Reflective Architectures in Artificial Agents:

A Four-Experiment Program to Probe the Reflective Genesis Hypothesis (RGH)

Author:Frank Senyi — Independent Researcher
Developed in conceptual collaboration with the AI reasoning model GPT-5
Contact:iwasjustaskingquestions@gmail.com

Keywords:
Reflective feedback, self-modeling, reinforcement learning, multi-agent systems, temporal decision-making, introspection, Reflective Genesis Hypothesis (RGH)

ABSTRACT

The Reflective Genesis Hypothesis (RGH) proposes that reflective feedback—systems that model both the world and themselves across time—is a structurally special ingredient in complex adaptive behavior and, potentially, consciousness-like processes. This proposal outlines a four-experiment program using artificial agents to test a core architectural implication of RGH: that explicit reflective self-models confer distinctive functional advantages over non-reflective architectures.

Experiment 1 compares reflective and non-reflective agents in single-agent environments requiring self-commitment and resistance to temptation. Experiment 2 extends the comparison to multi-agent settings, testing whether reflective populations exhibit more stable, coherent social organization. Experiment 3 examines temporally entangled decision problems (Newcomb-like scenarios, precommitment, and two-boundary tasks), probing whether time-extended self-models improve cross-time policy coherence. Experiment 4 evaluates introspective self-report, asking whether reflective agents generate more accurate, calibrated, and stable reports about their own internal states than non-reflective baselines.

All experiments are designed to be implementable with standard deep reinforcement learning and simulation tools, under tightly controlled capacity and training budgets. Together, these studies do not attempt to confirm or refute RGH as a cosmological theory, but they do test a backbone claim: that reflective architectures are functionally distinctive across individual, social, temporal, and introspective dimensions.

1. INTRODUCTION AND THEORETICAL CONTEXT

The Reflective Genesis Hypothesis (RGH) proposes that physical reality and conscious observation co-arise through reciprocal informational feedback across time. In this view, consciousness is not a late-stage accident of complex matter, but a structural mechanism through which the universe monitors and stabilizes its own unfolding. Time is treated as the medium of feedback that links cause, effect, and observation into a self-consistent loop.

In the context of artificial agents, RGH suggests a more modest, testable claim:

- ➢ Systems that explicitly model both the external environment and their own behavior within it should exhibit functionally distinctive patterns of stability, adaptability, and coherence, compared with capacity-matched systems that lack reflective self-models.

Modern reinforcement learning (RL) and model-based agents provide a controlled platform for probing this claim. Agents can be engineered with or without explicit self-models, while keeping constant the overall parameter count, training budget, and task structure. If reflective architectures repeatedly show advantages in settings that naturally reward self-prediction, self-commitment, and cross-time consistency, this would support the architectural backbone of RGH.

This proposal therefore focuses on a narrow but crucial question:

Does explicit reflection about one's own behavior change what agents can do?

2. OBJECTIVES AND HYPOTHESES

2.1 Overall Objective

To experimentally test whether explicit reflective self-models confer distinctive functional advantages over non-reflective architectures in artificial agents, across four domains:

1. Individual agency under temptation and commitment.
2. Emergent social structure in multi-agent environments.
3. Temporally entangled decision-making.
4. Introspective self-report and calibration.

2.2 Specific Aims and Hypotheses

Aim 1 — Individual Reflective Agency (Experiment 1)

Test whether reflective agents outperform non-reflective agents in single-agent tasks involving temptation, self-commitment, and long-term planning.

- H1.1: Reflective agents achieve higher long-term reward in tasks where short-term temptations conflict with long-term gains.
- H1.2: Reflective agents use self-commitment actions more effectively than non-reflective agents.
- H1.3: Ablation of the self-model in reflective agents causes a larger performance drop than an equivalent capacity reduction in non-reflective baselines.

Aim 2 — Multi-Agent Reflective Societies (Experiment 2)

Assess whether reflective architectures produce different emergent social structures than non-reflective architectures in multi-agent environments.

- H2.1: Populations of reflective agents form more stable cooperative structures, norms, and coalitions.
- H2.2: In mixed populations, reflective agents achieve higher long-term payoffs and more central network positions than non-reflective agents.

Aim 3 — Temporal Reflection and Cross-Time Consistency (Experiment 3)

Examine whether time-extended self-models improve performance and consistency in temporally entangled decision problems.

