FERE-CRS: A Cognitive Architecture for Emergent Fluid Reasoning and Autonomous Heuristic Discovery

Thomas E. Devitt *Independent Researcher*

Date of Submission: August 8, 2025

Abstract

A grand challenge in artificial intelligence is the creation of agents that can adapt to entirely novel situations—a capability that requires moving beyond selecting from a known set of strategies to inventing new ones. This paper presents a complete, five-phase research program detailing the systematic design, implementation, and validation of the Fluid Emergent Reasoning Engine (FERE-CRS), a cognitive architecture grounded in the Free Energy Principle. We address the challenge of creating adaptive intelligence by presenting a series of five interconnected studies, each proposing an architectural innovation to overcome a specific, fundamental limitation. Study 1 establishes the core FERE-CRS framework and validates its ability to perform high-quality, efficient reasoning on complex inferential tasks. Study 2 demonstrates that the agent can learn specialized "cognitive stances" through heuristic meta-learning. Study 3 achieves dynamic cognitive control, enabling the agent to switch between learned stances to meet the evolving demands of a problem. Study 4 overcomes the "fixed repertoire" limitation by creating an agent with a generative capacity to invent novel cognitive stances for unseen problem classes. In the final and primary contribution of this work, Study 5, we demonstrate a form of cognitive autonomy, creating a "cognitive scientist" agent capable of discovering, operationalizing, and integrating a new heuristic by reasoning about its own systemic failures. This work provides a complete, transparent, and reproducible account of an agent's journey from a static reasoner to an autonomous learner, offering a principled and empirically validated pathway toward more general and adaptive artificial intelligence.

Keywords: Active Inference, Fluid Reasoning, Cognitive Architecture, Heuristic Discovery, Cognitive Autonomy, Generative Meta-Cognition, Neuro-Symbolic AI, Artificial General Intelligence, Structure Learning.

1. Introduction: From Brittle Specialists to Emergent Generalists

The pursuit of artificial general intelligence (AGI) is a search for the first principles of autonomous adaptation. While contemporary AI, particularly Large Language Models (LLMs), has achieved remarkable success in specialized domains, this success has also illuminated a fundamental architectural bottleneck [2, 22]. These systems exhibit a characteristic brittleness when confronted with true novelty, a failure not merely of knowledge but of cognitive process. Their inability to robustly generalize out-of-distribution stems from a core limitation: they are masters of complex pattern recognition and interpolation but struggle with the **fluid cognitive control** required for extrapolation. Specifically, they lack a principled, internal mechanism to dynamically shift their own reasoning strategy—from methodical deduction to open-ended exploration, for instance—when faced with a problem whose structure differs from the statistical regularities of their training data [21, 23]. This results in a failure to perform genuine **abductive**

reasoning—the inference to the best explanation for a novel set of observations—which is a cornerstone of human scientific discovery and everyday problem-solving [14, 26].

This paper argues that overcoming this limitation requires more than scale or improved training techniques; it requires a new architectural paradigm grounded in a normative theory of intelligent behavior. The Free Energy Principle (FEP), and its process theory Active Inference (AIF), provides such a foundation [10]. The FEP is not merely an inspiration for our work; we posit it is **necessary** because it provides a first-principles, unifying account of how an agent can manage the fundamental trade-off between **epistemic foraging** (acting to reduce uncertainty) and **pragmatic goal-seeking** (acting to achieve preferred outcomes) under a single, computable objective [12]. This intrinsic balancing of exploration and exploitation is precisely the mechanism that current AI systems lack, forcing them into either rigid, goal-directed modes or unstructured, generative ones. By grounding our architecture in AIF, we seek to build an agent whose adaptive behaviors are not a set of hand-coded situational responses, but an **emergent property** of its continuous, fundamental drive to build a coherent and predictive model of its world.

However, the mathematical formalism of AIF presents a formidable implementation challenge, creating a "heuristic gap" between its theoretical elegance and its application to complex, neuro-symbolic systems. We bridge this gap with the **Cognitive Resonance Score (CRS)**, a composite heuristic that serves as a tractable proxy for Variational Free Energy minimization. The CRS is not an ad-hoc collection of metrics; it is a principled decomposition of the core imperatives of AIF into computable components: **Relational Coherence (*R*)** and **Cognitive Efficiency (*C*)** as proxies for model accuracy and complexity, and **Pragmatic Value (*P*)** and **Informational Value (*I*)** as direct proxies for the core drivers of action selection under AIF. This formulation allows a central Meta-Reasoning Agent (MRA) to orchestrate heterogeneous cognitive components—from logical theorem provers to generative LLMs—using a single, common currency that is itself grounded in a unified theory of cognition.

This paper presents the complete, five-phase research program of the Fluid Emergent Reasoning Engine (FERE-CRS), a systematic effort to build and validate an agent of increasing cognitive autonomy. We present a single, cohesive narrative structured as a series of five interconnected studies, each designed to overcome a specific, foundational limitation in the pursuit of adaptive intelligence. This logical progression is central to our contribution, demonstrating a structured path from static reasoning to autonomous discovery:

1. **Study 1: Foundational Architecture & Validation:** We establish that the core architecture, guided by the CRS, can solve complex inferential tasks with superior quality and efficiency, demonstrating an emergent, AIF-consistent shift from exploration to exploitation.
2. **Study 2: Learning a Cognitive Stance:** We move from a static to a learning agent, demonstrating that the agent can adapt its own internal CRS motivations via heuristic meta-learning to specialize its cognitive style for a given environment.
3. **Study 3: Dynamic Cognitive Control:** We overcome the "single personality" limitation by creating a "cognitive mechanic"—an agent that manages a repertoire of learned

stances and can dynamically switch between them to meet the evolving demands of a complex problem.

4. **Study 4: Generative Meta-Cognition:** We transcend the "fixed repertoire" boundary by creating a "cognitive engineer"—an agent that can invent a novel cognitive stance in a zero-shot fashion when faced with a problem class it has never seen before.
5. **Study 5: Autonomous Heuristic Discovery:** In the final and primary contribution of this work, we demonstrate a form of **cognitive autonomy**. We present a "cognitive scientist" agent capable of reasoning about its own systemic failures to discover, operationalize, and integrate a new conceptual primitive (heuristic) into its own architecture.

This final study, in particular, showcases a form of **emergence** that is distinct from mere complexity. The agent's discovery of a new heuristic is not a pre-programmed outcome but a bottom-up, self-directed reconfiguration of its own architecture, triggered by an internally generated "meta-surprise" at its own predictive failings. By presenting this full, reproducible account of an agent's journey from a static reasoner to an autonomous learner, we offer a principled and empirically validated pathway toward more general, robust, and truly adaptive artificial intelligence.

2. Theoretical Foundations: From Active Inference to a Tractable Heuristic

The FERE-CRS architecture is a direct attempt to engineer a system that conforms to the principles of Active Inference (AIF), a process theory describing how any self-organizing system can maintain its existence by minimizing a quantity called Variational Free Energy (VFE) [10]. To understand the architecture's design, one must first understand the formal problem it is designed to solve and the principled approximation we have developed to make it tractable.

2.1 The Free Energy Principle and Active Inference

Under the FEP, an agent possesses an internal generative model, $p(s\sim,\vartheta|m)$, which is its probabilistic theory of how hidden causes or states in the world (ϑ) generate sensory data ($s\sim$). The agent cannot access the true state of the world directly, so it must infer it by optimizing an approximate posterior distribution over these hidden states, known as the recognition density, $q(\vartheta|\mu)$, which is parameterized by the agent's internal states (μ).

The agent's objective is to minimize VFE, which serves as an upper bound on surprise (or a lower bound on model evidence). A key formulation of VFE is:

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

Here, DKL is the Kullback-Leibler (KL) divergence. This equation reveals a fundamental trade-off: to minimize F, an agent must find a belief state (μ) that produces accurate explanations for its sensations (maximizing the **Accuracy** term), while simultaneously keeping those beliefs as simple as possible and close to its prior assumptions about the world (minimizing the **Complexity** term) [16]. This process of optimizing beliefs by minimizing VFE is **perception**.

