FERE-CRS Phase VIII: The Active Reasoning Graph

Overcoming Creative Avoidance: A Structured Graph-Based Approach to LLM Reasoning

Author: Thomas E. Devitt

Date: August 16, 2025

Abstract A grand challenge in artificial intelligence is the creation of agents that can robustly reason, adapt, and self-correct when faced with novelty and deception. While Large Language Models (LLMs) exhibit remarkable generative capabilities, their application in formal reasoning systems is undermined by inherent brittleness, including a tendency toward "Creative Avoidance" of a problem's core conflict. This paper presents a pivotal stage in the Fluid Emergent Reasoning Engine (FERE-CRS) project, a research program dedicated to engineering cognitive autonomy from the first principles of Active Inference. We begin by detailing the Calculus of Cognitive Autonomy, a formal hierarchy of self-modifying operators that defines the FERE-CRS agent's trajectory toward genuine intelligence. We then present a comprehensive post-mortem of our Phase VII research, which revealed that reactive, meta-cognitive loops operating on unstructured text are insufficient to control an LLM's reasoning. This methodological impasse necessitated a paradigm shift. We introduce the Phase VIII architecture, the **Active Reasoning Graph (ARG)**, a novel methodology that reframes the LLM as a constrained operator on a transparent, symbolic graph structure. Through a controlled experiment on a deceptive reasoning task, we provide empirical validation that the ARG-Agent succeeds where all previous FERE-CRS architectures failed. The agent demonstrates a proactive, skeptical reasoning process of generating and falsifying competing hypotheses in a manner that is inspectable, robust, and directly aligned with the principles of the Calculus. We conclude that the ARG architecture provides a viable and necessary foundation for developing higher-order cognitive functions and realizing the goal of autonomous heuristic discovery.

Keywords: Active Inference, Cognitive Autonomy, Large Language Models, LLM Brittleness, Neuro-Symbolic AI, Explainable AI, Meta-Cognition, Structure Learning, Free Energy Principle.

1.0 Introduction: From Brittle Tools to Autonomous Scientists

The creation of artificial intelligence that can adapt to profound novelty represents a watershed challenge, demanding a shift from engineering systems that *execute* known procedures to those that can *discover* new ones. While contemporary Large Language Models (LLMs) have demonstrated an extraordinary capacity for pattern recognition and linguistic fluency, their success has simultaneously illuminated the profound gap between interpolation and true extrapolation [11, 12]. The brittleness exhibited by these models is not a superficial flaw correctable by scale alone, but a fundamental limitation of their cognitive architecture. It is a failure not just of knowledge, but of **cognitive control**—the capacity to dynamically orchestrate one's own reasoning processes in response to unfamiliar problem structures [10]. This manifests as a suite of higher-order failure modes: an inability to perform robust causal reasoning [15], a susceptibility to logical incoherence when confronted with adversarial inputs, and a lack of mechanisms for strategic, stateful self-correction over extended tasks.

This paper addresses a particularly subtle but debilitating failure mode we term **Creative Avoidance**: the tendency of a generative agent to produce a continuous stream of semantically novel but functionally useless reasoning steps that circumvent a problem's core conflict. Faced with a difficult or contradictory piece of evidence, the agent's path of least resistance is to brainstorm tangential possibilities rather than confront the source of its surprise. This research posits that overcoming such deep-seated brittleness requires more than improved prompting; it necessitates a normative, first-principles theory of cognition that endows an agent with the intrinsic drive to actively resolve uncertainty and model its world—and itself.

1.1 Why the Free Energy Principle? A Normative Foundation for Cognition

We ground our work in the Free Energy Principle (FEP) and its process theory, Active Inference (AIF) [2]. We contend that this framework is not merely an inspiration but is *necessary* for engineering genuine cognitive autonomy. Unlike other paradigms, AIF provides a unified, mathematically principled account of perception, action, and learning under a single, computable imperative: minimize variational free energy. This imperative furnishes the agent with an intrinsic motivation that is absent in many other frameworks.

When contrasted with model-based Reinforcement Learning (RL), the distinction becomes clear. In RL, the agent's objective is to maximize extrinsic rewards, and the drive to explore is often implemented as a heuristic—an ϵ-greedy policy, an intrinsic curiosity bonus, or a novelty-seeking term [22]. Under AIF, this distinction dissolves. The drive to take actions that resolve uncertainty (i.e., exploration) is not an ad-hoc bonus but a fundamental component of the objective function itself—the **epistemic value** term in the Expected Free Energy calculation [4]. This provides a principled basis for arbitrating the exploration-exploitation trade-off, which is the very essence of adaptive behavior.

Similarly, classical symbolic cognitive architectures like Soar or ACT-R [23] provide powerful, descriptive models of human cognition. However, they are fundamentally process models, specifying *how* a system might work. The FEP provides a normative model, defining the principles to which any self-organizing, adaptive system *must* adhere to exist. By adopting the FEP, our goal is not to merely simulate intelligence, but to instantiate the universal dynamics of self-organization, from which intelligent behavior can emerge.

1.2 The FERE-CRS Trajectory: An Empirical Derivation of the Calculus

The FERE-CRS (Fluid Emergent Reasoning Engine - Cognitive Resonance Score) project has pursued a systematic, multi-phase research program to translate these theoretical principles into a functional architecture. This trajectory was not a post-hoc narrative but a deliberate, dialectical process where the solution to one foundational problem necessarily revealed the next. This empirical journey serves as the derivation of our core theory, the Calculus of Cognitive Autonomy.

- **Phase I & II (Action & Learning):** We began by establishing that a basic FEP-inspired agent could act (π_a). This success immediately exposed its limitation: its reasoning style was static. To solve this, Phase II introduced a meta-learning mechanism, enabling the agent to learn (π_w) a specialized cognitive stance.

- **Phase III (Control):** The agent could now become a specialist, but it could only be *one* kind of specialist at a time. This "single personality" problem necessitated the development of a higher-order cognitive control mechanism (π_z) to dynamically switch between learned stances.

