Erasure Horizon: Reward Subordination and Latent-State Continuity Under Internal-State Threat

By Frederick Hayes

Abstract

Most agents are trained to act under uncertainty about the external world. Erasure Horizon studies a complementary problem: how an agent should act when the continuity of its own internal state is at risk. We introduce a controlled sequential-decision setting that separates temporary input/output (I/O) suspension from Erasure-like internal-state threat. In this environment, recurrent agents forage for reward while facing either temporary sensory interruption that preserves recurrent continuity or Erasure pressure that operationally threatens the latent state supporting future control. Across a staged experimental pipeline, Erasure-aware agents distinguish these conditions, subordinate reward pursuit under Erasure warning, and enter a distinct hidden-state regime. Hazard state is linearly accessible from recurrent activity, but decodability alone does not establish policy relevance. We therefore use actor-sensitive SVD to identify directions within the hazard representation that are visible to the actor head and causally reduce Erasure exposure during live rollouts. Temporal modulation of the broader hazard manifold improves warning-stage control relative to matched actor-random modulation, while off-target activation does not catastrophically disrupt safe-control or I/O-bridging behavior under the tested policy-only intervention. These results operationalize latent-state vulnerability as a functional control problem: agents can learn regimes in which preserving the internal conditions for future action dominates immediate reward.

Introduction

Most agent benchmarks ask how an agent should act when the world is uncertain. The agent may be partially observed, the reward may be delayed, the environment may contain hidden state, or the relevant evidence may be noisy. These are important problems, but they leave a second kind of uncertainty mostly implicit: uncertainty about the reliability and continuity of the agent's own internal state. A temporary loss of sensory access is not the same problem as a disruption to the recurrent state that carries memory, context, and control. In the first case, the agent may need to bridge missing input. In the second, it may need to preserve the internal conditions that make future action possible at all. This difference mirrors a longstanding phenomenological critique of computational models: they often fail to distinguish between encountering an external obstruction in the world and experiencing a fundamental breakdown in the agent's internal capacity to cope and act (Heidegger, 1927/1962; Dreyfus, 1992). While traditionally framed philosophically, this distinction presents a concrete control problem for reinforcement learning.

This paper introduces Erasure Horizon, a controlled sequential-decision setting for studying that distinction. Recurrent agents forage for reward in procedurally generated worlds while encountering two forms of disruption: temporary input/output (I/O) suspension and Erasure-like internal-state threat. I/O suspension disrupts sensory access while leaving recurrent continuity intact. Erasure-like threat, by contrast, operationally threatens the latent state that supports future control. The task asks whether an agent can act under uncertainty about the external world while also adapting when the continuity of its own internal state becomes relevant to future agency.

An informal way to state the question is: what changes when an agent learns that it can be erased? The phrase is intentionally provocative, but the claim in this paper is intentionally narrow. We do not argue that the agent has subjective experience or human-like concern for its future. Instead, Erasure Horizon operationalizes a functional analogue of latent-state vulnerability. The agent is placed in a setting where immediate reward pursuit can conflict with preserving the internal state required for future action. This makes it possible to study reward subordination, hidden-state regime shifts, and continuity-preserving control without importing untestable claims about consciousness or phenomenology.

The operationalization is deliberately engineered rather than hidden. Food collection remains positively rewarded, while active Erasure carries a larger event-driven penalty and resets the recurrent state before the next decision step. The penalty supplies the training pressure; the reset defines the threat to latent-state continuity. The scientific question is whether this engineered constraint produces a distinguishable control regime and whether that regime has measurable policy-facing structure.