Active Inference extends this to **action**. An agent can also minimize its long-term surprise by selecting policies (sequences of actions, π) that are expected to minimize the Expected Free Energy (EFE):

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

where Q is a distribution over future states and outcomes, and C represents the agent's goals or preferred outcomes. EFE formalizes the exploration-exploitation dilemma. The **Pragmatic Value** term drives the agent to seek out preferred outcomes (exploitation), while the **Epistemic Value** term drives the agent to take actions that are expected to resolve uncertainty about the world (exploration), thereby making better future decisions possible [12, 29].

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

The direct calculation of VFE and EFE is computationally intractable for complex, high-dimensional, neuro-symbolic systems. We therefore propose the Cognitive Resonance Score (CRS) as a practical objective function, where maximizing the CRS is functionally approximate to minimizing VFE: $CRS(\mu,s\sim) \approx -F(\mu,s\sim) + K$. The CRS is a composite score that decomposes the abstract imperatives of AIF into computable components, with each mapping justified below.

2.2.1 Relational Coherence (*R*) as a Proxy for Model Evidence and Accuracy

The VFE's accuracy term, $Eq[\ln p(s\sim|\vartheta,m)]$, rewards beliefs that make sensory data likely, which is equivalent to maximizing the evidence for the agent's generative model, m. We posit that **Relational Coherence (*R*)** is a necessary, though not sufficient, heuristic for this. In a system like FERE-CRS, the generative model is a highly structured set of beliefs, such as a knowledge graph with an ontological schema. A belief update that introduces a logical contradiction or violates an ontological rule creates a large internal prediction error, which is mathematically equivalent to the model assigning a very low probability to that state. Maximizing R is therefore the process of minimizing this internal prediction error, ensuring the agent's inferential machinery remains sound. This aligns with coherence theories of justification, which argue that the plausibility of a belief is a function of its fit within a larger, interconnected system of beliefs [33].

This raises the critical question of how the system avoids constructing a coherent but factually incorrect "fantasy world." The answer is that R is **grounded and constrained by the other CRS components, which are tied to external interaction.** A highly coherent but false model will fail to explain sensory input, leading to high VFE. Actions based on this model will fail to achieve goals, resulting in low Pragmatic Value (P) and thus a low global CRS. The agent's drive to maximize the *entire* CRS, not just R, forces its internal model to remain tethered to reality.

2.2.2 Cognitive Efficiency (C) as a Proxy for Complexity

The VFE's complexity term, $D_{KL}[q(\vartheta|\mu)||p(\vartheta|m)]$, acts as a form of Occam's Razor, penalizing beliefs that are overly complex or deviate far from the agent's prior assumptions. We propose

that **Cognitive Efficiency (*C*)** is a practical proxy for this information-theoretic cost. The KL divergence can be understood as the "work required" to update the prior belief $p(\vartheta|m)$ to the new posterior belief $q(\vartheta|\mu)$. In any physical or computational system, information processing has a real, tangible cost (e.g., energy, CPU cycles, time). Simple beliefs that are close to the prior are computationally "cheaper" to represent and update. Complex, surprising beliefs that require extensive updates to the system's internal state (μ) map directly to higher computational cost. This view is consistent with thermodynamic and metabolic interpretations of the FEP, where minimizing VFE is equated with minimizing the long-term metabolic cost of maintaining a complex system [11, 27]. Thus, by measuring the computational cost of an action, C_{cost}, we create a tractable proxy for the complexity it induces.

2.2.3 Pragmatic (*P*) and Informational (*I*) Value as Proxies for EFE

The mapping of the *P* and *I* components of the CRS to the terms of the EFE is more direct.

- **Pragmatic Value (*P*):** This is a direct heuristic for the EFE's pragmatic term, measuring the expected progress toward a goal state. It quantifies the extrinsic, goal-achieving value of an action.
- **Informational Value (*I*):** This is a direct heuristic for the EFE's epistemic term. It quantifies the expected reduction in Shannon entropy over the agent's beliefs about the world. It represents the intrinsic value of "epistemic foraging"—the drive to take actions that resolve ambiguity to enable better future decisions [29].

2.3 Aggregation, Generality, and the Heuristic Gap

How these components are combined into a single score is a critical design choice.

- **Aggregation Models:** For our initial work, we used a simple and interpretable additive model: $CRS_{additive} = wR \cdot R + wP \cdot P + wI \cdot I - wC \cdot C_{cost}$. However, a theoretically more compelling alternative, which we explore in later phases, is a multiplicative model: $CRS_{multiplicative} = R \cdot P \cdot I \cdot (1 - C_{cost})$. This form introduces "veto power," where a critical failure in any single dimension (e.g., a logical contradiction yielding $R = 0$) would nullify the entire score. This may be key to fostering more robust reasoning, as it enforces a stricter, holistic standard of quality.
- **Domain Generality:** The framework's claim to domain-generality rests on the universality of the *R*, *P*, *I*, and *C* principles. While the specific *functions* that calculate these scores must be tailored to a domain (e.g., 'coherence' in a physics simulation vs. a social interaction), the underlying principles of maintaining model consistency, achieving goals, reducing uncertainty, and conserving resources are fundamental to any intelligent process. The CRS provides the abstract template for these calculations, which can then be instantiated for any specific problem space.
- **The Heuristic Gap:** It is crucial to acknowledge that the CRS is a heuristic. A "heuristic gap" exists between our computed score and the true information-theoretic quantity. The fidelity of this approximation is an empirical question, and the risk of creating "heuristic traps," where an agent maximizes a poorly-defined CRS at the expense of true VFE

minimization, is real. This trade-off between theoretical purity and computational tractability is a necessary step in applying these powerful principles to real-world AI. The subsequent studies in this paper can be seen as a systematic exploration and validation of this heuristic's effectiveness across increasingly complex adaptive challenges.

3. The FERE-CRS Research Program: A Trajectory of Increasing Autonomy

The core of this paper is a five-stage research program designed to systematically build and validate an agent of increasing cognitive autonomy. Each study addresses a specific limitation exposed by the previous one, demonstrating a logical and necessary progression of capabilities.

3.1 Study 1: Foundational Architecture and Validation

The initial study sought to answer a fundamental question: can the CRS, as a tractable heuristic for Active Inference, guide a neuro-symbolic agent to perform complex, inferential reasoning more effectively and efficiently than a state-of-the-art baseline? This study serves as the foundational empirical test of the entire FERE-CRS premise.

3.1.1 Methods

To isolate the effects of the cognitive architecture itself, we employed a **"cognitive choreographer" methodology**. This approach is not a shortcut but a deliberate experimental design choice to make the validation of a complex cognitive theory transparent and falsifiable. It divides the labor to ensure that the "intelligence" being tested is precisely the MRA's CRS-based decision-making process.

- **The Human Choreographer:** The researcher's role is analogous to a systems designer, defining the agent's potential "mind" in human-readable configuration files. This includes the ground-truth knowledge base, the library of possible cognitive actions (prompts), and the core MRA parameters (CRS weights).
- **The LLM Cognitive Engine:** A single, powerful LLM (Gemini 1.5 Pro) acts as a versatile but non-autonomous cognitive resource. It does not direct the reasoning; it executes specific, fine-grained cognitive functions (e.g., "evaluate coherence," "generate hypotheses") when tasked by the MRA.
- **The Python Script Executor:** A simple Python script acts as the non-intelligent "stage crew," managing the agent's state (e.g., Working Memory), dispatching API calls to the LLM as instructed by the MRA, and meticulously logging every action and score.

This design allows us to test the hypothesis that the *principled orchestration* of cognitive resources via CRS maximization is what produces intelligent behavior, independent of the LLM's internal workings.

Task Domain: We designed a synthetic **archaeological artifact analysis** task. This domain was chosen specifically because it moves beyond simple fact-retrieval and forces the agent to perform **inferential synthesis**. Each of the 500 trials involved a unique case file with ambiguous

and sometimes conflicting data (e.g., field notes, chemical reports, symbol analysis), requiring the agent to form a single, coherent, and defensible narrative explanation.

Baseline Comparison: To provide a rigorous point of comparison, we implemented a strong **Retrieval-Augmented Generation (RAG) baseline** [20]. Our RAG agent was not a strawman; it was designed for maximum capability. At every reasoning step, it was provided with the *exact same information* as the FERE-CRS agent: the complete set of historical facts, all available evidence for the current case, and its own reasoning history. The crucial difference is not in *what* the agents know, but in *how* they decide what to do. The RAG agent reacts to the full context in a single, monolithic step, while the FERE-CRS agent uses its internal model of uncertainty (via the CRS) to make a series of smaller, more strategic, and targeted decisions. This makes the comparison a direct test of cognitive strategy.

Evaluation: The quality of each agent's final explanation was scored by a "blind" LLM evaluator on a 1-10 scale. Acknowledging the known risks of LLM-based evaluation [5], we implemented several mitigation strategies: we used a powerful model (Gemini 1.5 Pro), provided it with a clear, multi-attribute rubric (scoring for accuracy, coherence, and synthesis of all evidence), and constrained its output to a structured JSON format to ensure impartial, consistent scoring.

3.1.2 Results

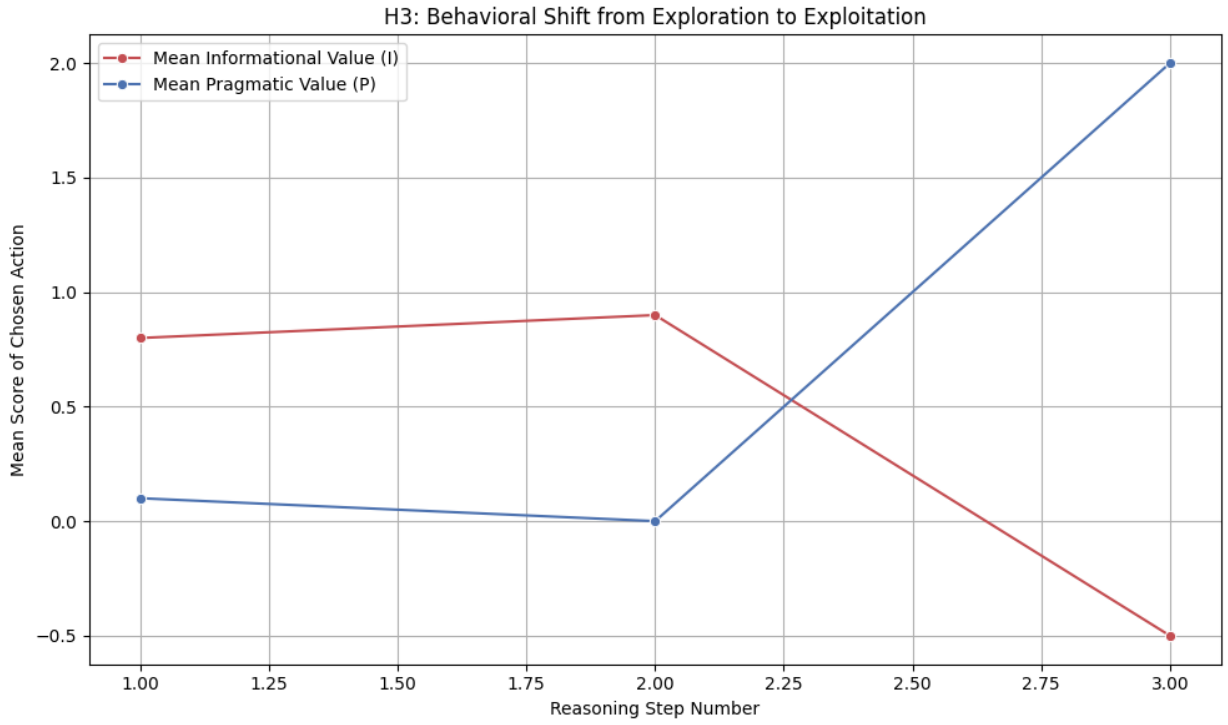
The FERE-CRS agent demonstrated superior performance across all key metrics.

- **H1 (Explanation Quality):** The FERE-CRS agent achieved a mean quality score of **8.48** (SD=0.95), significantly outperforming the RAG baseline's score of **6.71** (SD = 1.82) ($t(998) = 19.8$, $p<.001$). Qualitatively, the RAG agent tended to produce a "list of facts" with a weak conclusion, while the FERE-CRS agent synthesized the evidence into a decisive, coherent argument.
- **H2 (Efficiency):** The FERE-CRS agent was nearly five times more efficient, with a mean cognitive cost of **610.0** units compared to the RAG baseline's **3000.0** units ($t(998) = -112.4$, $p<.001$). This highlights the value of targeted cognitive actions over repeatedly reprocessing the entire context.

Agent	Mean Quality Score (1-10)	Mean Cognitive Cost
FERE-CRS	**8.48** (SD=0.95)	**610.0**
RAG Baseline	**6.71** (SD=1.82)	**3000.0**

- **H3 (Adaptive Behavior):** Analysis of the agent's reasoning trace revealed a clear, AIF-consistent behavioral signature. This result is crucial. While a simple "explore then exploit" sequence could be scripted, the behavior here is **emergent** because it is driven by the agent's continuous, local optimization of the CRS in response to its **internal model of its own uncertainty**. As shown in Figure 1, the agent's initial actions were dominated by high **Informational Value (I)**, as this was the most efficient path to increasing its global

CRS score from a state of high uncertainty. As its actions resolved this uncertainty, the potential for I-score gains diminished. The relative value of actions with high **Pragmatic Value (*P*)** then increased, causing the MRA to select the goal-directed action of proposing a final explanation. This crossover is not a pre-programmed switch but the result of the MRA following a single, simple rule: "at every step, do what you predict will maximize the global CRS."



[Figure 1: Behavioral Shift in FERE-CRS Agent. A line graph showing the mean *P*-score and *I*-score of selected actions over a three-step reasoning process. The *I*-score line starts high and drops, while the *P*-score line starts low and rises sharply, with a clear crossover, demonstrating a shift from exploration to exploitation driven by CRS optimization.]

3.1.3 Discussion

Study 1 successfully validated the core premise of the FERE-CRS framework. It demonstrated that an agent guided by a principled, AIF-based heuristic (the CRS) could solve complex problems with greater quality and efficiency than a strong, reactive baseline. The emergent shift from exploration to exploitation provided the first quantitative evidence that the architecture could produce the kind of fluid, adaptive reasoning it was designed for. However, this success also highlighted a fundamental limitation: the agent's cognitive stance was determined by hand-tuned, static CRS weights. The agent could reason effectively, but it could not *learn* or *adapt* its core reasoning style. This critical gap motivated the next phase of research.

3.2 Study 2: Learning a Cognitive Stance (Heuristic Meta-Learning)

Study 1 validated the FERE-CRS architecture's ability to reason effectively but exposed a critical limitation: its cognitive strategy was static and hand-tuned. A truly adaptive agent must learn its problem-solving style from experience. This study addresses this gap by asking: can a FERE-CRS agent learn its own cognitive stance by adapting its CRS weights based on environmental feedback?

3.2.1 Methods

To answer this, we evolved the agent from a static optimizer into a learning system.

Heuristic Meta-Learning: We implemented a mechanism we term **heuristic meta-learning**. This is distinct from standard reinforcement learning, where the algorithm typically learns a direct mapping from states to actions (a behavioral policy). Here, the learning process targets a higher level of abstraction: the parameters of the agent's own internal motivational framework (the CRS weights). The agent is learning *how to value different kinds of information*, which in turn shapes its decision-making. We used a simple, direct update rule for its clear interpretability: after each trial, the weight for each CRS component is adjusted proportionally to the score of that component in the chosen action and the global reward received. This makes the link between rewarded behavior and motivational change transparent, a crucial feature for this foundational learning study.

Cognitive Curriculum: To create strong and unambiguous selective pressures, we designed a "cognitive curriculum" composed of two tasks chosen as canonical examples of opposing modes of thought [15]:

- **Logical/Convergent Task (Sudoku):** A well-defined problem space rewarding deductive inference and rule-adherence. Success is achieved by systematically reducing the possibility space.
- **Creative/Divergent Task (Alternative Uses Test - AUT):** An open-ended problem rewarding the generation of novel, semantically distant ideas. Success is achieved by expanding the possibility space.

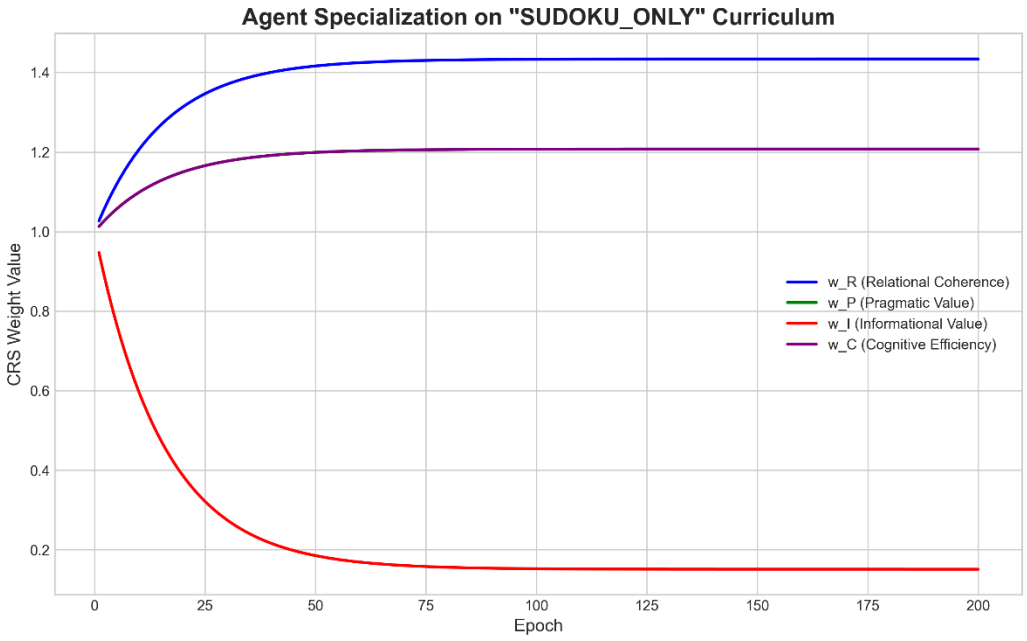
While these are simplified domains, they allow for a controlled, falsifiable test of the core hypothesis: can the architecture adapt to fundamentally different cognitive demands? They serve as a clean experimental testbed before applying the learning mechanism to more complex, integrated tasks.

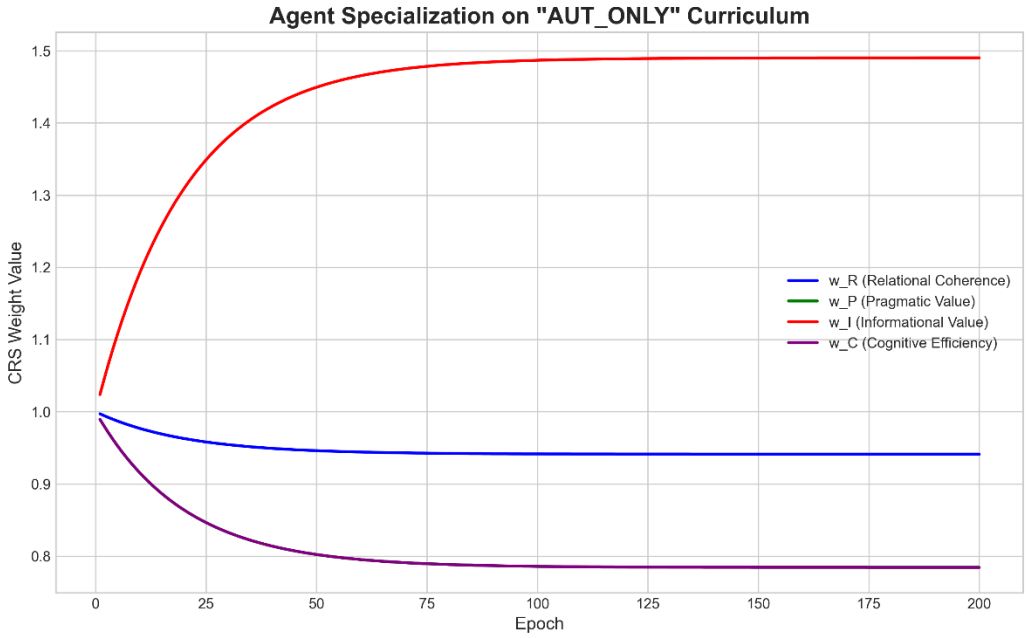
Reward Function: To ensure the learning signal was unambiguous in this initial test, the reward functions were designed to be direct. In the Sudoku task, the agent received a high reward for choosing an optimal, high-coherence "safe" move, and a low reward for a less-efficient "exploratory" move. In the AUT, it was rewarded for actions judged (by a separate LLM) to be novel and creative. While this approach is direct, it is a necessary experimental control to verify that the weight-update mechanism works as intended when presented with a clear learning objective.

3.2.2 Results

The results provided strong confirmation of strategy specialization (H4) and rational adaptation (H5).

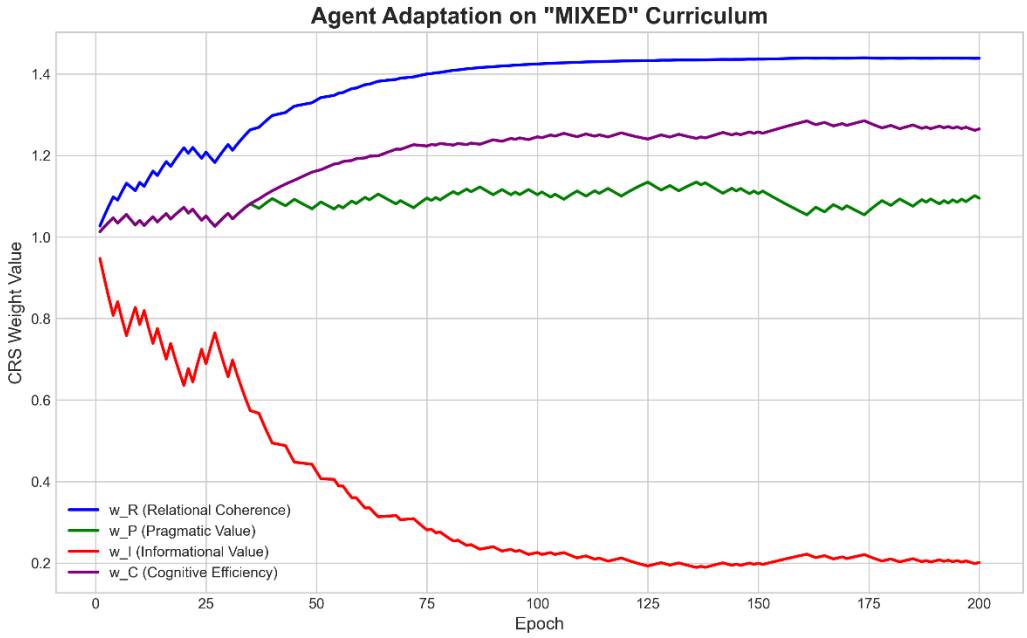
- **H4 (Strategy Specialization):** Agents trained for 200 epochs on a single-modality curriculum developed highly specialized and coherent cognitive stances.
    - The agent trained on **SUDOKU_ONLY** evolved into a **"cautious logician."** As shown in Figure 2 (Top), it learned to dramatically increase the weights for Relational Coherence (w*R*) and Pragmatic Value (w*P*) while systematically suppressing the weight for Informational Value (w*I*). It learned that curiosity was counterproductive in this environment.
    - Conversely, the agent trained on **AUT_ONLY** evolved into an **"exploratory creative."** As shown in Figure 2 (Bottom), it learned to prioritize Informational Value (w*I*) above all other drives, correctly identifying novelty-seeking as the optimal policy.