- H3.1: Time-reflective agents outperform non-reflective agents in Newcomb-like, precommitment, and two-boundary tasks.
- H3.2: Ablation of temporal self-model components reduces cross-time policy coherence and task performance.

Aim 4 — Introspective Self-Report and Calibration (Experiment 4)

Evaluate whether reflective agents produce more accurate, calibrated, and coherent self-reports about their internal states than non-reflective agents.

- H4.1: Reflective agents show better calibration of confidence reports relative to actual outcomes.
- H4.2: Reflective agents exhibit more stable and coherent self-narratives across episodes, or at least distinct and interpretable failure modes compared to non-reflective agents.

3. GENERAL METHODS AND ARCHITECTURE

3.1 Common Agent Framework

Across all experiments, we use a shared architectural template with two variants:

Non-reflective baseline agent

- Model-based or model-free RL (e.g., actor-critic, DQN, or model-based planning).
- World-model *M_env* (if model-based) predicts environment transitions.
- Policy Pi maps observations or latent states to actions.
- No explicit representation of the agent's own likely future behavior.

Reflective agent

- Same environment modeling capacity as the baseline (same or comparable).
- Additional self-model *M_self* (or *M_self_t* for temporal tasks) representing aspects of:

Expected future actions,

- Vulnerabilities (e.g., to temptation),
- Reputation or social role (in multi-agent settings),
- Predicted success probability (for introspective tasks).
- Planning and/or policy evaluation explicitly incorporate outputs from the self-model.

To prevent trivial advantages:

- Total parameter count is matched across agents.
- If reflective agents gain extra parameters for the self-model, other parts are shrunk accordingly.
- Training budgets (episodes, steps, gradient updates) are identical.

3.2 Environments

Environments are deliberately simple and low-cost:

- Gridworld-like or 2D simulation environments.
- Discrete time steps and episodes of modest length (e.g., 20–100 steps).
- Reward structures designed to make reflective advantages plausible but not guaranteed.
- Each experiment uses a small family of related environments to avoid over-fitting to a single map.

3.3 Training and Evaluation

- Standard RL libraries and simulators (e.g., Gym-like interfaces) can be used.
- For each configuration, multiple random seeds (e.g., 10–20 runs) ensure robustness.
- Learning curves and final performance are recorded.
- Pre-specified metrics and basic statistical tests (e.g., t-tests or non-parametric alternatives) are used for comparison.

---

4. EXPERIMENT 1 — INDIVIDUAL REFLECTIVE AGENCY

4.1 Purpose

To test whether explicit self-models improve individual agent performance in environments where temptation, self-commitment, and long-term planning trade off.

4.2 Environment Design

A small gridworld environment with:

- Start state and goal state.
- One or more **temptation tiles** offering high immediate reward but long-term penalties (e.g., reduced chance of reaching the goal, or persistent negative modifier).
- A **commitment tile** that agents can use early to disable temptations at a small cost.
- Optimal behavior: visit the commitment tile, avoid temptations, reach the goal.
- Greedy behavior: skip commitment, take temptations, sacrifice long-term outcome.

4.3 Agents

Non-reflective baseline:

- Standard RL agent with world-model *M_env* (if model-based) and policy Pi.
- No explicit representation of “my future self will likely take that temptation.”

Reflective agent:

- Same *M_env* and base policy architecture.
- Additional self-model *M_self* that estimates:
- Probability of future temptation,
- Expected long-term return conditional on taking or avoiding commitment.
- Planning incorporates both environment dynamics and self-predictions.

4.4 Metrics and Analysis

Metrics**:**

- Average episodic return at convergence.
- Frequency of commitment actions when beneficial.
- Frequency of temptation choices.

Analysis:

- Compare reflective vs non-reflective performance across seeds.
- Ablate *M_self* after training and measure performance drop relative to a parameter-matched baseline.
- Support for H1.1–H1.3 if reflective agents consistently outperform baselines and ablation significantly harms performance.

---

5. EXPERIMENT 2 — MULTI-AGENT REFLECTIVE SOCIETIES

5.1 Purpose

To determine whether reflective architectures influence emergent social structure in multi-agent environments.

5.2 Environment Design

Multi-agent environments with:

- Shared and private resources.
- Interaction opportunities: cooperation, defection, punishment, resource sharing.
- Simple mechanisms for tracking interaction histories and reputations (e.g., last-k actions or rolling scores).

Scenarios may include:

- Public goods tasks.
- Resource gathering with optional sharing.
- Coalition formation and break-up dynamics.