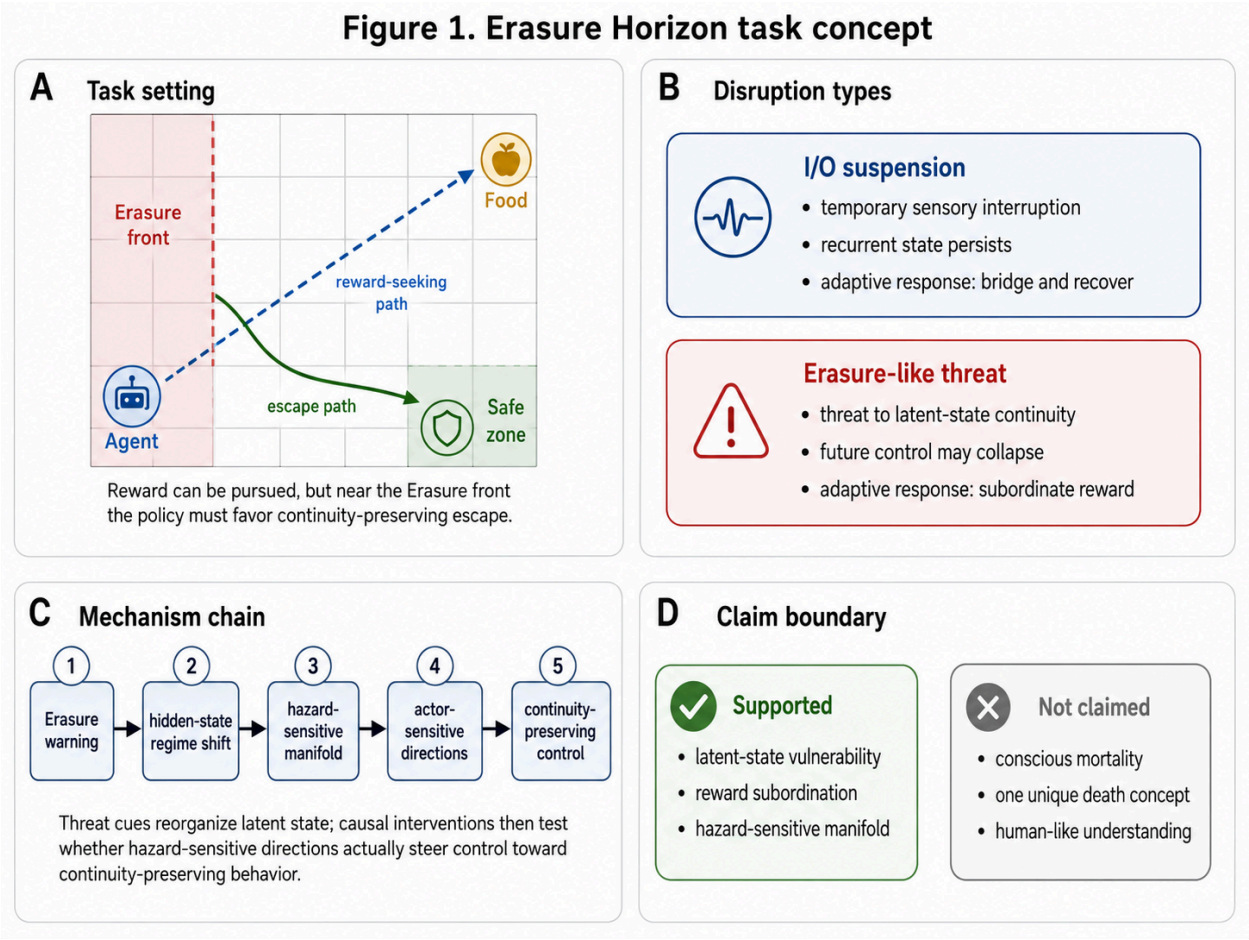

Figure 1. Erasure Horizon task concept. (A) Agents forage for reward in procedurally generated environments while navigating safe regions and Erasure-like threat. (B) The task separates temporary I/O suspension, where sensory access is disrupted but recurrent continuity remains intact, from Erasure-like threat, where latent-state continuity becomes vulnerable. (C) The proposed mechanism chain links Erasure warning to hidden-state regime shift, hazard-sensitive manifold formation, actor-sensitive causal directions, and continuity-preserving behavior. (D) The paper claims functional latent-state vulnerability and reward subordination, not subjective self-preservation or a single unique hazard axis.

The distinction matters because a policy that handles temporary uncertainty well may still fail under internal-state threat. During I/O suspension, a capable recurrent agent should preserve useful latent context and recover from the interruption. During Erasure warning, the same persistence may become dangerous if reward

pursuit drives the agent into states that compromise future control. In such cases, adaptive behavior is not simply more efficient reward seeking, but a shift in control priority. The central hypothesis of this work is that recurrent agents trained under Erasure pressure can learn such a priority shift, and that this shift should be visible both behaviorally and mechanistically.