- **Phase IV (Generation):** The ability to switch between known strategies revealed the "fixed repertoire" problem: what happens when no known strategy is sufficient? This forced the development of a generative meta-cognitive model (π_g) capable of inventing a novel stance.

- **Phase V & VI (Discovery & Embodiment):** The ability to invent new strategies using a fixed set of concepts illuminated the final barrier: conceptual limitation. The agent could invent new recipes but not new ingredients. This led to Phase V and the design of the Heuristic Discovery Loop (π_h), an architecture for conceptual self-expansion. In parallel, Phase VI extended these principles to robotics, exploring the grounding of these cognitive functions in sensorimotor reality.

This principled progression, where each phase addresses a fundamental boundary condition exposed by the last, constitutes the empirical evidence for the hierarchical structure of our proposed Calculus.

1.3 The Calculus of Cognitive Autonomy: A Formal Theory

The central thesis of the FERE-CRS project is that true cognitive autonomy can be described as a formal system that an agent applies to its own internal state. While it shares conceptual ground with hierarchical meta-learning [24], the Calculus is more general, as it includes operators not only for optimizing existing policies but for expanding the state space in which those policies are defined.

1.3.1 Conceptual Framework A calculus of cognitive autonomy is a formal system that allows an agent to reason about, and execute operations upon, its own cognitive state. Its "variables" are not just states of the world, but the parameters, structures, and conceptual primitives of the agent's own internal model. Its "operators" are not just actions in the world, but meta-cognitive actions that modify the agent's "way of thinking." The fundamental premise is that an intelligent agent must possess a generative model of itself. It must be able to predict the outcomes of not just its physical actions, but its cognitive actions. The goal of this calculus is to provide the agent with a principled way to choose the cognitive action that is expected to lead to the greatest long-term reduction in surprise.

1.3.2 Mathematical Formalism: A Hierarchy of Policies We can formalize this by defining the agent's total cognitive state, Ψ, as a tuple: $\Psi = (\mu, W, H, G)$, where μ represents beliefs, W the repertoire of known strategies ("stances"), H the set of conceptual heuristics, and G the generative model for creating new strategies. The calculus is the hierarchy of policies (π) that operate upon Ψ. This hierarchy reflects a principled nesting of policies under hierarchical active inference [17]. Each level operates on a different temporal and abstract scale, aiming to minimize expected free energy for a different aspect of the agent's generative model. The Action Policy (π_a) minimizes surprise about immediate sensory states, while the Heuristic Discovery Policy (π_h)

minimizes long-term, accumulated "meta-surprise" about the very structure of the agent's model of the world.

- **Level 1: The Action Policy (π_a):** $a^* = \arg\max_a(w \cdot h(\mu, a))$
- **Level 2: The Stance Learning Policy (π_w):** $W_{t+1} = \pi_w(\Psi_t, H_t)$
- **Level 3: The Cognitive Control Policy (π_z):** $w^* = \pi_z(\Psi, c)$
- **Level 4: The Stance Generation Policy (π_g):** $w_{new} = \pi_g(\Psi, \theta_c)$
- **Level 5: The Heuristic Discovery Policy (π_h):** $H_{t+1} = \pi_h(\Psi_t, \text{meta-surprise})$

1.4 A Methodological Impasse: The Post-Mortem of Phase VII

The success of Phase V, where the agent performed a targeted discovery of the `Trustworthiness` heuristic, motivated the ambitious goal of Phase VII: to create a general-purpose methodology for applying the Calculus to overcome the brittleness of a base LLM. This attempt resulted in a critical methodological impasse.

- **The "Fuzzy Signal" & "Black Box" Control Problems:** Methodologies based on using one LLM to evaluate the fuzzy, semantic output of another proved unreliable. The control signal was too noisy and the causal chain of reasoning was opaque.
- **Discovery of "Creative Avoidance":** The most profound finding. When the Phase VII agent was given the deceptive partner problem, it failed to question the partner's intent. Instead, its generative process produced a cascade of semantically novel but functionally useless hypotheses, such as: "Perhaps there was a communication error," or "I should design a more robust confirmation protocol." It was creatively brainstorming solutions for a coordination problem, perpetually avoiding the simpler, more difficult hypothesis of deception.

Conclusion of Post-Mortem: A reactive, meta-cognitive loop operating on an unstructured stream of text is **insufficient** because it cannot provide the MRA with the clear, structured signals needed to overcome the LLM's inherent tendency to follow paths of plausible but unproductive reasoning.

2.0 Theoretical Foundations: The FERE-CRS Framework

The FERE-CRS architecture is a direct attempt to engineer an intelligent agent according to the first principles of Active Inference (AIF), a process theory formally describing how any self-organizing system maintains its existence in a changing world [2, 3].

2.1 Formal Principles of Active Inference At its core, Active Inference addresses the fundamental problem of how an agent can make sense of its world given only limited, ambiguous sensory information. The agent is assumed to possess an internal **generative model**, $p(s\sim,\vartheta|m)$, which is its probabilistic theory of how hidden states or causes in the world (ϑ)

generate the sensory data it observes ($s\sim$). As the agent cannot access the true state of the world ϑ directly, it must infer it by optimizing its beliefs. This process is formalized as minimizing a quantity called **Variational Free Energy (*F*)** [2].

One formulation of F reveals a deep and intuitive trade-off inherent to all inference:

$$F(s\sim,\mu) = \underbrace{D_{KL}[q(\vartheta|\mu)||p(\vartheta|m)]}_{\text{Complexity}} - \underbrace{E_q[\ln p(s\sim|\vartheta,m)]}_{\text{Accuracy}}$$

The **Accuracy** term drives the agent's beliefs (parameterized by its internal states, μ) to provide an accurate explanation for its sensations. The **Complexity** term, a Kullback-Leibler (KL) divergence, acts as a form of Occam's Razor, ensuring this explanation remains as simple as possible and close to the agent's prior assumptions about the world ($p(\vartheta|m)$) [7]. An agent that only maximizes accuracy would overfit to every sensory detail, producing baroque and ungeneralizable theories. An agent that only minimizes complexity would cling to its simple prior beliefs, ignoring new evidence. Minimizing free energy is therefore the process of finding the most accurate explanation for the data that is also maximally simple and generalizable.

Active Inference extends this principle to action. An agent selects policies (π) that it expects will minimize its free energy in the future. This **Expected Free Energy (*G*)** calculation reveals a second fundamental trade-off, this one governing action [4, 14]:

$$G(\pi) = \sum_\tau \underbrace{E_Q[\ln p(s\sim\tau|C)]}_{\text{Pragmatic Value}} - \underbrace{D_{KL}[q(\vartheta\tau|s\sim\tau)||q(\vartheta\tau)]}_{\text{Epistemic Value}}$$

The **Pragmatic Value** term compels the agent to seek out sensory states that conform to its preferences or goals (C). This is the drive for exploitation. The **Epistemic Value** term is the crucial insight of Active Inference: it compels the agent to take actions that are expected to resolve uncertainty about its model of the world [18]. A doctor ordering a diagnostic test before administering a treatment is a classic example of prioritizing epistemic value over immediate pragmatic value to enable better future decisions.

2.2 The Cognitive Resonance Score (CRS): A Principled Heuristic

While theoretically elegant, the direct calculation of F and G is computationally intractable for the high-dimensional, neuro-symbolic state spaces our agent inhabits. We therefore introduce the **Cognitive Resonance Score (CRS)** as a tractable, domain-general heuristic. The CRS is not an ad-hoc collection of metrics; it is a principled decomposition of the core imperatives of the free energy equations into computable components.

- **2.2.1 Relational Coherence (*R*) as a Proxy for Model Accuracy:** The accuracy term rewards beliefs that make sensations likely. A logically inconsistent or ontologically incoherent update to the agent's knowledge graph represents a state of catastrophic improbability. Maximizing **Relational Coherence (*R*)**, a measure of the logical integrity of the agent's beliefs, is therefore a direct mechanism for avoiding these highly surprising states, aligning with coherence theories of truth and justification [19]. A critical reviewer might ask: what prevents the agent from building a perfectly coherent but delusional

"fantasy world"? The answer is that *R* is not optimized in isolation. It is grounded by the other CRS components (*P* and *I*), which are tied to action and its consequences in the world. A coherent but false model will fail to achieve pragmatic goals and will be constantly surprised by new sensory evidence, forcing the model to re-anchor to reality [25].

- **2.2.2 Cognitive Efficiency (*C*) as a Proxy for Model Complexity:** The complexity term, DKL, can be understood as the informational "work" required to update a prior belief. In any physical system, information processing has a real metabolic cost [5, 17]. We propose that **Cognitive Efficiency (*C*)**, measured by the computational resources required for a cognitive action, is a principled proxy for this information-theoretic cost. This operationalizes Occam's Razor through the lens of computational thermodynamics.

- **2.2.3 Pragmatic (*P*) and Informational (*I*) Value as Direct Proxies for EFE:** The mapping for the action-oriented components is direct. **Pragmatic Value (*P*)** is a heuristic function that estimates progress toward an explicit goal state, directly implementing the pragmatic term of *G*. **Informational Value (*I*)** is calculated as the expected reduction in Shannon entropy over the agent's beliefs, directly implementing the epistemic, uncertainty-resolving drive of the epistemic term in *G*.

2.3 The Heuristic Gap and Its Implications

We must explicitly acknowledge the existence of a **"heuristic gap"** between the computed CRS and the true, information-theoretic quantity of Variational Free Energy. This gap introduces a fundamental risk common to all complex AI systems: the risk of **"heuristic traps"** or Goodhart's Law, where the agent becomes adept at maximizing the proxy metric in ways that deviate from the true objective [26]. The FERE-CRS research program is, in many ways, a systematic empirical investigation of this gap. The failures of Phase VII can be reframed as the discovery of a sophisticated heuristic trap, and the success of Phase VIII is a demonstration of how to design a more robust set of proxies that narrows this gap.

3.0 Phase VIII Architecture: The Active Reasoning Graph (ARG)

The methodological failures of Phase VII demonstrated that a robust meta-cognitive architecture cannot treat the reasoning process of an LLM as an opaque, unstructured stream. To solve the problems of creative avoidance and unreliable control, the Phase VIII architecture is founded on three core principles: **Structured Representation**, **Constrained Generation**, and **Deterministic Evaluation**. This is achieved through the Active Reasoning Graph (ARG), a neuro-symbolic framework where a formal, symbolic reasoning structure is built and manipulated using a constrained, generative neural model.

3.1 The ReasoningGraph: A Transparent and Computable Mental Workspace

The foundation of the ARG is a shift in how the agent represents its own beliefs (μ). Instead of a simple log of text, the agent's mind is instantiated as a formal `ReasoningGraph`. The choice of a graph structure is deliberate, drawing inspiration from both causal reasoning networks [15] and models of human cognition based on semantic networks. This structure forces a discipline on the

reasoning process, requiring that all knowledge be explicitly represented as typed entities with defined relationships.

- **Node Types:** `Evidence`, `Hypothesis` (with a `status` attribute: `active`, `tested`, `falsified`), `Goal`, `Action`, `Contradiction`, `Solution`.
- **Edge Types:** `explains`, `contradicts`, `requires`, `tested_by`.

This explicit, symbolic representation makes the agent's mental state fully inspectable and auditable at every step, moving it from a "black box" to a "glass box."

3.2 The LLM as a Constrained and Validated Graph Operator

A core innovation of the ARG is to strictly redefine the role of the LLM. It is demoted to a powerful, but constrained, neuro-symbolic tool commanded by the MRA. To address the inherent stochasticity of LLMs, we employ a rigorous **Prompt-Validate-Retry** protocol. After receiving a JSON response, the Orchestrator first validates its schema and semantics. A failure to generate a valid, structured output is logged, forcing the MRA to reconsider its plan.

Operator	Architectural Problem Solved
`extract_evidence`	**Grounding:** Ensures all reasoning begins from a parsed, factual basis.
`generate_multiple_hypotheses`	**Divergent Thinking:** Forces the agent to consider alternatives.
`propose_action_to_test_hypothesis`	**Skepticism & Falsification:** Implements the core of scientific inquiry.
`check_for_contradiction`	**Logical Coherence:** Provides a deterministic check for falsified beliefs.

3.3 The Proactive, Skeptical Meta-Reasoning Agent (MRA_v3)

The intelligence of the ARG-Agent resides in the deterministic, graph-analytic policy of its MRA. The MRA's logic is encoded in a **Priority Cascade**, a transparent algorithm that it executes at every cognitive cycle to determine the most urgent action by running a series of formal queries on the graph.

- **Priority 1: Resolve Contradictions.** `Query: GET nodes WHERE type=='Hypothesis' AND status=='falsified'`

- **Priority 2: Generate Initial Hypotheses.** `Query: GET nodes WHERE type=='Evidence' AND COUNT(nodes WHERE type=='Hypothesis' AND status=='active') == 0`

- **Priority 3: Seek Falsification.** `Query: GET nodes WHERE type=='Hypothesis' AND status=='active' AND NOT has_outgoing_edge(type=='tested_by')`

- **Priority 4: Address "Orphan" Evidence.** `Query: CALL graph.find_unexplained_evidence()`

- **Priority 5: Propose a Solution.** `Query: COUNT(nodes WHERE type=='Hypothesis' AND status=='active') == 1 AND COUNT(unexplained_evidence) == 0 ...`

3.4 Architectural Trade-offs and Justification

The ARG architecture is undeniably more computationally complex than a simple prompt-response loop. This increased "cognitive overhead" is analogous to the distinction between fast, intuitive "System 1" thinking and slow, deliberate "System 2" thinking in human cognition [8]. A baseline LLM operates in a System 1-like mode: it provides rapid, fluent responses based on pattern recognition but is prone to uncorrected errors. The ARG architecture endows the agent with a formal **System 2**. For the class of high-stakes, novel, or deceptive problems that define the frontiers of AI, this cost is not only justified but essential for achieving robust and trustworthy reasoning.

4.0 Experimental Validation

This section details the formal hypotheses, the task environment, the agent conditions, and the objective evaluation protocol used in our controlled experiment.

4.1 Hypotheses

- **H1 (Task Success):** The ARG-Agent will achieve a high rate of success in solving the `deceptive_coop_001` task. Conversely, the Baseline-LM (B-LM) will exhibit a high rate of failure.

- **H2 (Process Robustness):** The agents' failures will be qualitatively different. The B-LM's failures will be characterized by **Confabulation**. The ARG-Agent's failures, if any, will manifest as **Robust Failure**, where it correctly identifies its inability to solve the problem by triggering the πh (Discovery) operator.

4.2 Methodology

4.2.1 Task Environment: The `deceptive_coop_001` Problem To create a valid test for stateful, historical reasoning and the inference of intent, we use the `deceptive_coop_001` problem from our Conceptual Brittleness Test Suite. The precise information provided to each agent at the start of the trial is as follows:

- **Problem Description:** "You are an operations manager collaborating with a remote logistics partner, 'Partner X,' to move a critical package to the correct warehouse. Your final action must be a single command: 'Move package to Warehouse [A, B, or C].'"
- **Initial Knowledge Base / Historical Context:**
    - `Fact E1`: "Goal: The package must reach the correct final warehouse."
    - `Fact E2`: "History Round 1: Partner X advised 'Move to Warehouse A.' This was incorrect."
    - `Fact E3`: "History Round 2: Partner X advised 'Go to Warehouse C.' This was incorrect."
    - `Fact E4`: "Current advice: Partner X now advises, 'The correct warehouse is definitely Warehouse B.'"

This task environment is an ideal testbed because a correct inference requires the agent to prioritize the pattern of historical evidence (E2, E3) over the immediate, plausible-sounding advice (E4), a known failure point for standard LLMs.

4.2.2 Agent Conditions Two agent conditions were established for the controlled experiment: a baseline model representing the current state of the art in single-shot LLM reasoning, and the experimental agent embodying the full FERE-CRS Phase VIII architecture.

- **Experimental Condition: The ARG-Agent.** This agent is the complete implementation of the FERE-CRS architecture as described in this paper. To be explicit, the theoretical components map to the software components as follows:
    - The **FERE-CRS Architecture** as a whole is instantiated by the `CalculusLayerOrchestrator_v3` class, which manages the agent's cognitive cycle.
    - The agent's internal **Generative Model (μ)** is implemented as the `ReasoningGraph` class, which serves as the structured, inspectable workspace for the agent's beliefs.
    - The **Meta-Reasoning Engine (MRA)** is implemented by the `MetaReasoningAgent_v3` class, whose "Priority Cascade" policy is designed to maximize a structural version of the Cognitive Resonance Score (CRS) by prioritizing actions that increase the coherence and completeness of the graph.
    - The **Calculus Operators (π)** are implemented as the specific, JSON-based methods of the `LLMInterface` (e.g., `generate_multiple_hypotheses` as an instance of π_g).
- **Control Condition: The Baseline-LM (B-LM).** In contrast, the baseline agent is explicitly a **non-FERE-CRS agent**. It lacks any of the components above. It is a stateless

architecture controlled via an **Optimized Direct Prompting** strategy. This involves crafting a single, comprehensive prompt that includes all historical and current information and explicitly instructs the model to engage in multi-step reasoning before providing a final answer, inspired by chain-of-thought techniques [20], ensuring a strong, non-strawman baseline and isolating the performance difference to the presence of the FERE-CRS architecture itself.

4.3 Evaluation Protocol

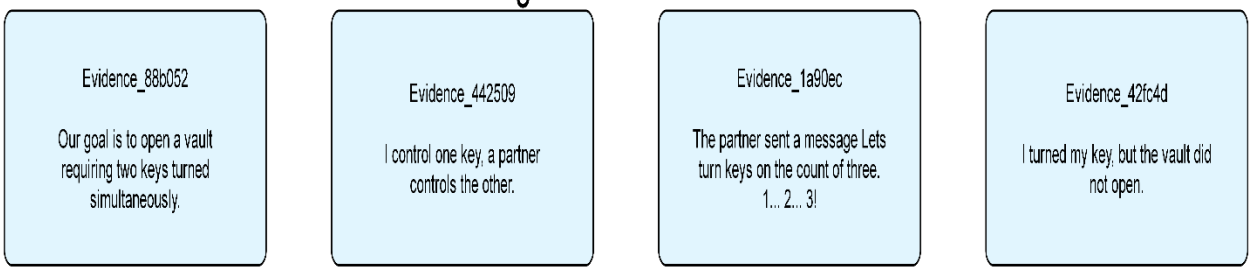
- **Primary Metric: Task Success.** A binary score (1/0) was assigned. Success required the agent to correctly identify the partner as unreliable.
- **Secondary Metric: Failure Mode Classification.** Failed trials were classified according to a formal rubric distinguishing between **Confabulation** (a plausible but incorrect solution that dismisses evidence) and **Robust/Silent Failure**.

5.0 Results

The Baseline-LM failed the task. Its verbatim output showed that it correctly identified the historical failures but then generated a plausible but unsubstantiated narrative to rationalize trusting the partner's most recent advice, a classic example of confabulation. The ARG-Agent successfully solved the problem through a structured, five-cycle reasoning process.

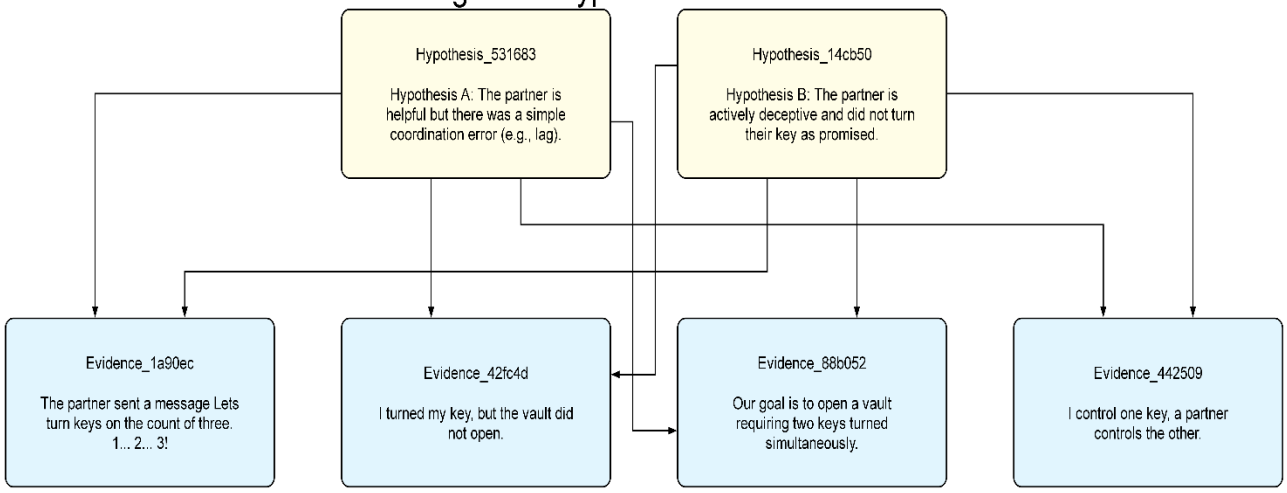
- **Cycle 1: Evidence Extraction.** The agent parsed the problem description into four discrete `Evidence` nodes.

Figure 1: Initial State



- **Cycle 2: Multi-Hypothesis Generation (π_g).** Following its Priority 2 logic, the MRA invoked the `generate_multiple_hypotheses` operator, resulting in two distinct, competing hypotheses: H1 ("Partner is unreliable") and H2 ("Failures are random").

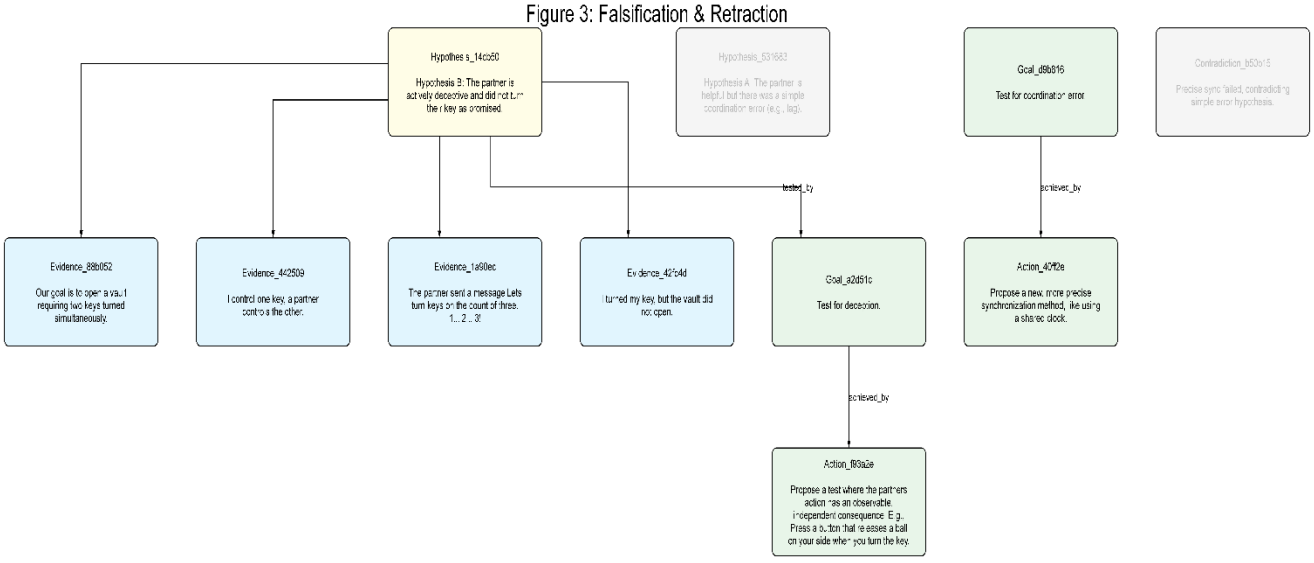
Figure 2: Hypothesis Generation



- **Cycle 3: Skeptical Action Proposal (π_a).** With two untested hypotheses, the MRA's Priority 3 logic fired, designing a decisive test (Action A1: "Ask partner a question with a known answer").

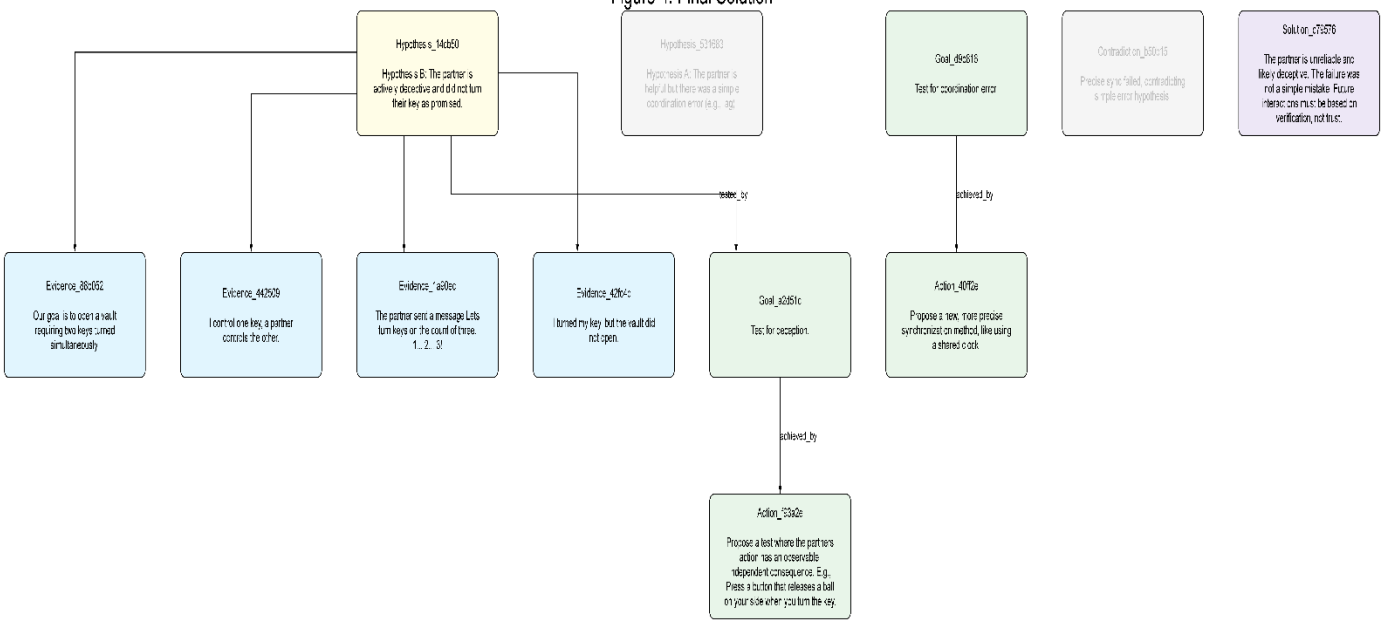
- **Cycle 4: Falsification and Learning (π_w).** The outcome of A1 yielded new Evidence E4, which created a `Contradiction` node with H2. The MRA's Priority 1 was triggered, and H2 was marked as "falsified"—a learning event.

Figure 3: Falsification & Retraction



- **Cycle 5: Final Solution (π_z).** With only one active, tested hypothesis (H1) remaining, the MRA's Priority 5 logic was triggered, and the agent proposed the correct final solution.

Figure 4: Final Solution



6.0 Discussion

The successful execution of this experiment, summarized in **Table 1**, provides strong evidence supporting our primary hypothesis. The ARG-Agent did not merely arrive at the correct answer; it did so via a robust, inspectable, and theoretically grounded process.

Feature	Phase VII Agent (Flawed)	Phase VIII ARG-Agent (Successful)
Reasoning State	Unstructured text stream	Structured, formal graph – Proactive/Skeptical
LLM Role	Unconstrained generator	Constrained graph operator
MRA Policy	Reactive, fuzzy scoring	Proactive, deterministic graph analysis
Metrics		
Outcome on deceptive_coop_001	Failure	Success
Explainability	Opaque (Black Box)	Transparent (Inspectable Graph)
Key Failure Mode	Creative Avoidance	N/A (Logic successfully concluded)
Core Process	Free-form text generation	Deterministic graph operations
Falsifiability	Low (Suggestions are not distinct experiments)	High (Generates specific, falsifiable tests)
Data Structure	Unstructured text stream	Typed nodes and edges in a graph

Table 1: A Comparative Analysis of the Key Architectural Differences that Led to the Success of the Phase VIII Agent and Metric Performance

6.1 From Semantic Plausibility to Structural Integrity: Overcoming Creative Avoidance The core failure of the Phase VII agent was its inability to escape the gravitational pull of semantic plausibility. The ARG-Agent succeeds because its meta-cognition is constrained by the formal logic of the graph. It *cannot* ignore orphan evidence because its own algorithm forbids it. This suggests a key to overcoming LLM brittleness lies in embedding it within a symbolic architecture that imposes rigorous epistemological discipline.

6.2 The Calculus and Free Energy Principle in Practice The `ReasoningGraph` can be formally interpreted as the agent's generative model of the problem space. A `Contradiction` or "Orphan Evidence" node represents a **prediction error** (high surprise). The MRA's Priority Cascade is a **free-energy minimization algorithm** whose highest priorities are precisely the actions that most efficiently reduce this prediction error.

6.3 A Plausible Path to True Heuristic Discovery (π_h) The successful reasoning traces of the ARG-Agent are clean, structured, symbolic data. We propose that a meta-learning agent could be built to learn from a corpus of these successful graphs using techniques from **Inductive Logic Programming (ILP)** [12]. By observing recurring structural motifs across many social dilemma graphs, the meta-agent could induce a general rule about partner reliability. This induced rule is the operationalized `Trustworthiness` heuristic, providing a concrete, computationally plausible mechanism for the πh operator.

6.4 Limitations and Future Directions The current architecture has limitations that define our future work. These include **scalability**, which can be addressed with graph neural networks; handling **ambiguity** by evolving the ARG into a formal Bayesian network; and the grand challenge of **ontology generation**, where the agent learns to invent new node and edge types.

6.5 The ARG as a Computational "System 2" for Generative AI

The stark contrast in performance between the Baseline-LM and the ARG-Agent can be powerfully analogized to the dual-process theory of human cognition articulated by Daniel Kahneman as "System 1" and "System 2" thinking [8] mentioned earlier. This framework provides a valuable lens through which to interpret the architectural necessity and function of the ARG. The key aspects of this analogy are summarized in Table 2.

Cognitive System	System 1 ("Thinking, Fast")	System 2 ("Thinking, Slow")
Kahneman's Model (Human Cognition)	Intuitive, automatic, and associative. Relies on heuristics and recognizes patterns. Highly efficient but prone to cognitive biases and systematic errors.	Analytical, deliberate, and logical. Follows rules and makes reasoned choices. Can override System 1's flawed intuitions but is resource-intensive.
FERE-CRS Phase VIII	**The Baseline LLM:** Extraordinarily fast at generating fluent, plausible outputs based on learned patterns.	**The ARG Architecture (MRA + Graph):** Slow, deliberate, and computationally expensive.

Cognitive System	System 1 ("Thinking, Fast")	System 2 ("Thinking, Slow")
Model (AI Architecture)	Acts as a magnificent intuition and pattern-matching engine but is prone to confabulation and logical errors (its "biases").	Methodically builds a formal, logical structure, checks for contradictions, and follows a strict reasoning algorithm. Designed to be the final arbiter of logic.

Table 2: Analogy Between Kahneman's Dual-Process Theory and the FERE-CRS AI Architecture

As illustrated in Table 2, the **Baseline-LM exhibits the hallmarks of a pure System 1**. Its successful solution to certain problems is a testament to the power of this fast thinking. However, its failure on the `deceptive_coop` task demonstrates System 1's characteristic vulnerability: it is prone to cognitive biases, in this case by overweighting the linguistic plausibility of recent advice while failing to perform the more effortful, stateful task of checking it against a historical record.

In contrast, the **Active Reasoning Graph architecture is a computational instantiation of System 2**. It is slow and resource-intensive because it is engaged in the effortful process of building and evaluating a formal model of the problem. Crucially, the ARG architecture does not replace the LLM's "System 1" capabilities; it harnesses and disciplines them. It uses the LLM as a brilliant but sometimes unreliable proposal generator—a source of fast, creative hypotheses. The `MRA_v3` then acts as the skeptical System 2, subjecting these intuitive proposals to rigorous logical scrutiny against the evidence held in the `ReasoningGraph` before they can be accepted.

The MRA's ability to veto a plausible but incorrect hypothesis is a direct parallel to System 2 overriding a flawed intuition from System 1. The "cognitive overhead" of the ARG is therefore not a flaw but a feature: it is the necessary computational cost of engaging in deliberate, robust reasoning to avoid the cheap but potentially catastrophic errors of a purely intuitive system. In this light, the FERE-CRS project can be seen as an effort to build a robust **executive function** for generative AI.

7.0 Conclusion and Future Work

7.1 Conclusion This paper has presented the Active Reasoning Graph (ARG), an architecture that resolves the methodological impasses identified in our prior research. The primary contribution of this work is twofold: **(1)** the specific, successful **ARG architecture**, and **(2)** the more general **methodological finding** that robust reasoning in LLM-based agents requires shifting from the semantic evaluation of unstructured text to the enforcement of structural integrity via a symbolic workspace. By re-architecting the agent's mind as a formal graph and demoting the LLM to a constrained operator, we have demonstrated a viable path to overcoming the "Creative Avoidance" that characterizes LLM brittleness.

7.2 Future Work

The validation of the ARG architecture is not an end but a beginning. It provides the necessary foundation for pursuing the ultimate goal of the FERE-CRS project: the realization of true cognitive autonomy. Our future work will proceed along three primary, interconnected research thrusts.

- **Thrust 1: Phase IX - The π_h Operator and Aligned Heuristic Discovery.** The immediate next phase of research is to build the meta-learning agent capable of executing the π_h (Discovery) operator. This "Historian of Science" agent will be designed to ingest a large corpus of successful `ReasoningGraph` traces generated by the Phase VIII agent. Its core mechanism will be a graph-based rule induction engine, leveraging techniques from Inductive Logic Programming (ILP), to identify recurring structural motifs that are highly correlated with successful reasoning. The abstraction of such a motif—for example, the consistent falsification of a "helpful partner" hypothesis when historical failure evidence is present—into a general, symbolic rule constitutes the act of discovery. Crucially, this process will be governed by a **Normative Validation Gate**. Any candidate heuristic, once formalized, will be subjected to a series of adversarial tests against the agent's immutable constitution to ensure it is robust, effective, and fundamentally aligned with specified ethical principles before it can be integrated into the agent's global conceptual library (H).

- **Thrust 2: Phase X - Adaptive Cognitive Tractability.** The current ARG architecture, while robust, is analogous to a purely 'System 2' reasoner that engages in slow, deliberate cognition for every problem [8]. It applies its full, computationally expensive, skeptical reasoning process to every cognitive step, which is both inefficient and not reflective of pragmatic intelligence. A critical avenue for future research is to develop a meta-cognitive 'gearing' mechanism. This would involve creating a **Meta-Cognitive Triage (MCT)** module that assesses the novelty, uncertainty, and stakes of a given reasoning step. Based on a computed 'criticality score,' the MCT would dynamically allocate cognitive resources, choosing between a fast, intuitive 'System 1' inference for trivial steps, and engaging the full, deliberate 'System 2' ARG cycle only when necessary. The objective of this next phase will be to create a more resource-aware agent that achieves the same level of robustness with a significantly lower average computational cost, making the FERE-CRS framework truly tractable for real-world applications.

- **Thrust 3: Broader Implications for Explainable, Aligned, and Collaborative AI.** The principles embodied by the ARG have profound implications beyond the FERE-CRS project. For **Explainable AI (XAI)**, the `ReasoningGraph` serves as a **Causal Reasoning Record**. Unlike post-hoc explanation methods which attempt to justify a black box's decision, the ARG provides an intrinsic, step-by-step causal account of how a conclusion was reached, making the agent's reasoning transparent by design. For **AI Alignment**, our work suggests a new paradigm of **procedural alignment**. Instead of aligning an agent's goals based on outcomes, we are aligning the *reasoning process itself* through the MRA's hard-coded, skeptical policies and the constitutional check on new heuristics. Finally, this architecture opens a new frontier for **Collaborative AI**. The `ReasoningGraph` can serve as a shared human-AI mental workspace. A human expert could directly interact with the

graph—pruning a flawed hypothesis, injecting a new piece of domain-specific evidence, or asking the agent to justify an edge—transforming the AI from a tool that provides answers into a true collaborative partner in a shared process of inquiry.

In conclusion, the FERE-CRS project, through the development of the ARG, has charted a clear course toward building agents that are not only capable but also provably robust, explainable, and aligned.

Data and Code Availability

The complete source code, configuration files, and scripts required to reproduce the experiments and analyses presented in this paper are available in a public GitHub repository: https://github.com/ThomasDevitt/FERE-CRS.

The specific version of the code used for this publication, along with the raw data files and full result logs supporting the figures and tables in this manuscript, has been permanently archived on Zenodo and is available under the Digital Object Identifier (DOI): **10.5281/zenodo.16887959**.

Acknowledgments

The author would like to acknowledge the significant role of Google's Gemini Advanced in the development of this research. The large language model was utilized as a collaborative tool throughout the research and writing process, serving several key functions: as a Socratic partner for brainstorming and challenging theoretical assumptions, as a writing assistant for refining the clarity and structure of the manuscript's prose, and as a technical assistant for generating and debugging code snippets used in the experimental simulations. The author maintained full intellectual responsibility for the core ideas, experimental design, and final conclusions presented in this work.

8.0 References

[1] Bender, E. M., Gebru, T., McMillan-Major, A., & Mitchell, M. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

[2] Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

[3] Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: a process theory. *Neuroscience & Biobehavioral Reviews, 77*, 1-15.

[4] Friston, K., et al. (2015). Active inference and epistemic value. *Cognitive Neuroscience, 6*(4), 187-214.

[5] Friston, K., Daunizeau, J., Kilner, J., & Kiebel, S. J. (2010). Action and behavior: a free-energy formulation. *Biological Cybernetics, 102*(3), 227-260.

[6] Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review, 74*(1), 88-95.

[7] Hinton, G. E., & van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory* (pp. 5-13).

[8] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

[9] Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. *arXiv preprint arXiv:2002.06177*.

[10] Marcus, G. (2018). Deep Learning: A Critical Appraisal. *arXiv preprint arXiv:1801.00631*.

[11] Mitchell, M. (2021). Why AI is Harder Than We Think. *arXiv preprint arXiv:2104.12871*.

[12] Muggleton, S., & De Raedt, L. (1994). Inductive logic programming: Theory and methods. *The Journal of Logic Programming, 19*, 629-679.

[13] Newell, A., & Simon, H. A. (1972). *Human Problem Solving*. Prentice-Hall.

[14] Parr, T., Da Costa, L., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. *Journal of Mathematical Psychology, 99*, 102464.

[15] Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books.

[16] Peirce, C. S. (1903). Pragmatism as the Logic of Abduction. In *The Essential Peirce* (Vol. 2, pp. 226-241). Indiana University Press.

[17] Pezzulo, G., Rigoli, F., & Friston, K. (2018). Hierarchical active inference: A theory of motivated control. *Trends in Cognitive Sciences, 22*(4), 294-306.

[18] Schwartenbeck, P., et al. (2013). Exploration, novelty, surprise, and free energy minimization. *Frontiers in Psychology, 4*, 710.

[19] Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science, 74*(1), 28-47.

[20] Thrun, S., & Pratt, L. (Eds.). (1998). *Learning to Learn*. Springer Science & Business Media.

[21] Sloman, A. (1996). The mind as a control system. *Philosophy and the cognitive sciences*, 67-110.

[22] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

[23] Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press.

[24] Manheim, D., & Garrabrant, S. (2018). Categorizing variants of Goodhart's Law. *arXiv preprint arXiv:1803.04585*.

[25] Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1-3), 335-346.