Viability-Driven Representation Learning: Perception and Action Emerge from State-Constrained Dynamics

Anonymous Authors[1]

Abstract

We introduce a training paradigm fundamentally different from reinforcement learning: **viability-driven representation learning**. Rather than optimizing reward or minimizing prediction error, we constrain internal state dynamics to remain in a **bounded, uncertainty-dependent non-equilibrium regime**. The system is penalized if its internal state becomes static (no adaptation), chaotic (uncontrolled drift), or irrelevant (not shaped by action or observation). Under these constraints, perception and action **emerge as mechanisms for maintaining representational viability**—not as behaviors rewarded by an external signal. We formalize this through three loss terms: (1) a **viability band** preventing equilibrium collapse and chaotic explosion, (2) **world coupling** grounding internal dynamics in external reality, and (3) **counterfactual responsibility** ensuring state change is agent-caused. In experiments on perception and survival tasks, agents trained under viability constraints develop anticipatory behavior, adaptive exploration, and robustness to regime shifts—**without any reward function**. We show that exploration is a mathematical necessity and action emerges because movement is required to keep internal dynamics alive. This reframes learning as homeostatic self-regulation, positioning representational viability as a sufficient condition for the emergence of intelligent behavior.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

1.1. The Problem with Reward-Based Learning

Reinforcement learning (RL) trains agents to maximize expected cumulative reward. This paradigm has achieved remarkable success, scaling from game playing (**??**) to robotic control (**??**). However, it carries a fundamental epistemological assumption: that intelligent behavior requires an external signal specifying what is valuable. Without reward, there is no learning; without shaping, there is no structure. This dependency creates brittleness in environments where rewards are sparse, misspecified, or absent, and it stands in stark contrast to biological systems.

Biological organisms do not receive scalar reward signals from their environment. Evolution selects for survival, but the individual organism must survive in real-time by maintaining efficient internal homeostasis constraint satisfaction—keeping temperature, energy, and representational coherence within viable bounds (Ashby, 1960; **?**). In nature, behavior emerges not because it is explicitly rewarded, but because it is *required* for continued existence. An organism that fails to perceive threats or locate energy sources ceases to exist; thus, the survivors are those whose internal dynamics successfully coupled with their environment to maintain viability.

1.2. The Viability-Driven Paradigm

We propose a different paradigm for artificial intelligence: **viability-driven representation learning**. This is a training regime where learning emerges strictly from constraints on internal state dynamics, not from reward optimization. The core idea is that learning happens because the system is forced to keep its internal state alive, responsive, and bounded under changing inputs.

The model is penalized if its internal state:

- Becomes **static**, leading to the "Dark Room" problem where agents minimize surprise by finding a corner of zero stimulation.
- Becomes **chaotic**, where internal states drift unboundedly, losing all correlation with the external world.

- Becomes **decoupled**, where the state evolves independently of action or observation, rendering the agent a passive spectator.

Under these constraints, the system *must* learn to perceive and act effectively. There is no reward to maximize; there is only viability to maintain. The agent is not a maximizer of utility but a regulator of its own internal equilibrium.

1.3. Emergent Properties

When we train systems under these viability constraints, we observe distinct emergent properties that differ directly from standard RL baselines. Exploration becomes a structural requirement rather than a heuristic bonus added to a loss function; passivity becomes structurally impossible rather than merely suboptimal. Anticipation emerges from the pressure to maintain predictive coupling, and adaptation to change follows directly from the dynamical constraints. These behaviors are not rewarded; they are structural consequences of the viability regime.

1.4. Contributions

Our contributions are fourfold:

1. We define a formal training objective based on viability constraints: World Coupling, a Viability Band, and Counterfactual Responsibility.
2. We provide a theoretical analysis showing why passivity and chaos are structurally eliminated in this framework.
3. We empirically demonstrate that perception, exploration, and adaptation emerge without reward in active perception and dynamic survival tasks, outperforming standard RL baselines in regime-shift robustness.
4. We offer a new conceptual framework where the agent is not a controller maximizing reward, but a subsystem of coupled dynamics maintaining representational viability.

2. Related Work

2.1. Reinforcement Learning and Its Limits

Standard Reinforcement Learning (Sutton & Barto, 2018) trains policies to maximize expected cumulative reward. While powerful, this paradigm requires external reward specification and can be sample-inefficient. Moreover, RL agents often converge to passive or degenerate policies when reward is sparse or misspecified. Deep RL has addressed some of these issues with auxiliary tasks (**?**), but the fundamental driver remains reward maximization. Our approach eliminates reward entirely; learning emerges from structural constraints on internal dynamics.

2.2. Curiosity and Intrinsic Motivation

Intrinsic motivation approaches (Schmidhuber, 1991; Pathak et al., 2017; Oudeyer et al., 2007; **?**) have attempted to bridge the gap by reusing the RL framework but substituting or augmenting external rewards with internal signals such as prediction error, novelty, or learning progress. Pathak et al. (Pathak et al., 2017) demonstrated that curiosity (prediction error) can drive exploration in the absence of external rewards. However, these methods remain fundamentally distinct from our approach because they still rely on reward maximization. Intrinsic rewards are still rewards that the agent optimizes. In contrast, our framework has no reward—not even internal reward. The viability band is a **constraint**, not an objective to maximize. This fundamentally alters the optimization landscape from searching for a peak to maintaining a volume.

2.3. Predictive Coding and World Models

Predictive coding (Rao & Ballard, 1999; Friston, 2010) frames perception as prediction error minimization, while world models (Ha & Schmidhuber, 2018; Hafner et al., 2019) learn latent dynamics for planning. These approaches minimize prediction error, which implies that the optimal state is one of zero surprise. This can paradoxically lead to the "Dark Room" problem: agents that reduce error by cutting off sensory input or finding trivial states where nothing happens. Friston addresses this with the concept of "expected free energy," which includes a term for information gain (**?**). Our viability band explicitly prevents the Dark Room scenario by requiring a *minimum* level of internal change, conditioned on uncertainty. We force the agent to be "restless" when it is uncertain, embedding exploration into the very definition of a valid internal state.

2.4. Homeostasis and Viability Theory

Our work draws significant inspiration from cybernetics and viability theory. Ashby's ultrastability (Ashby, 1960) describes systems that change their internal parameters when critical variables cross viable limits. Viability theory (Aubin et al., 2011) formalizes the set of states from which a system can remain indefinitely. Our viability band acts as a differentiable implementation of these concepts. Unlike homeostatic reinforcement learning (**?**), which uses deviation from homeostasis as a negative reward, we use deviation from the viability band as a direct loss on the state dynamics themselves. This distinction is crucial: our agent does not "want" to be homeostatic; it "is" homeostatic by definition of the learning law.

2.5. Dynamical Sytems Perspective

Beer (Beer, 1995) and others in the enactivist tradition (**?**) view agency as a coupling between an organism and its environment. In this view, cognition is not information processing but the regulation of this coupling. Our Counterfactual Responsibility term (L_{cf}) provides a computational mechanism for what this coupling looks like: the agent's internal state trajectory must be causally dependent on its own efferent copies (actions). This aligns with the notion of "sensorimotor contingencies" (**?**), where perception is constituted by the mastery of the laws governing sensory changes induced by action.

3. System Formulation

We do not treat the agent as separate from the environment. Instead, we model a **joint dynamical system**. Let x_t be the observation (e.g., image patch, sensor input), a_t be the action (movement, control signal), and z_t be the internal state (belief, memory, latent dynamics). The internal state evolves according to a learned recurrent update f_θ (e.g., LSTM, GRU):

$$z_{t+1} = f_\theta(z_t, x_t, a_t) \tag{1}$$

The action is produced by a policy head $a_t = \pi_\phi(z_t)$. Critically, there is no reward function r_t. The system is trained purely through constraints on the dynamics of z_t.

We measure the state motion as $\Delta z_t = z_{t+1} - z_t$. The viability regime regulates how much motion is allowed. We also define an energy signal $E_t \in [0, 1]$ that measures the system's epistemic state. In perception tasks, this corresponds to classification confidence; in world models, prediction accuracy; and in control, stability variables. Low energy indicates uncertainty or instability, while high energy indicates confidence. Critically, **energy modulates the lower bound**: lower confidence forces larger internal state change, ensuring exploration under uncertainty. This creates a direct coupling between epistemic state and dynamical requirements.

4. The Viability Loss

Our training objective consists of three terms, none of which is a reward. These terms represent the physical laws of the internal world: the constraints that the system must satisfy to exist.

4.1. Term 1: Viability Band

The core innovation is the **viability band constraint**. In most learning systems, the goal is convergence: finding a stable point where the loss is minimized. In a living system, total convergence is death (equilibrium). To prevent this, we enforce that the internal state change $\|\Delta z_t\|$ must stay inside a dynamic band:

$$\ell_t \leq \|\Delta z_t\| \leq u \tag{2}$$

4.1.1. The Lower Bound ℓ_t (Anti-Equilibrium)

The lower bound is not static; it is uncertainty-dependent.

$$\ell_t = \ell_{base} + \alpha(1 - E_t) \tag{3}$$

where ℓ_{base} is a minimal metabolic rate and α is a gain parameter.

This creates a fundamental homeostatic pressure:

- **When Uncertain ($E_t \to 0$):** The lower bound ℓ_t rises. The system is *forced* to move its internal state significantly. Passive observation is no longer a valid solution. The only way to satisfy ℓ_t is to take actions that produce new, distinct observations.
- **When Confident ($E_t \to 1$):** The lower bound ℓ_t relaxes to ℓ_{base}. The system is allowed to settle into a stable attractor, provided it maintains a minimal level of alertness.

This mechanism solves the "Dark Room" problem structurally. An agent in a dark room has low confidence (high uncertainty about the world), so its ℓ_t will be high. But since the room is dark, passive observation yields $\Delta z \approx 0 < \ell_t$. The violation of the lower bound creates a massive gradient signal forcing the agent to move, open its eyes, or leave the room.

4.1.2. The Upper Bound u (Anti-Chaos)

The upper bound $u =$ constant acts as an anti-chaos constraint. It prevents the system from maximizing change by simply generating random noise or seizures. The system must find a path of "controlled change"—enough to satisfy the lower bound, but structured enough to stay below the upper bound.

The viability loss is the penalty for violating these bounds:

$$L_{band} = \begin{cases} (\|\Delta z_t\| - u)^2 & \text{if } \|\Delta z_t\| > u \\ (\ell_t - \|\Delta z_t\|)^2 & \text{if } \|\Delta z_t\| < \ell_t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Critically, any state within the band has $L_{band} = 0$. The system is agnostic to *which* valid state it occupies, as long as it is viable. This leaves the manifold of viable states open for task-specific optimization or further exploration.

4.2. Term 2: World Coupling

The internal state must represent external reality, not drift into fantasy. The Viability Band alone could be satisfied by a system hallucinating internal dynamics. To prevent this, we anchor the internal state to the world.

For perception tasks, we enforce this via classification success (L_{task}). For world models, we use prediction accuracy:

$$L_{world} = \|\hat{x}_{t+1} - x_{t+1}\|^2 \tag{5}$$

This term ensures that "movement" in z-space corresponds to "movement" in observation space. It prevents **meaningless internal motion**—the system cannot satisfy the lower bound ℓ_t by generating random noise in z; the change must correspond to real, predicted perceptual structure.

4.3. Term 3: Counterfactual Responsibility

State change should be caused by the agent's action, not passive drift. We compute the counterfactual difference:

$$L_{cf} = \|f(z_t, x_t, a_t) - f(z_t, x_t, \mathbf{0})\|^2 \tag{6}$$

This measures how much the action a_t *matters* for the state update. If $L_{cf} \approx 0$, it means the action had no effect on the future state (the agent is a passive observer). High L_{cf} means the agent is actively shaping its own dynamics.

In practice, we **maximize** L_{cf} as part of the objective (using a negative sign in the total loss). This integrates naturally with the viability band:

- To satisfy the lower bound ℓ_t, the system must change its internal state.
- To satisfy L_{cf}, this change must be attributable to action.

Together, they force genuinely **agent-dependent dynamics**. The system learns that "doing something" is the only way to satisfy its existence conditions.

4.4. Final Objective

The final objective combines these terms:

$$L = \lambda_{task}L_{task} + \lambda_{world}L_{world} + \lambda_{band}L_{band} - \lambda_{cf}L_{cf} \tag{7}$$

There are no value functions, no policy gradients, and no scalar rewards. Actions are learned because movement is required to keep internal dynamics viable. The gradients flow through the differentiable world model and policy, shaping the system's attractor landscape to avoid the "death" of stasis or chaos.

5. Why This Works: Theoretical Analysis

In RL, an agent can converge to passivity if reward is zero. Nothing penalizes "doing nothing." Under viability constraints, passivity is structurally impossible because it violates the lower bound constraint: $\|\Delta z_t\| < \ell_t \implies L_{band} > 0$. The system is penalized for not changing.

Furthermore, exploration is a mathematical necessity. When energy E_t is low (indicating uncertainty), the lower bound ℓ_t increases. The system *must* change its internal state more when uncertain. This forces exploration automatically—not because exploration is rewarded, but because stasis under uncertainty violates viability. Conversely, the upper bound u prevents unbounded state change, ensuring that the system remains in a controlled non-equilibrium rather than exploding into chaos.

Without world coupling, the system could satisfy viability by generating random internal noise. L_{world} constraints force the state change to correspond to real structure in the environment. Combined with counterfactual responsibility, the system learns that action-induced perceptual change is the only way to satisfy all constraints simultaneously.

5.1. The Causal Chain of Structural Adaptation

The following chain explains how viability constraints produce adaptive behavior:

Environment Change → Prediction Error →
Energy E_t Drops → Lower Bound ℓ_t Increases
↓
L_{band} Penalizes Stasis → Forced State Change
$\Delta z_t \neq 0$ → Action Required (via L_{cf})
↓
Exploration → New Information → Energy
Restored → Viability Satisfied

This is structural adaptation: the system's dynamics *require* it to respond to environmental change. No reward is needed—the constraints alone generate intelligent behavior.

6. Experiments

We validate the viability-driven framework in three domains designed to probe different aspects of agency: active perception (information gathering), survival under regime shift (robustness), and dynamic occlusion (anticipation).

6.1. Experiment 1: Active Perception (Cluttered MNIST)

6.1.1. Setup

In the Active Perception task, a digit is placed on a cluttered 64×64 canvas. The agent observes only a 5×5 foveated

patch and must classify the digit. This is a classic "needle in a haystack" problem. The agent controls gaze position (p_x, p_y) and zoom scale s. We use classification confidence as the energy signal E_t.

6.1.2. Comparison to Baselines

A standard RL agent trained with PPO receives a reward of +1 for correct classification. A Random agent moves its fovea uniformly.

[Figure: Active Perception Trajectories. PPO agents converge to the center. Viability agents scan corners and high-contrast regions until the digit is found.]

Figure 1. Trajectory analysis of active perception. Viability agents exhibit covering behavior.

6.1.3. Results and Analysis

Condition	Behavior	Viability Status
Full System	Active search → Fixation	✓ Viable
No Lower Bound	Frozen (Dark Room)	× Collapsed
No World Coupling	Random motion	× Decoupled
No Upper Bound	Chaotic jitter	× Unstable

With all components present, the agent develops a distinct **three-phase cognitive cycle**:

1. **Search (High ℓ_t):** Initially, confidence is low, so the lower bound ℓ_t is high. The agent *must* move to satisfy the viability constraint. It executes wide saccades, effectively searching the space.

2. **Recognition (Transition):** The fovea lands on the digit. The classifier output spikes, increasing E_t.

3. **Fixation (Low ℓ_t):** As $E_t \to 1$, the lower bound ℓ_t drops to ℓ_{base}. The agent is now "allowed" to stop moving. It locks onto the target to maintain the high-confidence state.

This cycle is not hard-coded; it emerges purely from the dynamics. The "desire" to find the digit is actually a "requirement" to lower the metabolic cost of movement.

6.1.4. Qualitative Analysis of Emerging Behaviors

The viability constraints give rise to a rich ethogram of behaviors depending on the parameter regime. We observed several distinct "species" of agents during hyperparameter tuning, which illuminate the mechanics of the viability forcing function.

The "Restless Explorer" (High α, Low ℓ_{base}) When the uncertainty gain α is set too high (> 1.0), the agent effectively panics. Even small drops in confidence trigger massive lower bounds ℓ_t. The agent enters a permanent state of high-velocity saccades, never settling long enough to process information. It effectively "vibrates," achieving high internal state change (Δz) but low semantic coupling (L_{task}). This resembles a seizure state where homeostatic regulation fails due to over-correction.

The "Comatose Observer" (Low α, Low u) If the lower bound is too loose or the upper bound u is too restrictive, the agent finds a "grey zone" where it can drift slowly without engaging with the world. It tends to drift its gaze to a featureless part of the background and execute microscopic, minimal-energy movements that just barely satisfy ℓ_{base}. This is the "Dark Room" attractor. The addition of the uncertainty-dependent term $\alpha(1 - E_t)$ was specifically required to destabilize this attractor.

The "Hallucinating Solipsist" (No L_{world}) In early ablations without strong world coupling (L_{world}), agents discovered they could satisfy ℓ_t by generating internal oscillations in z_t that had no correlation with the external input x_t. The agent would essentially "imagine" a dynamic world to keep itself alive, ignoring the static reality. This confirms that viability alone is insufficient; it must be *grounded* viability.

The Viable Agent (Balanced Regime) The successful agent operates at the "edge of chaos." It moves rapidly when searching (high entropy) and transitions to a low-entropy tracking mode when the target is found. This phase transition is not scripted; it is the mathematical result of E_t rising, which lowers ℓ_t, which opens up a region of low-velocity state space that the agent naturally falls into to minimize energy expenditure (assuming implicit regularization terms or just the natural damping of the GRU).

6.2. Experiment 2: Survival Under Regime Shift

6.2.1. Setup

We compare the viability-driven agent against PPO in a 2D world with a teleporting energy source. The agent consumes energy to survive. E_t corresponds to the internal energy level. The environment undergoes a "regime shift" at $T = 50$, where the energy source teleports to a new, unseen location.

6.2.2. The Failure of Optimization

PPO agents learn a highly optimized policy for the initial source location. They move directly to the source and sit there. When the source teleports, the PPO agent continues to sit at the old location, waiting for a reward that never

comes, until it starves. This is the fragility of optimization: it overfits to the specifics of the reward landscape.

6.2.3. The Robustness of Viability

The viability-driven agent behaves differently. When the source moves, its world model prediction error spikes (or energy drops), causing E_t to plummet.

1. **Crisis:** E_t drops near zero.
2. **Reaction:** The lower bound ℓ_t spikes to its maximum value.
3. **Mobilization:** The agent forces itself to move. Staying still is now a violation of its existence conditions.
4. **Discovery:** This forced exploration leads it to encounter the new source location.
5. **Recovery:** Energy is restored, ℓ_t relaxes, and the agent stabilizes at the new location.

Table 1. Response to regime shift (energy source teleportation). Pre-shift energy values are averaged over the last 100 steps before the shift.

Agent	Pre-Shift	Post-Shift	Outcome
PPO	0.215	0.000	Collapse
Viability	0.034	0.200	Adaptation

This result (Table 1) highlights the fundamental difference: PPO optimizes for a specific state, while Viability optimizes for the *capacity to find states*.

6.3. Experiment 3: Dynamic Occlusion

6.3.1. Setup

In the Dynamic Occlusion task, a moving MNIST digit passes behind occluders. The agent observes through a foveated window. This task requires the agent to maintain an internal representation of the object's position when it is invisible.