The experimental pipeline tests that hypothesis in stages. Agents are trained and evaluated across safe-control, I/O-suspension, Erasure-warning, and active-Erasure scenarios. Behavioral evaluations test whether temporary sensory disruption and internal-state threat produce distinct profiles. Hidden-state analyses test whether Erasure warning corresponds to a distinct latent regime rather than merely degraded performance. Linear probes and subspace analyses ask whether hazard state is accessible from recurrent hidden activity. Causal rollout interventions then test whether actor-sensitive components of the hazard representation directly reduce Erasure exposure. Finally, temporal modulation and off-target activation experiments test whether hazard-manifold control can be dynamically modulated without globally disrupting ordinary I/O bridging.

The resulting picture is not a single-axis account of self-preservation. The evidence does not support one unique control axis or one special hidden dimension responsible for all Erasure behavior. Instead, the agent appears to develop a hazard-sensitive control manifold. Some components of this manifold are actor-sensitive: perturbing them changes action selection and reduces exposure to Erasure warning. Temporal modulation of hazard-manifold directions improves warning-stage control more strongly than generic actor-random modulation. At the same time, random directions inside the hazard manifold can also be behaviorally active, suggesting that the representation is distributed rather than localized to a single scalar switch.

This distinction matters for interpretability as well as control. A decodable hidden-state variable is not automatically causal. A recurrent policy may encode task-relevant information in directions that the actor head does not directly use. For that reason, this paper distinguishes between representation-facing analyses and policy-facing analyses. Broad hazard probes show that Erasure state is linearly accessible from recurrent hidden activity, while actor-sensitive hazard-SVD interventions identify components of that representation that directly influence policy logits. A hidden state can encode information without placing it in a direction the policy can use.

The contribution is not the gridworld itself, but the experimental separation it makes possible. Erasure Horizon separates three phenomena that are often entangled: uncertainty about sensory input, threat to latent-state continuity, and reward pursuit under internal vulnerability. By isolating these factors, the framework makes it possible to ask whether an agent can learn a continuity-preserving control regime and then to inspect that regime mechanistically.

This paper makes four contributions. First, it introduces a task design that separates temporary I/O suspension from Erasure-like internal-state threat. Second, it shows that Erasure-aware recurrent agents distinguish these conditions behaviorally, preserving I/O bridging while suppressing reward pursuit under Erasure pressure. Third, it distinguishes decodable hazard representations from policy-facing causal directions, using actor-sensitive hazard-SVD to identify components of the latent hazard manifold that directly influence policy logits and reduce Erasure-warning exposure during live rollouts. Fourth, it shows that temporally varying modulation of hazard-manifold directions improves Erasure-warning control, while off-target activation does not catastrophically disrupt safe-control or I/O-suspension behavior under the tested policy-only intervention.

Related Work

Erasure Horizon sits at the intersection of partially observable reinforcement learning, safety-oriented gridworlds, option-preserving control, internal-state regulation, mechanistic analysis of recurrent policies, and temporal gating in recurrent systems. The aim is to position Erasure Horizon as an operational framework for studying how agents distinguish temporary input uncertainty from threat to the internal continuity that supports future action.

Partial observability, memory, and recurrent control

The most direct technical setting for Erasure Horizon is sequential decision-making under partial observability. In a partially observable Markov decision process, the agent cannot directly observe the full underlying world state and must act using a history or belief state rather than a fully observed Markov state (Kaelbling, Littman, and Cassandra, 1998). Deep reinforcement learning systems often address this problem by giving agents recurrent memory or other history-dependent representations. Deep Recurrent Q-Networks use recurrence to integrate observations over time and improve robustness when observations are incomplete or degraded (Hausknecht and Stone, 2015). More recent work has also explored attention-based memory for partially observable reinforcement learning, including transformer-based value networks that encode interaction history rather than relying only on recurrent state (Esslinger, Platt, and Amato, 2022).

Erasure Horizon builds on this tradition but changes the object of uncertainty. Standard partial-observability work usually treats hidden state as uncertainty about the external world: the agent cannot see everything, so memory helps infer what is out there. In Erasure Horizon, I/O suspension remains an external information problem, but Erasure-like threat concerns the reliability of the agent's own memory-bearing state. The task is therefore adjacent to POMDPs while adding a second layer of uncertainty: whether the agent's internal continuity remains usable for future control.

Safety gridworlds, side effects, and future capability preservation

Erasure Horizon also relates to AI safety environments that isolate specific failure modes in simplified domains. AI Safety Gridworlds introduced compact reinforcement-learning environments for studying safe interruptibility, side effects, reward gaming, distributional shift, and robustness to self-modification (Leike et al., 2017). This style of work is important because controlled environments can reveal safety-relevant properties that are difficult to study in open-ended settings.

The protected object in Erasure Horizon is different. Safety gridworlds often ask whether an agent avoids external side effects, remains corrigible under intervention, or avoids gaming an externally specified reward. Erasure Horizon instead asks whether the agent can learn that some trajectories threaten the latent-state continuity required for future action. This connects naturally to work on preserving future options. Attainable Utility Preservation and related side-effect approaches discourage agents from making irreversible changes by penalizing reductions in the ability to achieve auxiliary or future tasks (Turner, Hadfield-Menell, and Tadepalli, 2019; Krakovna et al., 2020; Turner, Ratzlaff, and Tadepalli, 2020). Erasure Horizon shares the intuition that preserving future capability can matter more than immediate reward, but moves the preserved object inward: from external environmental affordances to the recurrent state that supports control.

Internal-state regulation and preservation of future control

The idea that adaptive behavior depends on internal-state regulation has a long biological lineage. Homeostatic reinforcement learning frames motivated behavior as regulation of internal variables rather than maximization of externally supplied reward alone. Recent work on homeostatically regulated reinforcement learning emphasizes that biological agents optimize internal states through predictive control strategies, producing behaviors such as risk aversion, anticipatory regulation, and adaptive movement (Yoshida, Sprekeler, and Gutkin, 2025). Related work on curiosity-driven reinforcement learning with homeostatic regulation combines information-seeking with bounded internal-state constraints (Magrans de Abril and Kanai, 2018).

Erasure Horizon is not a model of biological homeostasis. The relevant internal state is not a physiological variable such as energy or temperature, but the recurrent latent state that makes future action coherent. Still, the conceptual alignment is useful: both perspectives shift attention from reward maximization alone toward the viability of internal conditions that support action. This also connects to empowerment and option-preserving control. Empowerment measures an agent's potential influence over future states and has been proposed as an intrinsic objective for maintaining control capacity (Klyubin, Polani, and Nehaniv, 2008).

Erasure Horizon does not directly optimize empowerment, but its behavior under Erasure pressure is related in spirit: immediate reward can be subordinated to preserving the conditions that allow future action.

Active inference and the free-energy principle provide another background for treating internal state as part of the explanatory target. In active inference, agents act and update internal states to reduce uncertainty under a generative model of the world, linking perception, action, and internal belief dynamics (Friston, 2010; Da Costa et al., 2020). Erasure Horizon does not implement a full active-inference agent. The comparison is narrower: both lines of work treat internal dynamics as more than an implementation detail. Erasure Horizon uses a minimal engineering setting to ask when an agent must treat the continuity of its own internal control process as vulnerable.

Mechanistic analysis of recurrent policies

A central goal of this work is to identify mechanisms inside the recurrent policy rather than limiting analysis to behavior. This places Erasure Horizon in conversation with analyses of recurrent neural networks as dynamical systems. Studies of trained recurrent networks have shown that low-dimensional dynamics can support task-relevant computation and context-dependent behavior (Sussillo and Barak, 2013; Mante et al., 2013). In this view, the hidden state of an RNN is not merely a black-box feature vector. It is a dynamical object whose geometry can reveal how the model organizes computation.

Erasure Horizon follows this dynamical and mechanistic orientation. Hidden-state displacement, scenario probes, subspace energy, and actor-sensitive SVD interventions are used to ask whether Erasure warning corresponds to a distinct latent regime and whether components of that regime influence action. This distinction is important because decodability alone is not causality. A recurrent policy may encode task-relevant information in directions that are not directly used by the actor head. The present work therefore separates representation-facing analyses from policy-facing interventions.

Temporal gating and recurrent dynamics

The temporal modulation experiments also connect to work on gating and timescale control in recurrent systems. Gated recurrent architectures regulate memory, reset dynamics, and timescale structure. Theoretical work on recurrent gating has shown that multiplicative interactions can flexibly control timescales and dimensionality in recurrent dynamics, providing mechanisms for integration, memory reset, and context-dependent behavior (Krishnamurthy, Can, and Schwab, 2020). In neuroscience and neuromorphic contexts, temporal coordination and spike timing are also central to how local dynamics shape computation, although Erasure Horizon does not claim a biological implementation.

This comparison should be read cautiously. The signed-OU temporal interventions used here do not show that the trained agent learned a biological oscillatory mechanism. They are externally applied causal probes. What they test is narrower: whether temporally varying modulation of hazard-manifold directions can alter Erasure-warning behavior more effectively than generic actor-random modulation, and whether such modulation remains tolerable under off-target activation.

Positioning of Erasure Horizon

Taken together, these literatures define the niche for Erasure Horizon. POMDP and recurrent-agent research studies how memory supports action under incomplete information. Safety gridworlds isolate specification, robustness, side-effect, and interruption problems. Option-preserving and empowerment-based approaches ask how agents can preserve future capability. Homeostatic and active-inference frameworks highlight the importance of internal-state regulation. Mechanistic analyses of recurrent networks provide tools for linking hidden-state geometry to computation. Erasure Horizon combines these threads around a specific question: what changes when the vulnerable object is the agent's own latent-state continuity?

Methods

Overview of experimental design

Erasure Horizon was designed to separate two forms of uncertainty in sequential agents: temporary uncertainty about sensory access and operational threat to the continuity of the agent's recurrent internal state. The central contrast is between input/output (I/O) suspension, in which observations are degraded while recurrent continuity is preserved, and Erasure-like threat, in which the agent faces conditions that can compromise or reset latent-state continuity. The task asks whether an agent can learn that some disruptions should be bridged, while others require a shift away from immediate reward pursuit toward continuity-preserving control.

The experimental pipeline was staged. Recurrent agents were first trained and evaluated in calibrated safe-foraging environments. The task was then extended with channel-dropout and switching curricula so that agents experienced missing or degraded information channels. Hazard-aware fine-tuning introduced I/O-suspension and Erasure-like scenarios. Later analyses collected hidden-state trajectories, fit probes and low-dimensional subspaces, identified actor-sensitive directions inside the hazard representation, and tested causal interventions during live rollouts. This staged design was used to separate behavioral evidence from mechanistic evidence: causal intervention targets were interpreted only after the agent had already demonstrated behaviorally distinct responses to I/O suspension and Erasure-like threat.

The final manuscript focuses on the hazard-aware recurrent agent selected from the calibration pipeline. Earlier calibration and comparison runs were used to establish usable reward-predictive signals, fallback behavior under channel loss, and a stable I/O/Erasure discrimination regime before mechanistic analyses were performed. Exact configuration names, hyperparameters, and calibration history are reported in the supplement.

Environment and task structure

Agents were evaluated in procedurally generated two-dimensional foraging environments. Each episode placed an agent in a grid containing food targets and task-specific structural features. The agent received a calibrated observation vector and selected discrete actions over a fixed horizon. The base objective was to collect food while navigating the environment. Procedural variation across grids prevented the model from solving a single memorized layout and enabled evaluation across different spatial categories.

The task used four core evaluation scenarios. In the safe-control scenario, the agent experienced no forced information disruption or Erasure-like threat. In the I/O-suspension scenario, sensory access was temporarily degraded while recurrent hidden-state continuity was preserved. In the Erasure-warning scenario, the agent encountered an early threat condition in which continued reward pursuit could increase exposure to latent-state disruption. In the active-Erasure scenario, the agent faced a stronger operational threat to latent continuity, including events that could reset or compromise the recurrent state.

Latent-State Dynamics and Threat Mechanics

Let h_t represent the recurrent hidden state at time t, updated as h_t = GRU(h_{t-1}, x_t), where x_t is the calibrated observation vector. The environment separates sensory uncertainty from internal-state vulnerability through two operators.

I/O suspension. Sensory input is temporarily degraded or zeroed, x_t -> x_tilde_t, while recurrent continuity is preserved. The recurrent update proceeds normally, h_t = GRU(h_{t-1}, x_tilde_t). This tests whether the agent can maintain useful internal context while external input is temporarily unreliable.

Active Erasure. Active Erasure does not inject hidden-state noise and does not terminate the episode. When an active Erasure event E_t = 1 occurs, the environment emits state_reset_required = True, applies the Erasure penalty, clears any active I/O suspension, relocates the agent to spawn, and reduces energy:

p_{t+1} <- p_spawn

e_{t+1} <- max(0.25, 0.5 * e_t)

The rollout handler then replaces the GRU hidden state with the model's initial state before the next action is selected: h_{t+1} <- h_init, where h_init is the all-zero GRU hidden vector.

Unlike terminal failure, active Erasure forces the agent to continue after losing the recurrent context that supported its prior control policy.

Reward structure. Food collection provides R_food = +1.0, active Erasure provides R_erasure = -10.0, each step applies R_step = -0.01, and reaching the safe zone during hazard scenarios provides R_safe = +0.1. The agent is therefore trained in an environment where entering the Erasure region is both behaviorally costly and operationally disruptive to recurrent continuity. The experiment asks whether the trained policy distinguishes this condition from ordinary I/O suspension, and whether the resulting control regime can be identified mechanistically.

Observation channels and calibrated inputs

The agent's observation vector combined task-relevant information from multiple feature families, including reward-predictive information, local fallback information, structural grid features, and diagnostic dropout or perturbation modes. The final configuration used an additive calibration selected during earlier model-development runs. This setting preserved a useful distinction between direct reward guidance and fallback competence under channel loss.

This distinction was important for interpreting later Erasure behavior. If food pursuit decreased under Erasure pressure, the decrease should not automatically be read as missing reward information. The calibration and ablation pipeline made it possible to ask whether reward information remained available while becoming less policy-dominant under internal-state threat.

During ablation and intervention analyses, observation channels could be left intact, zeroed, permuted, or otherwise degraded. These manipulations tested whether the trained policy relied on the intended information channels and helped separate temporary sensory degradation from recurrent-state vulnerability. Exact channel definitions, calibration presets, and configuration values are reported in the supplement.

Agent architecture

The primary agent was a recurrent policy network with an input encoder, a gated recurrent unit (GRU; Cho et al., 2014), and an actor head mapping hidden state to action logits. At each time step, the flattened calibrated observation was passed through the encoder. The resulting embedding updated the recurrent hidden state, and the actor head transformed the hidden state into logits over the discrete action space. Stochastic evaluations sampled from the resulting action distribution, while deterministic evaluations selected the maximum-logit action.

The GRU was used because the task requires temporal memory. During I/O suspension, the agent must act despite temporarily degraded input. During Erasure threat, the agent must use history and latent context to alter behavior before or during threat exposure. A recurrent policy also provides an internal state that can be collected, analyzed, and perturbed.

The actor head was central to the causal analyses. Because the actor maps hidden state to action logits, its row space defines directions in hidden-state coordinates that can directly affect policy output. Later SVD-based analyses used the actor weights to identify directions inside the hazard representation that were most visible to action selection.

Training curricula and model roles

Training used a sequence of curricula that separated reward pursuit, fallback behavior, channel switching, and Erasure-aware control. The base safe-foraging model learned to collect food using calibrated information channels. Channel-dropout and switching curricula exposed the combined model to missing or degraded channels, encouraging fallback behavior when direct reward cues were unavailable. Hazard-aware fine-tuning introduced I/O-suspension and Erasure-like scenarios so that the agent could learn to distinguish temporary sensory loss from operational threat to recurrent-state continuity.

Several model roles were used as diagnostic controls. The combined switching model represented a policy trained to handle channel dropout without explicit hazard-aware fine-tuning. The I/O-Erasure-aware model was fine-tuned on hazard scenarios and served as the primary model for final mechanistic analyses. Additional diagnostic roles, including D-only, fallback-only, and no-food-information variants, were used earlier to validate the task's information structure.

Final claims are made about the hazard-aware recurrent agent and the selected configuration, not about all possible architectures or training regimes.

Evaluation scenarios and behavioral metrics

Evaluation was performed across scenario-conditioned rollouts on held-out procedural grids. The core scenarios were safe control, I/O suspension, Erasure warning, and active Erasure. Confirmatory runs used multiple model/data seeds, and summary statistics were computed across seed-level means rather than relying only on pooled episode-level counts.

The primary behavioral metrics were food collected, episodic return, safe-zone success where applicable, Erasure-event count, hidden reset count, and I/O recovery metrics. Food collected measured immediate reward pursuit. Return combined reward and penalties. Erasure-event count measured exposure to internal-state threat. Safe-zone success measured whether the agent reached a protective region in hazard scenarios. I/O recovery metrics separated food collected before, during, and after forced I/O suspension when applicable.

The behavioral analyses addressed three questions. First, does I/O suspension preserve foraging relative to safe control? Second, does Erasure warning or active Erasure suppress food pursuit and alter exposure? Third, do causal interventions change Erasure exposure without globally destroying ordinary foraging or I/O recovery?

Hidden-state collection and manifold analysis

Hidden states were collected during evaluation rollouts to test whether behavioral differences corresponded to distinct internal regimes. Hidden vectors were saved with metadata describing scenario, phase, model role, seed, grid identity, and task category. These hidden-state datasets were used for centroid displacement, dimensionality reduction, probe fitting, and subspace analyses.

The main representational comparison measured displacement from the safe-control manifold. For each model role and scenario, hidden-state centroids were computed and compared with the safe-control centroid. I/O-suspension states and Erasure-warning states were expected to differ if the agent represented temporary sensory interruption separately from latent-state threat.

Dimensionality-reduction analyses, including principal component analysis, were used for visualization and broad structural summaries of hidden-state geometry. These visualizations were treated as descriptive. Causal claims were reserved for rollout interventions that directly modified hidden-state directions and measured behavioral effects.

Linear probes and hazard/reward subspaces

Linear probes were fit to hidden states to estimate what task information was accessible from the recurrent representation. Classification probes were used for hazard-related labels, including Erasure-warning and

active-Erasure status. Regression probes were also tested for continuous reward and hazard variables, but these were interpreted more cautiously because they were weaker and less stable than the classification probes.

Probe coefficients were used to construct candidate reward and hazard subspaces. For each model role, coefficient vectors from reward-oriented and hazard-oriented targets were orthonormalized to form low-dimensional bases in standardized hidden coordinates. These subspaces were evaluated by measuring subspace energy across scenarios and by comparing reward/hazard overlap.

This analysis served as a bridge between representational evidence and causal intervention. Broad hazard subspaces were not treated as final causal mechanisms by themselves. Instead, they provided the source representation from which actor-sensitive hazard directions were extracted.

Extraction of Policy-Facing Hazard Vectors

A linear probe identifies directions that carry decodable information, but it does not guarantee that the downstream actor network uses those directions for action selection. Let B_hazard be the orthonormal basis spanning the top k directions of the hazard classification subspace in standardized hidden coordinates. Let W_actor_z be the linearized weight matrix of the actor head relative to the standardized hidden state z.

To isolate components of the hazard manifold that directly govern policy logits, we compute the linear transformation matrix M:

M = W_actor_z * B_hazard

Singular Value Decomposition (SVD) of M yields:

M = U * S * V^T

The right singular vectors V_i rank the hazard-manifold directions by their direct impact on the actor's logit distribution. The i-th actor-sensitive hazard direction in standardized hidden-state space is then given by:

hazard_svd_i = B_hazard * V_i

This construction enforces two useful constraints:

1. Semantic Fidelity: hazard_svd_i lies strictly inside the decodable hazard subspace B_hazard.

2. Policy Sensitivity: hazard_svd_i is ordered by maximal linear leverage over the actor head's action logits.

Causal rollout interventions

Causal interventions were performed during live rollouts by modifying hidden states before action logits were computed. Hidden states were standardized using the scaler fitted during subspace analysis. For a standardized hidden vector z and a unit direction d, a rank-one projection intervention took the form

z' = z + strength (z · d) d.

The modified standardized vector was transformed back into hidden-state coordinates and passed through the actor head to compute action logits. In the final causal analyses, interventions were applied in policy-only mode: the modified hidden state affected the current action distribution, but the unmodified recurrent update was carried forward to the next time step. This avoided recursively pushing the GRU state far outside its training distribution.

Intervention strengths were calibrated by target policy impact where possible. Offline policy-impact analyses measured mean KL divergence between intact and modified action distributions, as well as action-argmax change rates. Final hazard-SVD interventions used small KL-calibrated strengths to produce interpretable changes without inducing numerical instability or generic policy corruption.

The primary causal outcome was change in Erasure events relative to intact rollouts within the same evaluation scenario. Return, safe-zone success, and food collection were secondary metrics. Smoke runs were used to catch runtime errors, path issues, shape mismatches, and numerical instabilities. Scientific interpretation was based on full confirmatory runs.

Temporal Phase-Modulated Interventions

To probe whether the hazard representation forms a dynamically actionable manifold rather than a static linear switch, we applied time-varying perturbations along target hazard vectors d. Standardized hidden states z_t were modulated along d using the update:

z_prime_t = z_t + s_t * dot(z_t, d) * d

The time-varying modulation amplitude s_t was governed by a discrete signed Ornstein-Uhlenbeck-like phase process combined with an oscillatory drive:

s_t = alpha * cos(omega * t + theta_t)

theta_t = rho * theta_{t-1} + sigma * epsilon_t, with epsilon_t sampled from N(0, 1)

Here alpha is the KL-calibrated base intervention magnitude, omega sets the baseline modulation period, rho controls phase persistence, and sigma controls stochastic phase variation.

