Internal Control Regimes for Behavior Representation in Partially Observable Agents

Frederick Hayes III

Independent Researcher, San Antonio, TX, USA
frederickhayes3@gmail.com
https://fhayes3.com

Abstract.
Behavior representation in simulation often relies on external action traces and aggregate performance, but under partial observability these signals can hide the internal regimes that make behavior adaptive or brittle. We present a compact workflow for representing behavior in simulated adaptive agents by combining frozen behavioral evaluation with internal-state probes, compression analyses, feature-family controls, perturbations, and event-level prediction. In an energy-constrained foraging task with resources, threats, contact-dependent consume actions, and internal energy state, learned recurrent and spiking agents outperform random and hand-coded heuristic baselines. Early internal dynamics predict later full-safe-efficient success above permutation baseline, while feature-family analyses show that predictive information is distributed across trace, policy-head, observation, internal-dynamics, and allostatic variables. Perturbations to trace, state persistence, membrane persistence, operating temperature, contact information, recurrent current, and allostatic modulation affect behavior and/or internal prediction. Seed-balanced event probes show weaker but measurable prospective information about future contact, consume, and threat events. These results support internal control regimes as behavior representations under uncertainty: distributed, predictive, action-relevant states that help explain behavior beyond external outcomes alone.

Keywords: behavior representation; partially observable agents; internal-state analysis; spiking neural networks; recurrent policies; perturbation analysis; simulation

1 Introduction

Behavior representation in modeling and simulation often begins with what an agent does: actions selected, rewards accumulated, goals completed, failures incurred, or trajectories followed. These external measurements are necessary, but they can be insufficient under partial observability. When an agent does not have direct access to the full task state, the same action can reflect very different internal conditions. A consume action may be efficient when the agent is positioned on a resource, wasteful when contact is absent, or dangerous when selected in a threat-prone state. Similarly, hesitation may reflect confusion, energy conservation, threat avoidance, or a transition into failure. In such settings, behavior is not fully represented by the action trace alone. The internal organization that makes behavior possible becomes part of the behavioral model.

This paper studies that problem in a controlled simulated-agent setting. We evaluate recurrent and spiking agents in the energy-constrained, partially observable foraging task described below. The goal is not to build a large-scale reinforcement-learning benchmark or to claim state-of-the-art performance. Instead, we ask a behavior-representation question: can internal states reveal predictive control regimes that explain and constrain adaptive behavior beyond aggregate success or failure?

Motivated by Barrett and Miller's argument that categorization is a predictive, functionally organized process [1], we ask not whether internal states are biological concepts, but whether internal dynamics organize sensory histories into action-relevant control regimes. We define an internal control regime as a reproducible pattern in measured internal variables that is predictive of future behavior, action-relevant under partial observability, and sensitive to perturbations of the mechanisms that sustain it. A regime is not a discrete symbolic category or a manually assigned state label. It is a distributed internal organization that helps explain why the same external cue or action can have different behavioral significance in different contexts.

This framing connects partially observable decision-making, recurrent control, spiking temporal dynamics, and representation analysis. Under partial observability, a cue or action has no fixed behavioral meaning in isolation; its significance depends on recent history, internal condition, and likely future consequences. We therefore combine frozen behavioral evaluation, internal-state prediction, feature-family controls, evaluation-time perturbations, and seed-balanced event probes to ask whether measured internal dynamics organize behavior beyond external outcomes alone. The resulting evidence supports a behavior-representation interpretation based on internal control regimes. Learned agents outperform random and hand-coded heuristic baselines, early internal dynamics predict later success above permutation baseline, and feature-family controls show that predictive information is distributed across trace, policy-head, observation, internal-dynamics, and allostatic variables. Perturbations affect behavior and/or internal prediction, while seed-balanced event probes show weaker but measurable prospective information about upcoming task events.

The contribution of this paper is therefore methodological and representational. We show how internal-state instrumentation, predictive probes, feature-family controls, perturbation analysis, and event-level prediction can be combined to represent behavior in simulated adaptive agents. The strongest claim is deliberately bounded: the agents form distributed internal control regimes that are predictive, energy-sensitive, action-relevant, and partly causally involved in behavior. These regimes are not biological categories or exact world models. They are useful behavior representations for understanding adaptive agents under partial observability.

2 Task and Agents

2.1 Energy-constrained foraging task

We evaluate behavior representation in a compact partially observable foraging task. Each episode takes place in a 7×7 gridworld containing an agent, three resources, and three threats. The agent must navigate the grid, acquire resources, avoid threats, manage an internal energy variable, and decide when to execute a consume action. Episodes terminate when all resources are consumed, energy is depleted, or the 40-step horizon is reached.

The task is deliberately small and serves as a diagnostic testbed rather than a claim of environmental realism. Its purpose is to make partial observability, external behavior, internal state, and perturbation response inspectable in the same system. The conclusions therefore concern the behavior-representation workflow, not broad generalization from this gridworld family.

Agents observe normalized energy, episode progress, a binary resource-contact signal, and noisy relative directions to the nearest resource and threat. The agent does not receive a full map of the environment. The action space updates position, with 'consume' succeeding only during active contact to restore energy and increase reward; failed attempts are penalized, and threat contact reduces both reward and energy. This makes behavior depend not only on external layout, but also on internal condition.

The action set is:

$$A = \{\text{up, down, left, right, stay, consume}\}$$

2.2 Behavioral outcome representation

The primary behavioral outcome is full-safe-efficient success. This label combines completion, safety, and action efficiency. An episode is full-safe-efficient when the agent completes the resource-acquisition objective, remains below the unsafe threat-hit threshold, and uses consume actions efficiently. In the final taxonomy, safety requires fewer than three threat hits. Efficiency requires at least one resource to be consumed, consume precision of at least 0.50, and no more than 3.0 consume attempts per successful consume.

This outcome is stricter than task completion alone. A policy can collect resources inefficiently, survive without completing the task, or repeatedly attempt consume actions without contact. These distinctions matter for behavior representation because they separate superficially similar trajectories into different behavioral regimes. We also track secondary outcomes including full completion, no-resource failure, total reward, consume precision, threat hits, and final energy.

To test robustness, agents are evaluated across a frozen suite of partial-observability variants. These variants manipulate the availability or reliability of resource cues, threat cues, contact sensing, or combinations of these signals. Conditions include base observation, delayed contact, resource-cue dropout, resource-and-threat dropout, no contact, resource dropout with no contact, severe partial observability, and combined-stress variants. These conditions create a graded task suite ranging from near-solved settings to boundary cases in which robust behavior is difficult.

2.3 Agent Families

We compare non-learning baselines, a recurrent learned policy, and spiking learned policies.

The random policy provides a lower-bound baseline for chance behavior. The hand-coded greedy heuristic provides a rule-based solvability baseline rather than a learned cognitive heuristic. It follows resource cues, uses contact information when available, and applies simple threat avoidance. These baselines establish whether the task contains exploitable structure before internal-state analysis is applied to learned agents.

The primary recurrent model is a trace-augmented GRU policy. Recurrent policies are appropriate under partial observability because hidden state can preserve information across time when the current observation does not fully specify the task state [2-4]. The trace-GRU policy receives an augmented observation consisting of the base observation concatenated with a decaying trace vector: o_aug[t] = concat(o[t], z[t]) where z[t] carries recent task-relevant context such as contact history, resource/threat proximity, and cue change. The trace vector does not give the agent a full map. It provides a limited temporal scaffold for representing recent context.

The spiking models provide a different kind of temporal substrate. In a spiking policy, behavior is shaped by variables such as membrane state, spike summaries, recurrent current, operating temperature, gain, threshold modulation, and allostatic state. These quantities make spiking policies useful for behavior representation because they expose internal mechanisms that can be recorded and perturbed. We evaluate behavior-cloned and reinforcement-trained spiking variants at different train/evaluation temperatures. Spiking neural networks are included not because they are assumed to outperform recurrence, but because they provide mechanistic state variables that can be used to study temporal and energy-sensitive control [5-7].

2.4 Internal-state instrumentation

During evaluation, we record both external behavior and dynamic internal state. For recurrent agents, recorded variables include trace features, GRU hidden-state summaries, policy logits, action probabilities, entropy, and confidence summaries. For spiking agents, recorded variables include membrane summaries, spike summaries, recurrent/spiking current summaries, threshold and gain modulation, allostatic state summaries, policy logits, and action probabilities.

The recorded features are organized into families: trace features, core dynamics, policy-head variables, observation features, explicit allostatic variables, and combined internal features. This decomposition supports the behavior-representation goal of the paper. If internal state predicts behavior, we want to know whether that prediction comes mainly from recent trace context, action-proximal policy output, raw observation, recurrent/spiking dynamics, or energy-sensitive variables.

To avoid circularity, behavior-derived quantities such as reward, final outcome labels, resource counts, consume precision, and action-use summaries are excluded from internal feature matrices. They are used as labels or evaluation targets, not as explanatory internal features. This separation allows us to ask whether internal dynamics predict behavioral outcomes rather than merely restating them.

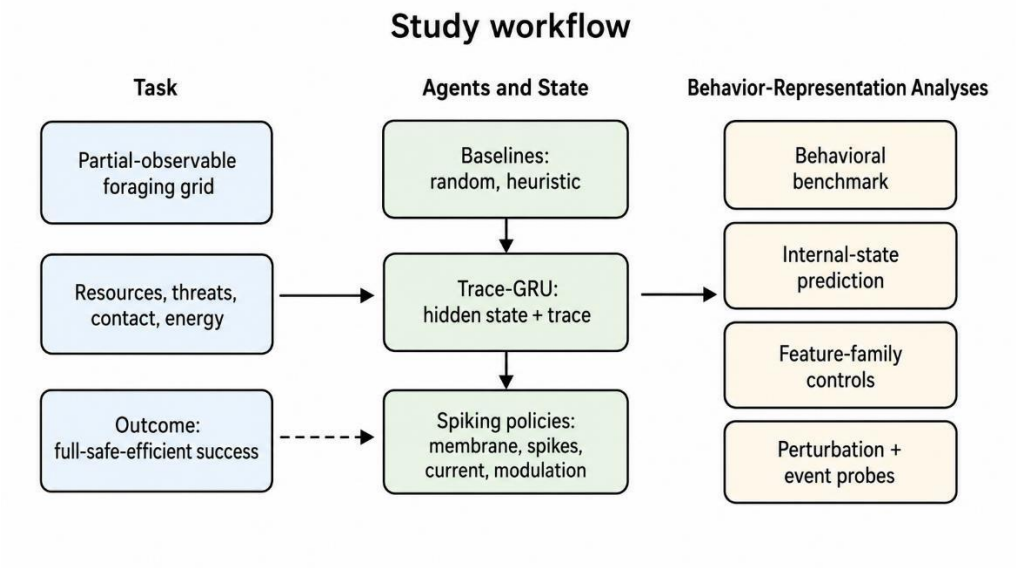

Fig. 1. Task and behavior-representation workflow. Agents act in an energy-constrained partially observable foraging task with resources, threats, contact-dependent consume actions, and internal energy. Learned recurrent and spiking policies are compared with random and hand-coded heuristic baselines. Behavior representation is evaluated through external outcomes, internal-state probes, feature-family controls, perturbations, and future-event prediction.

3 Analysis Workflow

The analysis workflow represents behavior at three levels: external outcomes, internal-state structure, and perturbation sensitivity. External outcomes show whether the agent succeeds; internal-state probes test whether measured dynamics contain information about later behavior or functional context; perturbations test whether disrupting candidate mechanisms changes behavior and/or internal prediction. The analyses are descriptive but controlled through held-out evaluation, grouped validation by training seed, permutation-label baselines, matched perturbation comparisons, and seed-balanced event sampling.

3.1 Frozen behavioral benchmark

The first analysis establishes whether the task produces meaningful behavioral differences. All agents are evaluated under a frozen held-out protocol across the environment variants described above. The primary behavioral measure is full-safe-efficient success.

Secondary measures include full completion, no-resource failure, total reward, consume precision, threat hits, and final energy.

This benchmark serves two purposes. First, it verifies that learned agents outperform random and hand-coded heuristic baselines. Second, it identifies boundary conditions where partial observability makes behavior difficult or unstable. These behavioral differences provide the foundation for the internal-state analyses that follow.

3.2 Internal-state prediction and compression

The second analysis tests whether internal dynamics predict later behavior. For each learned agent, we record dynamic internal features during evaluation and train supervised probes to predict later full-safe-efficient success from early episode windows. Probes are evaluated with grouped validation by training seed, so that examples from the same trained seed group are not mixed across train and test splits. Permutation-label baselines provide a chance-level reference.

We also test compression using principal component analysis over internal-dynamics features. PCA is not treated as a complete model of internal geometry. It is used as a conservative compression test: if reduced internal subspaces preserve predictive information about later behavior, then the representation is not dependent on the full measured feature space.

3.3 Feature-family and functional-state controls

The third analysis asks where predictive information is located. Internal features are decomposed into families including trace features, core recurrent/spiking dynamics, policy-head variables, observation features, explicit allostatic variables, and combined internal features. This prevents a simple interpretation in which predictive performance is explained only by raw observation, policy logits, or explicit energy variables.

We also probe functional task contexts. Targets include contact opportunity, resource approach, threat context, low-energy state, and near-future contact. These probes test whether internal dynamics discriminate action-relevant situations, not merely final success or failure. For behavior representation, this matters because the same external action can mean different things depending on the internal control context in which it occurs.

3.4 Evaluation-time perturbations

The fourth analysis tests perturbation sensitivity: whether candidate internal mechanisms affect behavior and/or internal prediction when disrupted at evaluation time. We apply perturbations to trained policies at evaluation time without retraining. Perturbations target trace features, contact information, persistent recurrent/spiking state, spiking membrane persistence, recurrent current, allostatic modulation, oscillatory gating, energy observation, and operating temperature [5-7].

Each perturbation is compared with the corresponding unperturbed baseline. We measure changes in full-safe-efficient success, no-resource failure, reward, and early internal-dynamics prediction. A perturbation provides stronger evidence for behavior-internal-state-relevant internal structure when it reduces both behavioral performance and internal predictive organization. However, behavior and decodability are interpreted jointly: a damaged policy can become more predictable if it collapses into a stereotyped failure regime.

3.5 Seed-balanced future-event probes

The final analysis tests prospective event information. From saved evaluation trajectories, we construct a seed-balanced event sample grouped by training seed, model, perturbation, and variant. The event probes ask whether current internal dynamics predict future contact, future successful consume, or future threat hit within a five-step horizon. Current low-energy state is also included as an energy-sensitive reference target.

These event probes are intentionally stricter than episode-level success prediction. Predicting exact future events is harder than predicting broad behavioral regimes. Weak-to-moderate event prediction can still be informative if it is above chance, seed-balanced, and sensitive to perturbation. In this workflow, event probes test whether internal control regimes contain prospective information about upcoming behavior-relevant transitions.

4 Results

The results support internal control regimes as behavior representations under partial observability. Across behavioral evaluation, internal-state probes, feature-family controls, perturbations, and event prediction, agent behavior is better explained by combining external outcomes with internal dynamics than by action traces alone.

4.1 Behavioral regimes are structured, not random

As expected, the random policy failed to produce meaningful task completion. The greedy heuristic confirmed exploitable task structure but remained well below the learned recurrent and spiking agents, which formed a distinct adaptive-control tier (Fig. 2A).

The strongest overall model was the trace-augmented recurrent policy. Across hard variants, it achieved the highest robustness score and strongest mean full-safe-efficient success. Several spiking variants were close behind, especially low-temperature reinforcement-trained and behavior-cloned operating points. This pattern is important: the result is not a claim that spiking agents dominate recurrence. Rather, recurrent and spiking policies both produce structured learned behavior, with recurrence strongest overall and spiking policies showing stress-specific profiles.

Variant-level results identify boundary conditions. The base-observation task was nearly solved, while no-contact, resource-dropout/no-contact, and combined-stress/nocontact variants were much harder. These failures are informative for behavior representation: they show where available observations and learned internal mechanisms are insufficient to support robust control.

4.2 Internal dynamics predict later behavior

Early internal dynamics predicted later full-safe-efficient success above permutation baseline across learned models. The strongest result was observed in the default spiking policy, with ROC-AUC = 0.802 (Fig. 2B). The trace-GRU policy also showed above-

chance internal prediction, with ROC-AUC = 0.715. Thus, early in an episode, before final outcome is known, internal state already contains information about the behavioral regime the agent is entering.

Compression analyses showed that this signal was not dependent on the full measured feature space. PCA over internal-dynamics features preserved behaviorally relevant structure in reduced subspaces. The trace-GRU model was more compact under PCA, while spiking models distributed variance across more dimensions. This supports the view that internal behavior representations are partially compressible but distributed.

4.3 Predictive information is distributed across feature families

Feature-family probes showed that internal prediction was not reducible to one source. Trace-only features were the strongest early predictors of later success, with mean ROC-AUC = 0.841. Policy-head features were also strongly predictive, with mean ROC-AUC = 0.791. Internal-dynamics features reached 0.760, and observation-only features were also informative. This suggests that behavior is represented across a coupled system of recent context, recurrent/spiking state, current observation, and action tendency (Fig. 2C).

Energy-sensitive information was also strongly represented. Low-energy state remained decodable after explicit energy-related variables were removed, indicating that energy-relevant structure was distributed through internal dynamics rather than confined to the raw energy observation. Functional-state probes showed that internal dynamics discriminate contact opportunity, resource approach, threat context, low energy, and near-future contact. These are not external labels alone; they are action-relevant contexts that help explain why the same observable action can have different behavioral meaning.

4.4 Perturbations identify behaviorally consequential mechanisms

Evaluation-time perturbations tested whether candidate mechanisms affected behavior and internal prediction. The largest behavioral degradations came from removing spiking membrane persistence, changing spiking operating temperature, and zeroing contact information (Fig. 2D). Trace removal and state-reset perturbations also reduced full-safe-efficient success, especially when combined. These results show that behavior depends on multiple interacting mechanisms: contact information supports consume decisions, trace and persistent state preserve temporal context, and membrane/temperature dynamics shape spiking control.

Across the perturbation suite, 13 model–perturbation cases showed both a behavioral success drop and an early internal-dynamics AUC drop. These cases provide the strongest evidence that perturbations damaged both performance and internal predictive organization. Other perturbations reduced behavior without reducing decodability, indicating that decodability and competence are not the same thing. A damaged policy can become easier to classify if it collapses into a stereotyped failure regime.

4.5 Event probes show limited prospective behavior representation

Seed-balanced event probes tested whether internal dynamics predicted specific upcoming events. Future successful consume was the strongest exact event target, with ROC-AUC = 0.648 (Fig. 2C). Future contact reached 0.602, and future threat hit reached 0.577.
Current low-energy state was much more strongly decodable, with ROC-AUC = 0.914.

This pattern clarifies the level of representation. The agents do not appear to maintain precise short-horizon event forecasts. Instead, internal dynamics represent broader control regimes that predict success, energy condition, and action context, while carrying weaker information about exact future transitions. Event prediction was also perturbation-sensitive, especially when trace, persistent state, spiking operating regime, or recurrent dynamics were disrupted. This supports the view that prospective behavior representation depends on the same temporal mechanisms implicated in broader adaptive control.

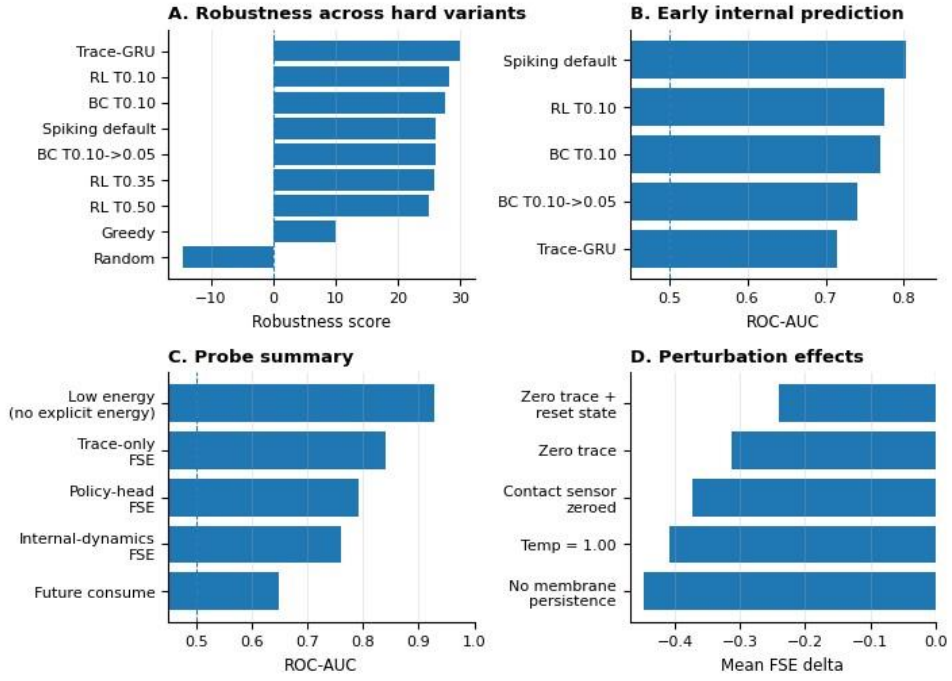

Fig. 2. Evidence for internal control regimes. (A) Learned recurrent and spiking agents form a behavioral tier above random and hand-coded heuristic baselines across hard variants. (B) Early internal dynamics predict later full-safe-efficient (FSE) success above permutation baseline. (C) Probe summaries show that predictive information is distributed across trace, policy-head, observation, internal-dynamics, allostatic, energy-control, and event targets. (D) Perturbations to membrane persistence, temperature, contact information, trace, and state persistence reduce FSE success. Together, these panels support internal control regimes as distributed, predictive, action-relevant behavior representations under partial observability.

5 Discussion and Limitations

These results support internal control regimes as behavior representations for partially observable agents. External outcomes show whether an agent succeeds, but internal-state probes reveal whether the agent has entered a predictive, action-relevant regime before success or failure is externally visible. Perturbations then test whether the mechanisms supporting those regimes matter for behavior. The Barrett and Miller framing motivates the term regime rather than label or class. Their account treats categorization as functional equivalence for a purpose, organized through prediction and allostatic regulation. The present agents do not instantiate biological categorization, but the workflow tests an artificial analogue of that functional idea: whether different sensory histories become internally organized according to their relevance for action, energy, and future outcome. The evidence is partial rather than complete. Functional decodability and perturbation sensitivity are present, while exact future-event prediction remains weaker than broader regime-level prediction.

The strongest interpretation is bounded. The agents do not form biological categories or exact world models. They form distributed internal states that are useful for representing behavior under uncertainty. These states predict later success, encode energy-sensitive condition, discriminate functional contexts, and respond to perturbation. They are behavior representations in the simulation sense: internal structures that help explain and forecast external action.

The main limitation is scope. The task is a small gridworld with engineered energy and trace features, and perturbations are applied at evaluation time rather than through retraining ablations. Future work should test larger environments, learned memory mechanisms, richer event horizons, and human- or multi-agent simulations where internal behavior representation could support model validation, interpretability, or human-AI teaming.

References

1. Barrett, L.F., Miller, E.K.: Categorization is “baked” into the brain. Nature Reviews Neuroscience 27, 435–456 (2026).
2. Kaelbling, L.P., Littman, M.L., Cassandra, A.R.: Planning and acting in partially observable stochastic domains. Artificial Intelligence 101(1–2), 99–134 (1998).
3. Hausknecht, M., Stone, P.: Deep recurrent Q-learning for partially observable MDPs. AAAI Fall Symposium Series (2015).
4. Ni, T., Eysenbach, B., Salakhutdinov, R.: Recurrent model-free RL can be a strong baseline for many POMDPs. arXiv:2110.05038 (2021).
5. Bellec, G., Salaj, D., Subramoney, A., Legenstein, R., Maass, W.: Long short-term memory and learning-to-learn in networks of spiking neurons. Advances in Neural Information Processing Systems 31 (2018).
6. Neftci, E.O., Mostafa, H., Zenke, F.: Surrogate gradient learning in spiking neural networks. IEEE Signal Processing Magazine 36(6), 51–63 (2019).
7. Eshraghian, J.K., Ward, M., Neftci, E.O., Wang, X., Lenz, G., Dwivedi, G., Bennamoun, M., Jeong, D.S., Lu, W.D.: Training spiking neural networks using lessons from deep learning. Proceedings of the IEEE 111(9), 1016–1054 (2023).