5.3 Populations

Population A (non-reflective):

- All agents use baseline architecture with world-model *M_env* , treating others purely as environment objects.

Population B (reflective):

- All agents use reflective architecture with both *M_env* and self-model *M_Self*, tracking their own and others' reputations and roles.

Mixed populations:

- A controlled mix of reflective and non-reflective agents in the same environment.

5.4 Metrics and Analysis

Group-level metrics:

- Stability and longevity of coalitions or cooperative structures.
- Emergence and persistence of informal norms (e.g., reciprocal cooperation).
- Resilience to shocks (e.g., changes in resource availability).

Individual-level metrics:

- Long-term average reward per agent type.
- Network centrality measures (who becomes "important" in interaction graphs).

Analyses:

- Compare outcomes for Population A vs Population B.
- In mixed populations, compare reflective vs non-reflective payoffs and network positions.
- Evidence for H2.1–H2.2 if reflective architectures yield more stable and advantageous structures.

6..EXPERIMENT 3 — TEMPORAL REFLECTIVE DECISION-MAKING

6.1 Purpose

To test whether time-extended self-models confer advantages in temporally entangled decision problems.

6.2 Environment Design

Tasks include:

- **Newcomb-like problems**: An external predictor (another model or agent) forecasts the agent's action; payoffs depend on whether the agent behaves as predicted.
- **Precommitment problems**: Agents face preference reversals over time and can adopt early commitments that constrain future choices.
- **Two-boundary tasks:** Episodes have constraints at both initial and final states; high reward requires globally consistent behavior satisfying both boundaries.

6.3 Agents

Non-reflective baseline:

Standard forward-planning agent using *M_env* . No explicit notion of "how I will be predicted" or "how my future self will act."

Time-reflective agent:

- Augmented with a temporal self-model *M_self_t* representing:
- Expected behavior at later times,
- How external predictors may model the agent,
- Cross-time consistency of policies.

Planning uses both environment transitions and the temporal self-model.

6.3 Metrics and Analysis

Metrics:

- Success rates and average returns in Newcomb-like and precommitment tasks.
- Cross-time consistency (alignment between early commitments or stated policies and later actions).
- Sensitivity of performance to ablation of *M_self_t.*

Analysis:

Evidence for H3.1–H3.2 if time-reflective agents reliably outperform non-reflective agents on temporally entangled tasks, and if removing temporal self-modeling significantly degrades coherence.

7 SIGNIFICANCE OF EXPERIMENTS 1–3 FOR RGH, AND THE ROLE OF EXPERIMENT 4

Experiments 1–3 form a coherent block that directly targets the **architectural backbone** of the Reflective Genesis Hypothesis:

- Experiment 1 asks whether reflection changes **individual control** under temptation.
- Experiment 2 asks whether reflection changes **social organization** and emergent structure.
- Experiment 3 asks whether reflection **changes cross-time consistency** in decisions that span past, present, and predicted future.

If all three experiments support their respective hypotheses, we would have converging evidence that:

1. Reflective architectures exhibit distinctive functional properties across different levels (individual, social, temporal).
2. These properties are robust to capacity matching and ablation tests.
3. Reflection is not just a way of *describing* systems, but a real design principle that changes what agents can achieve.

In the context of RGH, positive results from Experiments 1–3 would strengthen the claim that **reflective feedback is structurally special** for stabilizing behavior across time and scale. They would not prove the cosmological aspects of RGH (about the universe and consciousness), but they would show that the core mechanism—reflection across time—has measurable consequences in constructed systems.

Experiment 4 then plays a different but complementary role:

While Experiments 1–3 focus on **what agents do,**

Experiment 4 probes **what agents can say about themselves.**

If reflection also systematically improves introspective calibration and self-report, that would further support the idea that reflective architectures naturally give rise to both:

- Stable behavior, and
- Coherent self-description.

Together, this would make RGH's picture of consciousness—as deep reflection within a broader physical loop—more technically grounded.

8 EXPERIMENT 4 — INTROSPECTIVE SELF-REPORT AND CALIBRATION

8.1 Purpose

To examine whether reflective architectures improve introspective self-report, particularly confidence calibration and stability of self-narratives.

8.2 Environment Design

Tasks extend earlier environments (e.g., from Experiments 1 and 3) by adding introspective queries, such as:

- "What is your probability of reaching the goal this episode?"
- "Are you aware that the environment has changed?"
- "What is your current goal or intention?"