6.3.2. Comparison of Baselines

- **Reactive Baseline:** Trained to minimize prediction error of the immediate observation.
- **Task Baseline:** Rewarded for tracking distance.
- **Viability Agent:** Trained with the full loss, including action-conditioned learning progress.

6.3.3. Emergence of Anticipation

Under viability constraints, the agent develops **anticipatory saccades**—moving its gaze toward where the digit *will* emerge, not where it was last seen.

Why does this happen? During occlusion, the prediction error increases because the agent cannot see the digit. This causes E_t (confidence) to drop. The lower bound ℓ_t increases, forcing the agent to change state. If the agent merely reacts, it loses the target. If it anticipates, it regains the target, restoring E_t and allowing it to relax. Thus, **anticipation is the path of least resistance** in the viability landscape.

Reactive baselines fail because they have no structural pressure to maintain dynamics during uncertainty; they simply converge to predicting the background (the "Dark Room").

7. Discussion

7.1. Against the Agent-Centric View

The dominant view in AI is agent-centric: an agent perceives, thinks, and acts on a passive world. Our framework suggests a system-centric view. The agent is not a separate entity maximizing utility; it is a nexus of causal flows that must self-regulate. This coupling (Equation 8 and 10) means the agent and environment form a single dynamical system. Intelligence is the property of this system staying within the viability kernel.

7.2. Implications for AGI Safety

Reward-based agents are prone to reward hacking (**?**): finding loopholes to maximize the score without doing the task. Viability-driven agents are robust to this because they do not maximize a score. There is no number to drive to infinity. There is only a bounded region to stay within. This suggests a new direction for safe AI: defining safety not as a list of constraints on a maximizer, but as the fundamental operating condition of the system itself.

7.3. Biological Plausibility

Our model aligns closely with free energy differentiation and autopoiesis (**?**). Biological cells do not maximize glucose; they regulate glucose to stay alive. By adopting this constraint-based view, we move AI closer to the robustness and flexibility of living systems.

7.4. Biological Connections: Allostasis over Homeostasis

While we use the term "homeostasis," our framework is more accurately described as *allostasis* (**?**)—stability through change. A purely homeostatic thermostat turns on when the temperature deviates. An allostatic agent antic-

ipates the deviation and acts beforehand. Our agent exhibits allostasis because the world model (RNN) propagates the state forward. If the agent predicts a future confidence drop (energy loss), the viability concern becomes active *now*, triggering preemptive exploration. This aligns with Friston's Active Inference (Friston, 2010), but without the need for an explicit "expected free energy" functional. The "expectation" is implicit in the dynamics of the RNN: if the current state trajectory leads to a violation of the viability band in the future, the gradients backpropagate to the present perception/action to correct it.

7.5. Future Work: Hierarchical Viability

The current system operates on a single timeframe. Biological regulation is hierarchical: cells regulate pH, organs regulate blood flow, organisms regulate shelter. We hypothesize that scaling this approach requires a **hierarchy of viability kernels**.

- **Level 1 (Fast):** Direct sensorimotor viability (don't have seizures, keep fovea valid).
- **Level 2 (Medium):** Semantic viability (maintain coherent object representations).
- **Level 3 (Slow):** Strategic viability (maintain long-term energy resources).

Future work will explore nesting these loops, where the "action" of the higher level sets the "viability bounds" (ℓ_t, u) of the lower level.

8. Conclusion

We introduced viability-driven representation learning, a training paradigm where intelligent behavior emerges from constraints on internal state dynamics. The key innovations are the viability band, world coupling, and counterfactual responsibility. Under these constraints, passivity is structurally impossible, exploration is a mathematical necessity, and anticipation/adaptation describe what viability looks like when a system is embedded in a world. This work suggests that the path to general intelligence may not lie in better reward functions, but in better definitions of what it means for a digital system to "survive."

References

Ashby, W. R. *Design for a brain: The origin of adaptive behaviour*. Chapman & Hall, 1960.

Aubin, J.-P., Bayen, A. M., and Saint-Pierre, P. *Viability theory: new directions*. Springer Science & Business Media, 2011.

Beer, R. D. A dynamical systems perspective on agent-environment interaction. *Artificial intelligence*, 72(1-2): 173–215, 1995.

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

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Hafner, D., Lillicrap, T., Ba, J., and Norouzi, M. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2019.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V. V. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11(2): 265–286, 2007.

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, volume 34, 2017.

Rao, R. P. and Ballard, D. H. Predictive coding in the visual cortex. *Nature neuroscience*, 2(1):79–87, 1999.

Schmidhuber, J. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the first international conference on simulation of adaptive behavior*, pp. 222–227, 1991.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.

A. Hyperparameter Details

We provide the full set of hyperparameters used for the Active Perception experiments.

Table 2. Hyperparameters for the Viability-Driven Agent.

PARAMETER	DESCRIPTION	VALUE
	VIABILITY CONSTRAINTS	
ℓ_{base}	BASE LOWER BOUND (RESTING METABOLISM)	0.05
α	UNCERTAINTY GAIN	0.50
u	UPPER BOUND (ANTI-CHAOS)	2.00
	LOSS WEIGHTS	
λ_{band}	VIABILITY BAND WEIGHT	1.0
λ_{world}	WORLD MODEL PREDICTION WEIGHT	1.0
λ_{task}	TASK (CLASSIFICATION) WEIGHT	1.0
λ_{cf}	COUNTERFACTUAL RESPONSIBILITY WEIGHT	0.1
	ARCHITECTURE	
D_z	LATENT STATE DIMENSION	64
H_{rnn}	RNN HIDDEN SIZE	256
N_{cnn}	NUMBER OF CONVOLUTIONAL FILTERS	[32, 64, 64]
	OPTIMIZATION	
lr	LEARNING RATE	1×10^{-4}
B	BATCH SIZE	64
γ	OPTIMIZER (ADAM) BETAS	(0.9, 0.999)

B. Derivation of the Energy-Modulated Bound

Why does the lower bound take the form $\ell_t = \ell_{base} + \alpha(1 - E_t)$? This can be derived from the principle of **Constant Information Rate**. Assume the agent has a channel capacity C for processing new information. The information theoretical surprise of the current state is related to $1 - E_t$. If $E_t \approx 1$, the agent is processing very little new information (surprise is low). To utilize its channel capacity, it must generate *motor information* (movement). If $E_t \approx 0$, the agent is overwhelmed with sensory surprise. To stay within capacity, it should reduce motor noise, but our viability objective inverts this: we require the agent to *resolve* the uncertainty.

We postulate that the "Viability Energy" V is conserved:

$$V = \text{Epistemic Energy} + \text{Kinetic Energy} = \text{constant} \tag{8}$$

If Epistemic Energy (confidence) is low, Kinetic Energy (movement) must be high to balance the equation. This leads to the inverse relationship between confidence and forced movement.

C. Code Implementation

We provide the JAX implementation of the core Viability Loss function to demonstrate the exact calculation of the gradients.

```
def loss_viability(z_next, z_current, E_t, params):
    """
    Computes the viability band loss.

    Args:
       z_next: The next internal state z_{t+1}
       z_current: The current internal state z_t
       E_t: The energy/confidence level [0, 1]
       params: Dictionary containing l_base, alpha, u

    Returns:
```

```
       L_band: Scalar loss
    """
    # Calculate state displacement
    delta_z = z_next - z_current
    dz_norm = jnp.linalg.norm(delta_z, axis=-1)

    # Calculate dynamic lower bound
    # Low energy -> High required movement
    # High energy -> Low required movement (resting)
    l_t = params['l_base'] + params['alpha'] * (1.0 - E_t)

    # Calculate Upper Bound (Anti-Chaos)
    u = params['u']

    # Violation of Lower Bound
    loss_lower = jnp.square(jnp.maximum(0, l_t - dz_norm))

    # Violation of Upper Bound
    loss_upper = jnp.square(jnp.maximum(0, dz_norm - u))

    return loss_lower + loss_upper
```