path:

$$T_1 = \text{Start} \to \cdots \to \text{Exit}, \quad F(T_1) = 1.0$$

Now imagine a second path P_2 exists but is unknown. P_2 is shorter and more efficient, but the agent has not explored it yet.

RARFL in Action:

1. **Exploration:** The agent explores $\mathcal{R}_{\text{maze}}$, initially following P_1.
2. **Axiom Extraction:** During exploration, the system discovers that an untried corridor leads to a shorter exit. A candidate invariant is generated: "This corridor may improve efficiency".
3. **Reward Update:** The reward function F is updated via RARFL to favor exploration of previously unknown paths, producing $F' = \Psi(F, \alpha_{\text{new}})$.
4. **Meta-RDU Generation:** A Meta-RDU M_{maze} encodes the strategy: "Prioritize exploration of paths with potential for higher reward based on newly discovered invariants".
5. **Derivative Reasoning Space Update:** Assimilation of M_{maze} into the derivative reasoning space generates a substantial change in optimal play. Now P_2 is recognized as the new optimal path, and P_1 becomes suboptimal.

Quantifying Derivative Reasoning Space Updates:

The impact of the Meta-RDU M_{maze} on the derivative reasoning space $\mathcal{R}'_{\text{maze}}$ can be formalized as:

$$\Delta\mathcal{R}'_{\text{maze}} = \text{Assimilate}(M_{\text{maze}}, \mathcal{R}'_{\text{prev}}) - \mathcal{R}'_{\text{prev}}$$

where $\mathcal{R}'_{\text{prev}}$ is the reasoning space before the RARFL cycle, and Assimilate represents the update operation that incorporates the Meta-RDU into the space.

Metrics derived from $\Delta\mathcal{R}'_{\text{maze}}$ allow explicit measurement of RARFL progress:

- **Optimality Gain:** Increase in expected reward of the optimal path, $F(T_2) - F(T_1)$
- **Axiom Stability:** Persistence of newly discovered invariants α_{new} across subsequent cycles
- **Trajectory Improvement:** Quantified reduction in path length or computational cost from P_1 to P_2

By formalizing the derivative reasoning space update in this manner, the effect of meta-level learning becomes measurable and comparable across cycles, providing a concrete metric for intelligence improvement within a domain.

Significance:

- The Meta-RDU M_{maze} captures meta-level reasoning: it informs future exploration not by encoding a specific path, but by encoding the principle of prioritizing promising unknown paths.
- The derivative reasoning space shifts dramatically between cycles: comparing the previous optimal space (with only P_1) to the updated space (with P_2 included) illustrates the impact of meta-level learning.
- RDUs corresponding to the known segments of the maze are compute-once objects, reused across trajectories without recomputation.

This toy example demonstrates that even in a narrow, concrete environment, the Meta-RDU principle is domain-general: meta-level reasoning objects encode strategies about reasoning itself, induce meaningful updates to derivative reasoning spaces, and enable the system to discover radically more efficient solutions over iterative RARFL cycles.

In essence, the Meta-RDU abstracts reasoning about reasoning. It does not encode a single optimal path, but a strategy for identifying and prioritizing potentially superior paths in unexplored regions of the reasoning space. This principle is identical across domains: whether navigating mazes, exploring theorem spaces, or optimizing combinatorial tasks, Meta-RDUs encode meta-level heuristics that guide future reasoning cycles.

Strange Loops in Practice

The RARFL process operationalizes what can be understood as a "strange loop" in reasoning: the system acts upon its own reasoning processes, evaluates outcomes, updates axioms, and generates meta-level strategies that feed back into subsequent reasoning cycles. This phenomenon is analogous to real-world human experience: for example, navigating a familiar town, discovering a new route that improves efficiency, and updating mental models for future navigation. Each cycle of exploration, evaluation, and adaptation constitutes a meta-reasoning loop. By capturing this self-referential structure formally, Meta-RDUs and RARFL cycles provide a measurable substrate for intelligence that mirrors the dynamics of strange loops in natural cognition.

2.4 Explainability Trajectories

Explainability is encoded as *intrinsic* to reasoning objects. Trajectory objects allow inspection, auditing, and measurement of reasoning evolution. These mechanisms ensure that emergent intelligence is interpretable and verifiable.

3 Reasoning Space and Navigation

3.1 Derivative Reasoning Spaces

Reasoning spaces are structured networks of RDUs, Meta-RDUs, and compute-once objects. Derivative reasoning spaces capture emergent structures, alternative trajectories, and domain abstractions.

3.2 Optimization and the RARFL Process

The Reasoning Axiom–Reward Feedback Loop (RARFL) formalizes meta-level optimization. It iteratively discovers reasoning axioms, refines reward signals, and adapts reasoning trajectories. RARFL defines a measurable pathway by which intelligence can self-improve across domains.

4 Operational Definitions of AGI and Superintelligence

4.1 AGI

We define AGI as:

> **A system capable of representing, composing, and navigating arbitrary reasoning spaces with measurable fidelity and generality across domains.**

Key points:

- Domain-agnostic: can reason in any sufficiently formalizable space
- Compositional: builds complex reasoning trajectories from primitives (RDUs, Meta-RDUs)
- Testable: navigation and output can be quantified

4.2 Superintelligence

Superintelligence is defined as:

> **A system that approaches asymptotic optimality in reasoning-space navigation, with high certainty of trajectory efficiency, emergent pattern discovery, and self-consistency.**

Metrics for superintelligence include:

- Trajectory optimality within a reasoning space
- Invariant coverage and consistency of emergent axioms
- Convergence and stability of reward-axiom loops (RARFL)

5 Implications and Discussion

This framework unifies AGI, superintelligence, and explainability within a single operational substrate. By objectifying reasoning and providing measurable structures for improvement, we establish the first scientific basis for:

- Defining and testing AGI formally
- Quantifying superintelligent behavior
- Designing reproducible, self-optimizing reasoning architectures

Furthermore, the framework enables rapid adoption by human and machine agents, ensuring that intelligence is both measurable and auditable.

6 Conclusion

We have formalized intelligence as a property of structured reasoning spaces. RDUs, Meta-RDUs, compute-once objects, explainability trajectories, and RARFL cycles together constitute a substrate from which AGI emerges naturally, and superintelligence can be operationally defined and measured. This approach lays the foundation for a reproducible, auditable, and experimentally testable science of reasoning.