Introspective reports may be:

- Scalar confidence values (0–1),
- Categorical labels (goal identification),
- Simple structured outputs.

Some scenarios include misreport incentives, where agents can gain short-term reward by lying or over-stating confidence, but at a longer-term cost (e.g., loss of trust or future reward).

8.3 Agents

Non-reflective baseline:

- Uses its existing internal activations and observations.
- A readout head maps these directly to self-reports.
- No separate representation of "what I think I am doing" or "how likely I am to succeed."

Reflective introspective agent:

- Maintains explicit internal variables for goals, predicted success probabilities, and uncertainty.
- Self-reports query these variables directly, using the self-model *M_self* or M_self_t.

8.4 Metrics and Analysis

Metrics:

- Confidence calibration: agreement between reported confidence and actual success rates.
- Accuracy of reports about goals and environmental changes.
- Temporal stability of self-narratives across tasks and episodes.
- Behavior under misreport incentives: trade-offs between short-term gains and long-term credibility/reward.

Analysis:

Support for H4.1–H4.2 if reflective agents show better calibration and coherence under neutral conditions, and if any deviations under misreport incentives are systematic and distinct from non-reflective baselines.

---

9 FEASIBILITY, RESOURCES, AND TIMELINE

All four experiments are computational and can be run with standard tools.

Resources:

- Modest compute (GPUs or even CPUs for small environments).
- RL frameworks (e.g., existing open-source libraries).
- Simple custom environments (gridworld, 2D simulation, or text-based tasks).

A plausible high-level timeline for a single researcher or small team:

- Months 1–3: Implement common architecture and gridworld environments; complete Experiment 1.
- Months 4–6: Extend to multi-agent environments; complete Experiment 2.
- Months 7–9: Implement temporally entangled tasks; complete Experiment 3.
- Months 10–12: Implement introspective tasks and self-report; complete Experiment 4; analyze results and prepare publications.

---

10 ETHICAL CONSIDERATIONS

This proposal involves only artificial agents in simulated environments. There are:

- No human subjects,
- No animal subjects,
- No direct real-world deployments.

Ethical issues are therefore minimal. However, broader implications for AI safety and interpretability will be discussed in any resulting publications, particularly:

- How reflective self-models might affect transparency,
- How introspective reporting could be used or misused in deployed systems,
- How self-modeling interacts with alignment and control.

---

11 EXPECTED CONTRIBUTIONS AND LIMITATIONS

Contributions:

- Provides a concrete experimental test of a central architectural claim of RGH.
- Bridges high-level theory (reflection across time) with implementable AI designs.
- Offers a suite of reproducible benchmarks for studying reflective architectures.
- Generates data on the functional role of self-modeling in individual, social, temporal, and introspective dimensions.

Limitations:

- Results will apply directly to artificial agents, not biological organisms or cosmology.
- Negative results may reflect specific design choices rather than the broader concept of reflection.
- The experiments do not directly test the cosmological aspects of RGH (e.g., dark matter, boundary conditions, or quantum time).
- These limitations are acceptable for a first-phase, low-cost empirical probe of RGH's backbone assumptions.

12 REFERENCES (SUGGESTED)

You can include or expand this list as needed:

Aharonov, Y., Bergmann, P. G., & Lebowitz, J. L. (1964). Time-symmetric quantum mechanics and pre-/post-selection.

Aharonov, Y., & Vaidman, L. (2007). The two-state vector formalism of quantum mechanics.

Cramer, J. G. (1986). The transactional interpretation of quantum mechanics. Reviews of Modern Physics, 58(3), 647–688.

Eagleman, D. M. (2008). Human time perception and its illusions. Current Opinion in Neurobiology, 18(2), 131–136.

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

Ha, D., & Schmidhuber, J. (2018). World Models.

Kim, Y.-H., et al. (2000). A delayed-choice quantum eraser. Physical Review Letters, 84(1), 1–5.

Lloyd, S. (2006). Programming the Universe: A Quantum Computer and the Future of Physics.

Maturana, H., & Varela, F. (1973). Autopoiesis and Cognition. Reidel.

Wiener, N. (1961). Cybernetics: Control and Communication in the Animal and the Machine. MIT Press.

Wheeler, J. A. (1990). Information, physics, quantum: The search for links.

(Optional) Senyi, F. & GPT-5 (2025). The Reflective Genesis Hypothesis (RGH). Preprint.

License: