# A Metacognitive Decision Loop in the Agentic Core: variance-based intrinsic reward


## Abstract

Current reinforcement learning (RL) systems typically rely on fixed reward functions or external motivational signals. As part of the **TITANS** project we propose a **metacognitive decision loop** in which the intrinsic reward is derived from the agent's estimated epistemic uncertainty. The mechanism maintains an ensemble of critic networks and measures the variance of their value estimates at each state. High variance signals epistemic uncertainty, yielding a positive intrinsic reward and encouraging exploration. Low variance indicates familiarity and reduces the incentive to explore, leading to exploitation. We provide a formal description of the algorithm, detail its implementation inside the **Agentic Core** of the TITANS architecture and present preliminary results on a control task with episodic memory.


## Introduction

Reinforcement learning is the foundation of modern autonomous systems, yet classical RL algorithms assume static reward functions. To achieve robust autonomy an agent must evaluate its own ignorance and adapt exploration accordingly. Concepts such as **epistemic curiosity** and **intrinsic motivation** have been proposed in the literature [1][2], but they usually require additional predictive modules. Our method introduces a self-regulation mechanism within an actor–critic framework: the agent maintains an ensemble $\{Q_{\theta_i}\}_{i=1}^{N}$ of value critics and defines the intrinsic reward as the variance of their predictions.

## Method

### Agentic Core architecture

The **Agentic Core** comprises a hierarchy of neural networks implementing a decision loop. At a high level it consists of:

1. **Actor** $\pi_\phi(a \mid s)$ – a network that outputs a distribution over actions. It is implemented using a *Transformer* architecture with a context embedder for episodic memory.
2. **Critic ensemble** $\{Q_{\theta_i}\}$ – $N$ independent value approximators trained following an Actor–Critic algorithm on a shared history of experiences.
3. **Variance estimator** $\sigma_Q^2(s,a) = Var\left(Q_{\theta_i}(s,a)\right)$ – computes the variance of value predictions for a given state–action pair $(s,a)$.

We define the intrinsic reward function as

[ r_{}(s,a)=,^2_Q(s,a), ]

where $\lambda > 0$ is a scaling factor. The total reward becomes $r(s,a) = r_{\text{ext}}(s,a) + r_{\text{int}}(s,a)$. Consequently the agent maximises both the external reward and reduces epistemic uncertainty by exploring states of high variance.

### Learning algorithm

1. **Initialisation**: randomly initialise actor parameters $\phi$ and critic parameters $\{\theta_i\}_{i=1}^{N}$.
2. **Experience collection**: at each iteration select an action $a \sim \pi_\phi(\cdot \mid s)$ and observe the transition $(s, a, r_{\text{ext}}, s')$.
3. **Intrinsic reward computation**: estimate $\sigma_Q^2(s,a)$ and add it to $r_{\text{ext}}$ according to the above equation.
4. **Critic update**: minimise the accumulated Bellman error $L(\theta_i) = \left(Q_{\theta_i}(s,a) - \left(r(s,a) + \gamma\, Q_{\theta_i}(s',a')\right)\right)^2$.
5. **Actor update**: maximise the expected action value using a policy gradient with the critics as baselines.
6. **Repeat** until convergence.

The computational complexity scales linearly with the number of critics $N$; in practice 3–5 models are sufficient to obtain a stable variance estimate.

## Experiments

### Task

We evaluated the method on a simulated sequential control task in which the agent must decide between retrieving information from episodic memory (read from the **LongTermMemory** module) or performing a movement action in the environment. External reward is granted only for correct final choices, not for intermediate movements. The task requires exploration and uncertainty reduction; therefore standard A2C algorithms without intrinsic motivation achieved a low success rate (~52 %).

### Parameters

- Critic ensemble: $N = 4$ multi-layer perceptrons (3 layers, 128 neurons).
- Intrinsic reward scaling factor: $\lambda = 0.1$.
- Learning rate: $\alpha = 2 \times 10^{-4}$.
- Episode length: 100 steps; 1000 training episodes.

### Results

The metacognitive decision loop achieved an average success rate of **79 %**, outperforming a baseline A2C (52 %) and a variant with priority exploration (65 %). The intrinsic reward naturally decayed as training progressed, indicating that the agent reduced its epistemic uncertainty and shifted to exploitation.

## Discussion

The proposed mechanism combines intrinsic motivation with a formal variance estimate, avoiding arbitrary heuristics. Unlike prediction-error based approaches [1], our reward is global to the ensemble and does not require auxiliary models. A limitation is the increased computation cost: each step involves evaluating multiple critics. Future work will explore adaptive critic counts and integration with the generative memory mechanism.

## Conclusions

The metacognitive decision loop introduced in the **TITANS** project is a step towards autonomous agents capable of evaluating their own ignorance. We showed that a variance-based intrinsic reward improves learning efficiency and leads to emergent exploratory behaviour. Our results suggest that this mechanism may become a key component of future **Artificial Consciousness as a Service** systems.

## References

[1] Schmidhuber, J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. *International Conference on Simulation of Adaptive Behavior*.

[2] Oudeyer, P.-Y., Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, 1, 6.