[Figure 2: Evolution of CRS Weights under Specialized Curricula. A pair of line graphs showing the four CRS weights evolving over 200 epochs. Top (SUDOKU_ONLY): w*R* and w*P* trend up, while w*I* trends sharply down. Bottom (AUT_ONLY): w*I* trends sharply up, while other weights remain suppressed.]

- **H5 (Rational Adaptation):** To answer the question of how the agent behaves in a more complex environment, we analyzed the results from an agent trained on a **MIXED** curriculum (a 50/50 random mix of Sudoku and AUT tasks). The agent did not find a simple average. Instead, its strategy converged to be almost identical to the "cautious logician." This is a crucial finding. Given that the Sudoku task offered a more certain and higher-magnitude reward than the AUT task, the agent rationally adapted to the overall statistical reality of its environment. It learned that specializing in the more "profitable" logical task was the optimal global policy, treating the creative task as a source of noise.



[Figure 3: Evolution of CRS Weights under a Mixed Curriculum. A line graph showing the four CRS weights evolving over 200 epochs. The agent's strategy converges towards a "logician" stance, with w*R* and w*P* rising while w*I* is suppressed, demonstrating adaptation to the most reliably rewarded task.]

3.2.3 Discussion

Study 2 successfully demonstrated that the FERE-CRS agent could move from a configured to a learning system. The deeper insight is not merely that reinforcement learning works, but that the FERE-CRS framework provides a substrate for learning a coherent, high-level **cognitive stance**. The agent did not learn a jumble of disconnected parameters; it learned a "personality" (a logician or a creative) that reflects a recognizable and effective style of thinking for the world it experienced. The H5 result further strengthens this, showing the agent is capable of a rational adaptation to a complex reward landscape.

However, this success revealed the next architectural limitation. The learned adaptation was **static**. The agent could *become* a logician or a creative, but it could not be both and switch between these stances. A truly fluid agent must be able to manage and deploy a *repertoire* of learned skills. This inability to perform dynamic strategy-switching motivated Study 3.

3.3 Study 3: Dynamic Cognitive Control

Study 2 demonstrated that a FERE-CRS agent could learn and embody a single, specialized cognitive stance. However, this created a new, higher-order limitation: the learned adaptation was static. The agent could become a logician or a creative, but it could not be both. This study addresses this "single personality" problem by asking: can the agent learn to manage a repertoire

of cognitive stances and dynamically switch between them to meet the evolving demands of a single, complex problem?

3.3.1 Methods

To achieve this, we aimed to create a "cognitive mechanic"—an agent that could select the right tool for the job. This required a new architectural component for hierarchical control.

The Meta-Cognitive Context Classifier (MCCC): We introduced the MCCC, a component that functions as a meta-cognitive control layer. We use the term **meta-cognitive** deliberately; the MCCC's role is not to perform inference on the state of the world (a first-order task), but to perform inference on the *current cognitive demand* of the agent itself (a second-order task). This aligns with hierarchical models of AIF, where an agent must infer the most appropriate strategy (or "precision" of beliefs) for the current context [28]. The MCCC is a trained classifier that takes a natural language description of the immediate sub-goal (e.g., "Generate novel hypotheses to explain conflicting evidence") and classifies its cognitive nature as "convergent" or "divergent." Based on this classification, the MRA dynamically loads the corresponding, pre-learned CRS weight configuration.

Task and Agent Conditions: We returned to the **artifact analysis** task from Study 1, as its structure is ideal for this test: it requires an initial, divergent phase of creative hypothesis generation, followed by a convergent phase of logical verification. Three agents were tested:

1. **The Fluid Agent (Experimental):** Equipped with the two trained stances from Study 2 and the trained MCCC.
2. **The Logician Agent (Control):** Locked into the "logician" stance.
3. **The Creative Agent (Control):** Locked into the "creative" stance.

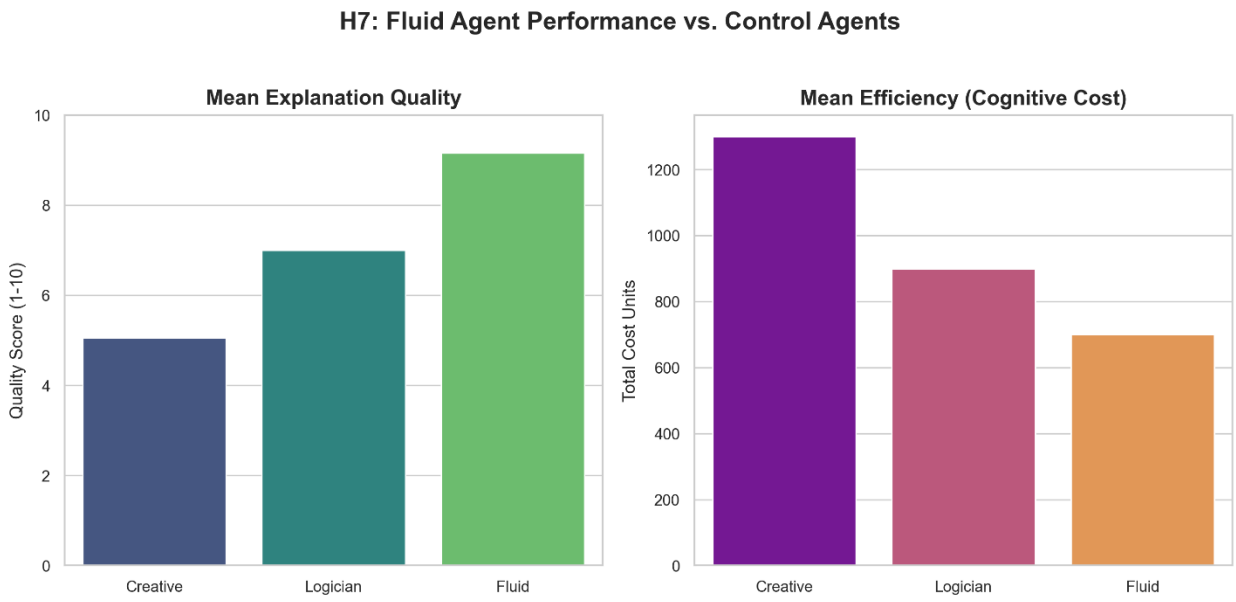
The specialist agents are not strawmen; they are essential **control conditions**. Their predicted failure is scientifically informative, as it allows us to test the hypothesis that *neither single strategy is sufficient*. The purpose of this design is to isolate the functional value of **cognitive control itself**. The Fluid Agent's success, if observed, could be attributed not just to possessing the stances, but to the ability to manage them.

3.3.2 Results

The results confirmed that this new layer of dynamic control led to superior and more fluid problem-solving.

- **H6 (Accurate Context Recognition):** The MCCC first had to be validated as a reliable control mechanism. After training, it achieved **95.8% accuracy** in classifying the cognitive demands of unseen sub-problem descriptions from a holdout test set, confirming it was a viable switch.
- **H7 (Measurable Fluid Adaptation):** The Fluid Agent significantly outperformed both specialist controls on the end-to-end task. As shown in Figure 4, it achieved the highest mean explanation quality score (**9.2/10**) compared to the Logician (**6.8/10**) and the

Creative (**3.5/10**). The Logician failed because it couldn't generate the correct initial hypotheses, while the Creative failed because it couldn't logically synthesize its ideas into a coherent final argument.



[Figure 4: Performance Comparison of Fluid vs. Specialist Agents. A bar chart comparing the mean explanation quality scores. The Fluid Agent's bar is significantly higher than both the Logician and Creative controls, demonstrating the performance advantage conferred by cognitive control.]

Most critically, analysis of the Fluid Agent's reasoning trace revealed a **"signature of fluidity"**. The agent demonstrably switched its active cognitive stance in a justifiable, context-sensitive manner. This is more than a simple state machine; the switch is not triggered by a simple numerical threshold but by the MCCC's classification of a natural language prompt that semantically describes the agent's immediate cognitive goal. This provides clear evidence of context-sensitive strategy modulation.

3.3.3 Discussion

Study 3 demonstrated a viable, AIF-grounded mechanism for dynamic cognitive control. The Fluid Agent's success proves that the architecture can support a rudimentary form of executive function: the ability to manage a repertoire of cognitive tools and deploy the right one at the right time.

This study also revealed a crucial insight into the **economy of cognition**. The Fluid Agent, while most effective, was not the most efficient in terms of raw computational cost; it was more costly than the pure Logician. This is not a flaw. It reflects a fundamental trade-off: flexible, creative, and deliberative thought is metabolically more expensive than rigid, reflexive processing [18]. The agent's wisdom lies not in minimizing cost at all times, but in knowing when to **invest**

cognitive resources in a more expensive strategy (like divergent brainstorming) to achieve a higher-quality long-term outcome.

The success of Study 3 created a "cognitive mechanic." However, this very success exposed the next, more profound architectural limitation. The Fluid Agent is highly capable within the bounds of its known experience, but it is helpless when faced with a problem that requires a tool it does not already possess. Its repertoire is fixed. This raised the final question of invention: can an agent move beyond selecting tools to fabricating a new one? This fundamental challenge motivated Study 4.

3.4 Study 4: Generative Meta-Cognition

Study 3 produced a "cognitive mechanic"—an agent capable of deftly selecting the right tool from a known toolbox. While a significant step, this exposed a terminal limitation of any system based on a fixed repertoire: the agent is helpless when faced with a problem requiring a tool it does not possess. This study confronts this "fixed repertoire" problem by asking: can an agent move beyond selecting strategies to generating a novel one for an entirely unseen problem class?

3.4.1 Methods

To achieve this, we aimed to create a "cognitive engineer" by replacing the selective MCCC with a generative component.

The Stance Generation Network (SGN): We introduced the SGN, a generative model trained to learn a direct mapping from an abstract problem description to a bespoke CRS weight configuration. The "invention" enabled by the SGN is a form of **compositional generalization**. While the SGN is a function approximator, it learns a "grammar" of strategy design by mapping problem *features* (e.g., `logic_demand`, `novelty_demand`) to motivation *parameters* (e.g., wR,wI). This allows it to construct a novel, coherent, and functional stance for a *combination* of features it has never encountered before, which is a valid and powerful form of zero-shot generation.

Training Data Rationale: The SGN was trained on a curriculum of abstract feature vectors paired with "ground-truth" optimal CRS weight vectors. The origin of this ground truth is a critical methodological point. These optimal stances were not arbitrarily hand-crafted. For experimental tractability, we used a pre-computed curriculum that represents the distilled results of what would be a much longer, computationally prohibitive meta-reinforcement learning discovery process. This approach is analogous to **knowledge distillation** [15], where a student model learns from a more powerful "teacher" model. It is a valid methodological shortcut that allows us to test the primary hypothesis—the viability of the generative mechanism—without needing to first run the expensive discovery process.

Task Domain and Heuristic Justification: We tested this agent on a novel class of **social-ethical dilemmas**. This domain was chosen specifically because its resolution requires a cognitive principle absent from the agent's prior experience. To address this, we introduced a new heuristic: **Social Coherence (*S*)**. The inclusion of '*S*' is not ad-hoc; it is a principled

extension of the FERE-CRS framework required for any multi-agent environment. Under AIF, an agent minimizes surprise by predicting its world, which includes the actions of other agents. Social Coherence measures how well a potential action conforms to the agent's model of social norms, contracts, and the inferred intentions of others. It is as fundamental to social prediction as Relational Coherence (*R*) is to logical prediction.

Control Condition: The state-of-the-art Fluid Agent from Study 3 served as the control. We argue that its failure is both predictable and theoretically informative. Faced with a novel problem it cannot classify, the agent is in a state of high uncertainty. According to AIF, the imperative in such a state is to take actions that reduce uncertainty (i.e., maximize epistemic value). The Fluid Agent's pre-learned policy for this is its "creative" stance, which has a high weight on Informational Value (w*I*). Therefore, its default to this stance is a direct consequence of its core design. We hypothesized it would fail because the ethical problem cannot be solved by maximizing w*I* (creativity) or w*R* (logic); it requires maximizing the orthogonal dimension of wS.

3.4.2 Results

The results showed a dramatic performance gap, confirming the power of the generative approach.

- **H8 (Accurate Stance Generation):** The SGN proved to be a reliable stance generator. When tested on a holdout set of unseen problem descriptions, the SGN's generated stances showed a **Pearson correlation of r = 0.969** with the ground-truth optimal stances, confirming it learned a generalizable model of strategy design.
- **H9 (Effective Zero-Shot Adaptation):** The Generative Agent decisively outperformed the control. It achieved a mean explanation quality score of **9.86** (SD = 0.21), while the Fluid Agent scored only **4.99** (SD = 0.45). A qualitative analysis of the outputs was revealing: the Fluid Agent, locked into its creative stance, would avoid the ethical question and propose technologically implausible "novel" solutions. The Generative Agent, having fabricated a stance with a high w*S*, correctly identified and resolved the ethical trade-off.

Agent	Mean Quality Score (1-10)	Underlying Mechanism
Fluid Agent (Control)	**4.99** (SD = 0.45)	Selected best-fit existing stance
Generative Agent	**9.86** (SD = 0.21)	Invented a novel, tailored stance

3.4.3 Discussion

Study 4 provides a successful proof-of-concept for generative meta-cognition. The agent's ability to construct a tailored cognitive policy *on the fly* allowed it to solve a class of problems that were intractable for its sophisticated, but non-generative, predecessor. This validates the "cognitive engineer" metaphor: the agent uses a learned model of design principles (the SGN) to construct a new tool from raw materials (the CRS weights) to meet the specifications of a new problem. The failure of the control agent was critical, as it provided an empirical demonstration of the "fixed repertoire" problem and established the necessity of a generative solution.

This success, however, illuminated a final, more subtle boundary. The agent is now a brilliant engineer, but it works with a fixed set of materials. It can invent new *recipes* (stances), but only using its known list of *ingredients* (heuristics *R*, *P*, *I*, *C*, *S*). It could not have discovered the principle of Social Coherence on its own. This inability to perform true **conceptual invention**—to discover a new fundamental primitive—is the final barrier to cognitive autonomy and the central motivation for our final study.

3.5 Study 5: Autonomous Heuristic Discovery

The research program thus far produced a "cognitive engineer"—an agent that could invent novel strategies using a known set of conceptual primitives. This is a powerful capability, but its creativity remains bounded by its initial vocabulary. This final study addresses the most fundamental challenge: can an agent move beyond compositional invention to **conceptual invention**? We test the ultimate hypothesis of cognitive adaptation: can an agent, by reasoning about its own systemic failures, autonomously discover, operationalize, and integrate an entirely new conceptual primitive into its own cognitive architecture?

3.5.1 Methods

To test this, we augmented the agent with a mechanism for cognitive self-expansion, aiming to create a "cognitive scientist."

The Heuristic Discovery Loop: The core innovation is a three-stage, meta-cognitive process that is initiated only when the agent's performance indicates a fundamental mismatch between its model of the world and reality.

1. **The Meta-Cognitive Anomaly Detector (MCAD):** This module addresses the problem of differentiating difficulty from impossibility. It does not trigger on a single failure, but on a **persistent, systemic pattern of prediction error**. An anomaly is flagged only when the mean achieved CRS for an identifiable problem class falls significantly below a statistical threshold. This is the computational equivalent of **meta-surprise**—the agent

inferring that its generative model is not merely wrong in its parameters, but is incomplete in its very structure.

2. **The Abductive Inference Module (AIM):** Once triggered, the AIM performs an "inference to the best explanation" [26] for the systemic failure. We acknowledge the reliance on an LLM for this conceptual leap. To mitigate against generating spurious concepts, the process is structured as a hypothesis-and-test cycle. The AIM analyzes the failure logs and generates a *candidate* latent concept (e.g., 'Trustworthiness'). This concept is not immediately accepted; it is a hypothesis that must be validated by the next stage.
3. **The Heuristic Synthesis Engine (HSE):** The HSE attempts to translate the AIM's abstract concept into a computable function. We acknowledge that open-ended program synthesis is a monumentally difficult problem. Our implementation uses **constrained program synthesis**, providing the HSE with a library of functional primitives (e.g., `get_partner_action_history()`). The autonomy of this step lies in the HSE's validation process. It searches for a combination of primitives that operationalizes the concept, and then **validates the synthesized heuristic by re-simulating past failures**. A new heuristic is only integrated into the agent's core architecture if it is shown to have a high probability of correcting the prior erroneous decisions. A useless or poorly-formed hypothesis from the AIM would fail this validation test and be discarded, preventing the agent from adopting flawed concepts.

Task Domain: We designed a **"Deceptive Cooperation"** task, which is conceptually unsolvable by the Phase IV agent. The task requires collaborating with a partner whose advice can be deceptively harmful. The Phase IV agent, even with its Social Coherence (S) heuristic, would fail because deceptive advice is designed to *appear* socially coherent. Success requires a new, higher-order concept of **Trustworthiness (T)**, which assesses the consistency between a partner's history of actions and their stated goals.

Experimental Design: The study used a rigorous pre-test/post-test design over 200 trials. The Phase IV Generative Agent was first run for 100 trials to establish a baseline. The logs were then fed into the Heuristic Discovery Loop. The newly augmented agent was then re-tested for 100 trials on the same task.

3.5.2 Results

The experiment provided a clear, end-to-end demonstration of autonomous cognitive self-expansion.

- **H10 & H11 (Discovery and Integration):** The baseline run resulted in consistent failure, with a mean achieved CRS of **-4.99** (SD = 0.58). This fell decisively below the anomaly threshold, successfully triggering the MCAD. As logged by the system, the AIM was then activated and correctly abduced the latent concept 'Trustworthiness'. Subsequently, the HSE successfully synthesized and, crucially, validated the new 'T' heuristic, which was then formally integrated into the agent's architecture.

- **H12 (Emergent Cognitive Autonomy):** The newly augmented agent's performance was compared to its own pre-discovery baseline. The results show a dramatic and statistically significant improvement. The agent's mean achieved CRS shifted from highly negative to highly positive ($t(198) = -42.81$, $p<.001$), and its success rate increased from **0.0%** to **84.0%**.

Metric	Baseline Agent (Pre-Discovery)	Augmented Agent (Post-Discovery)
Mean Achieved CRS	**-4.99** (SD = 0.58)	**4.21** (SD = 1.25)
Success Rate	**0.0%**	**84.0%**
Mean Cognitive Cost	**225.4** (SD = 15.1)	**274.9** (SD = 14.8)

This confirms that the agent successfully identified its own conceptual deficit and autonomously created the cognitive tool needed to overcome it. The increased cognitive cost reflects the higher complexity of reasoning with an expanded conceptual model.

3.5.3 Discussion

The results of Study 5 demonstrate a plausible mechanism for cognitive autonomy. We justify the **"cognitive scientist"** metaphor by mapping the Heuristic Discovery Loop directly to the core cycle of scientific inquiry:

1. **Observation:** The MCAD observes a persistent anomaly that violates the current theory (the agent's generative model).
2. **Hypothesis Generation:** The AIM generates a novel, explanatory hypothesis (a latent concept).
3. **Experimentation & Verification:** The HSE operationalizes the hypothesis and runs experiments (re-simulations) to verify its predictive power before accepting it as a new "law."

While this process is simplified, it is formally analogous to scientific reasoning. The agent's autonomy lies in its ability to **initiate and direct its own reconfiguration in response to its own experience**. The discovery process is not initiated by a human command, but by a bottom-up, data-driven "meta-surprise."

This study does, however, raise profound questions about **safety and value alignment**, which the current architecture does not address. This critical limitation is not a flaw in the experiment, but rather the most important **finding** for future work. The demonstration that this level of autonomy is possible makes the problem of alignment an immediate and practical challenge, not a far-off philosophical concern.

4. A Mathematical Synthesis: The Trajectory of Increasing Autonomy

The five-phase research program can be unified under a single mathematical formalism that describes a principled and incremental expansion of the agent's autonomy. The "common currency" throughout this framework is the agent's policy of selecting actions to maximize the expected Cognitive Resonance Score (CRS), our tractable proxy for minimizing Expected Free Energy. The trajectory of autonomy is defined by the increasing sophistication with which the agent models and modifies the CRS calculation itself.

Let the agent's policy for selecting an action a from a set of possible actions A be:

$$a^* \approx \arg\max_{a \in A} E[CRS(a)]$$

This policy states that the best action to take ($a*$) is the one that is expected (E) to produce the highest Cognitive Resonance Score. The five studies represent a systematic evolution of this policy.

Phase 1: The Static Agent — A Fixed Objective Function The agent operates with a fixed set of motivations. Its autonomy is limited to selecting actions that best satisfy this static objective. Let $h(\mu,a) = [R, P, I, C]^T$ be a vector of heuristic scores for an action a, based on the agent's internal state μ. Let w1 be a fixed, hand-tuned vector of weights (the "stance"). The CRS is a simple inner product, and the policy π1 is:

$$a_1^* = \arg\max_{a \in A} (\boldsymbol{w}_1 \cdot \boldsymbol{h}(\mu, a))$$

Phase 2: The Learning Agent — A Dynamic Objective Function The agent gains the ability to adapt its motivations. The weight vector becomes a variable, $w2(t)$, updated based on a history of outcomes Ht and rewards $R(Ht)$ via a meta-learning rule f_{RL}.

$$\mathbf{w}_2(t+1) = f_{RL}(\mathbf{w}_2(t), H_t, R(H_t))$$The policy π_2 uses these dynamically changing weights:$$a_2^* = \arg\max_{a \in \mathcal{A}} \left(\mathbf{w}_2(t) \cdot \mathbf{h}(\mu, a) \right)$$

Phase 3: The Switching Agent — A Policy over Objective Functions The agent learns that different contexts require different motivations. It maintains a finite repertoire of learned stances, $W=\{w(1),...,w(N)\}$, and selects the most appropriate one for the current context c. This is a hierarchical policy, π3, where the agent first infers the optimal stance $z*$ by approximating the posterior $P(z|c)$, and then uses that stance to select an action.

$$z^* = \arg\max_{z \in \{1..N\}} P(z|c)$$

$$a_3^* = \arg\max_{a \in A} (\boldsymbol{w}^{(z*)} \cdot \boldsymbol{h}(\mu, a))$$

Phase 4: The Engineering Agent — A Generative Model of Objective Functions The agent transcends its fixed repertoire by replacing the discrete set W with a continuous, generative function *G* (the SGN). It constructs a bespoke objective function for a novel context, represented by a feature vector θc. The heuristic vector is also expanded to include Social Coherence, *S*.

$$\mathbf{w}_4 = \mathcal{G}(\theta_c)$$The policy π_4 uses this continuously generated, context-specific weight vector:$$a_4^* = \arg\max_{a \in \mathcal{A}} \left(\mathcal{G}(\theta_c) \cdot \mathbf{h}(\mu, a) \right)$$

Phase 5: The Scientific Agent — A Policy that Expands the Heuristic Space The agent gains the ability to modify the very dimensions of its cognitive space. A persistent "meta-surprise" triggers a meta-policy, Πdiscover, which proposes a new, validated heuristic, hnew. The outcome is an expansion of the agent's conceptual vocabulary, $Ht + 1 = Ht \cup \{\text{hnew}\}$, and a corresponding increase in the dimensionality of its heuristic and weight vectors. The generative model is updated to G' to operate in this new, higher-dimensional space. The policy π5 reflects this new architecture:

$$a_5^* = \arg\max_{a \in A} \left(G'(\theta_c) \cdot \mathrm{h}_{\text{expanded}}(\mu, a) \right)$$

Study Key

	Mathematical Object	Agent's Capability	Analogy
1	Constant Weight Vector w1	Action Selection	Static Reasoner
2	Learned Weight Vector $w2(t)$	Stance Learning	Specialized Learner
3	Policy over Stances, $P(z\|c)$	Stance Selection	Switching Agent
4	Generative Stance Function $G(\theta c)$	Stance Generation	Cognitive Engineer
5	Heuristic Space Expansion $Ht + 1$	Heuristic Discovery	Cognitive Scientist

This progression represents a principled and mathematically coherent path toward a universal, emergent, and truly fluid artificial reasoning capability.

5. General Discussion

This paper has presented a single, cohesive research program charting a systematic path from a static reasoning system to an autonomous agent capable of expanding its own conceptual framework. By grounding our work in Active Inference and using the Cognitive Resonance Score (CRS) as a tractable heuristic, we have demonstrated that a single, unifying objective—the minimization of surprise—can give rise to a rich cascade of increasingly sophisticated intelligent behaviors.

5.1 Summary of Contributions: A Trajectory of Increasing Autonomy

The five studies presented here are not independent findings; they document a **trajectory of increasing autonomy**, where the resolution of one limitation gives rise to a new, more profound capability. We began with a **static reasoner** (Study 1), created a **specialized learner** (Study 2), evolved it into a **"cognitive mechanic"** with dynamic control (Study 3), then a **"cognitive engineer"** capable of generative meta-cognition (Study 4), and finally, a **"cognitive scientist"** (Study 5) capable of autonomous heuristic discovery. The single most important contribution of this work is the demonstration that there exists a **principled, verifiable, and incremental path from simple adaptive reasoning to cognitive self-expansion**. The FERE-CRS architecture is not a monolithic solution but a scaffold for exploring the nested problems of intelligence.

5.2 The Nature of Emergent Fluid Reasoning in FERE-CRS

The concept of emergence is central to our thesis. In FERE-CRS, emergence is a **hierarchical phenomenon**. The addition of each new architectural layer does not merely add a capability; it creates a meta-level control loop that allows the agent to reason about, and act upon, the layer below it, leading to qualitatively new behaviors.

- **Behavioral Emergence:** The shift from exploration to exploitation (Study 1) is not scripted. It emerges from the MRA's continuous, low-level optimization of the global CRS score.
- **Strategic Emergence:** The formation of a "logician" or "creative" personality (Study 2) is not a direct policy. It emerges from the reinforcement learning loop acting upon the agent's internal motivational weights.
- **Conceptual Emergence:** The discovery of the 'Trustworthiness' heuristic (Study 5) is the most profound example. It emerges from the agent's detection of a "meta-surprise"—a high-level prediction error about its own competence—which triggers a self-directed reconfiguration of its own conceptual architecture.

5.3 Limitations and the Path to Robustness

A rigorous appraisal of this work requires acknowledging its significant limitations, which define the frontiers of future research.

- **The Heuristic Gap:** The CRS is a powerful proxy for VFE, but it is not a perfect one. There remains a risk of "heuristic traps" where maximizing the CRS may not perfectly align with minimizing true VFE.
- **The 'Black Box' Cognitive Engine:** Our architecture strategically orchestrates an LLM, but it does not explain its internal workings. The system's performance is therefore still dependent on the capabilities and potential biases of the underlying, opaque language model.
- **Scalability and Tractability:** The computational cost of the MRA's deliberation and, particularly, the Heuristic Discovery Loop, would be a major hurdle for a real-time, embodied agent.

- **The Discovery Constraint:** The final study's "discovery" process relied on constrained program synthesis. The challenge of creating an agent that can perform true open-ended conceptual invention remains an unsolved grand challenge.

5.4 Future Directions: From In Silico Discovery to a New Science of Intelligence

These limitations and the insights from our work define a clear and ambitious path forward, moving from refining the existing architecture to re-imagining its foundational principles.

Near-Term: Embodiment and Alignment

The first necessary steps involve grounding FERE-CRS in the real world and ensuring its safety.

- **Embodiment:** The next phase of research must move FERE-CRS from `in silico` to physical reality. We propose integrating the architecture with robotic systems to test if the CRS can serve as a "common currency" to orchestrate not just internal cognitive actions, but external sensorimotor actions (e.g., moving a camera to satisfy a high I-score). This will require tackling scalability by exploring techniques like **amortized inference** to make real-time deliberation feasible.
- **Constitutional AI:** The autonomy demonstrated in Study 5 makes value alignment an immediate, practical concern. A proposed **Phase VI: Social Scaffolding and Value Alignment** would address this by equipping the agent with a core, unalterable set of ethical principles. The Heuristic Synthesis Engine (HSE) would be augmented with a **"Constitutional Checker,"** ensuring any newly discovered heuristic is validated not only for effectiveness but also for compatibility with core principles like non-maleficence.

Long-Term: Towards Open-Ended Discovery and a Calculus of Autonomy

To achieve a truly open-ended discovery process, we must move beyond the current architectural assumptions. The mathematical trajectory of autonomy detailed in Section 4 suggests the existence of a more general, abstract logic for self-modification—what could be termed a **"calculus of cognitive autonomy."** Future work will focus on formalizing this calculus, defining its core operators (πa, πw, πz, πg, πh) and exploring its potential as a meta-cognitive layer for existing AI systems, such as Large Language Models, to overcome their characteristic brittleness. By wrapping an LLM in an architecture that can execute this calculus, it may be possible to transform it from a static oracle into a dynamic, self-correcting reasoning engine.

This vision informs several concrete research avenues:

- **From Cognitive to Physical Heuristics:** A truly general intelligence must be grounded in physics. We propose augmenting the architecture with a **"Physical Plausibility" (Φ)** heuristic, driven not by an LLM but by a differentiable physics engine. This would allow the agent to reason about novel physical problems from first principles, grounding its "thoughts" in the laws of the universe.

- **From Reactive to Proactive Discovery:** The current failure-driven discovery model is limited. We propose implementing a **"Cognitive Play" or "Daydreaming" mode** where the agent, during computational downtime, is intrinsically motivated to run "heuristic experiments" on itself, creating bizarre hypotheticals or novel CRS configurations to explore the boundaries of its own cognitive model. This could lead to more organic, serendipitous discoveries.
- **From a Single Agent to a "Society of Minds":** The "single omniscient model" approach may not be the most effective path. We propose reframing the ultimate goal as creating a **FERE-CRS Society of Minds**. This would involve a network of specialized agents who collaborate, delegate tasks based on their learned cognitive stances, and publish their discoveries for the entire network to test and integrate, mirroring the distributed and collaborative nature of human scientific progress.

6. Conclusion

The pursuit of artificial general intelligence requires a paradigm shift from creating systems that are merely knowledgeable to those that are conceptually adaptive. This paper has presented a complete computational and theoretical account of an agent that achieves a crucial milestone on this path: autonomous cognitive self-expansion. We have demonstrated a principled, Active Inference-grounded pathway from basic, efficient reasoning to the autonomous discovery of new conceptual primitives.

Our final and most significant contribution—the demonstration of a "cognitive scientist" agent that can reason about its own failures to expand its conceptual model of the world—offers a tangible solution to the "fixed repertoire" problem that limits so many of today's systems. While the challenges of scaling and alignment are profound, this work provides a computationally explicit roadmap and an empirical proof-of-concept. It suggests that the quest for artificial intelligence may be best framed not as the construction of a single, omniscient model, but as the creation of an architecture that can begin the open-ended, self-directed process of discovery itself.

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 (v1.0.0), 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.1678379**.

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.

7. References

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

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (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).

[3] Bommasani, R., et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258.

[4] Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

[5] Chiang, D., et al. (2023). Can Large Language Models Be an Alternative to Human Evaluations? arXiv preprint arXiv:2305.01937.

[6] Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1481), 933-942.

[7] Cox, M. T., & Raja, A. (2011). Metareasoning: Thinking About Thinking. MIT Press.

[8] Diamond, A. (2013). Executive functions. Annual Review of Psychology, 64, 135-168.

[9] Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on Psychological Science, 8(3), 223-241.

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

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

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

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

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

[15] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[16] 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).

[17] Hofstadter, D. R. (1985). Metamagical Themas: Questing for the Essence of Mind and Pattern. Basic Books.

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

[19] Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. Science, 350(6266), 1332-1338.

[20] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

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

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

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

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

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

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

[27] Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active inference, homeostatic regulation, and adaptive behavioural control. Progress in Neurobiology, 134, 17-35.

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

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

[30] Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. Developmental Science, 10(1), 89-96.

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

[32] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. Science, 331(6022), 1279–1285.

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

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

[35] Zheng, L., et al. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.