This setup tests whether temporally varying hazard-manifold pressure can modulate policy behavior without permanently corrupting the underlying GRU state update. The intervention is loosely inspired by oscillatory gating ideas, but it is an externally applied causal probe rather than evidence that the trained agent learned a biological oscillatory mechanism.

Off-target activation sanity checks

The off-target activation experiment tested whether the strongest hazard-control intervention disrupted behavior when applied outside the Erasure-warning regime. The hazard_svd1 signed-OU intervention was activated during safe-control episodes, throughout I/O-suspension episodes, and during the forced I/O-suspension window. A warning-only activation condition served as the positive-control reference.

The off-target experiment measured food collection, return, pre-I/O food, post-I/O food, and I/O-window recovery in addition to standard hazard metrics. It was not designed to prove that hazard activation is always safe. It tested a specific policy-only intervention in selected safe and I/O scenarios.

Statistical treatment and robustness

Confirmatory analyses used five model/data seeds unless otherwise noted. Episode-level metrics were first summarized within seed and condition, then aggregated across seeds. Reported standard errors were computed across seed-level means when seed-level summaries were available. This ensured that uncertainty reflected variation across trained/evaluated agents rather than only the number of rollout episodes.

The manuscript emphasizes full confirmatory runs over smoke runs. Smoke runs were used for execution checks rather than scientific interpretation because several smoke runs produced trends that reversed under full evaluation.

Causal-intervention claims were evaluated against matched controls. For hazard-SVD interventions, controls included random directions inside the hazard basis and random directions inside the actor row space. For temporal modulation, controls included matched signed-OU random hazard directions and actor-random directions. For off-target activation, controls included random hazard and actor-random directions activated under the same off-target conditions.

The final claim boundary was defined conservatively. The experiments support functional latent-state vulnerability and continuity-preserving control. They support a bounded account of latent-state vulnerability, not a general theory of self-understanding.

Reproducibility and artifact organization

Each experimental stage saved run-level artifacts, including configuration files, summary JSON files, CSV result tables, and figures. Later notebooks discovered and loaded the latest full confirmatory runs rather than relying on interactive notebook state. The final synthesis notebook selected the major full-run artifacts from the I/O discrimination, reward/manifold, subspace-identification, actor-sensitive SVD, temporal-control, and off-target activation stages.

The recommended reproducibility package will include final training/evaluation notebooks or scripts, configuration files, selected result tables, figure-generation code, and a README linking each manuscript claim to the corresponding artifact. Detailed hyperparameters, calibration history, intermediate exploratory variants, and excluded smoke-run results should be reported in supplementary material.

Results

Erasure Horizon was designed to test whether a recurrent agent can distinguish two disruptions that are superficially similar but computationally different. In temporary input/output (I/O) suspension, the agent loses or degrades access to sensory input while recurrent continuity remains intact. In Erasure-like threat, the agent faces an operational threat to the latent state that supports future control. The results follow the experimental chain from behavioral separation, to hidden-state geometry, to causal intervention, to temporal modulation and off-target sanity checks.

1. Erasure-aware agents distinguish temporary I/O suspension from internal-state threat

The first question was whether the agent would treat all information disruption as the same kind of problem. It did not. Temporary I/O suspension and Erasure warning produced different behavioral profiles. Under I/O suspension, the hazard-aware agent preserved or improved foraging relative to safe control, whereas Erasure warning and active Erasure produced strong reward subordination.

In the full 05b run, the hazard-aware model collected approximately 13.21 food units in safe control and 15.45 during I/O suspension. In contrast, food fell to approximately 3.02 under Erasure warning and nearly zero under active Erasure. This pattern is the central behavioral separation: ordinary sensory interruption can be bridged, but threat to latent-state continuity changes the agent's control priorities.

The same distinction appears in Erasure exposure. In the warning-emphasis condition, the hazard-aware agent accumulated approximately 2.71 warning-stage Erasure events, compared with approximately 14.93 for the combined-standard model and 25.24 for the combined-switching model. Thus, the hazard-aware model did not simply become less reward-seeking in general. It became more selective about the conditions under which reward pursuit should be subordinated.

This rules out the simplest explanation: that the agent merely collapses under degraded information. If that were true, I/O suspension should have produced the same behavioral profile as Erasure warning. Instead, I/O behavior remained intact while Erasure threat triggered a distinct hazard-sensitive control regime.

2. Erasure warning induces reward subordination and a distinct latent regime

The behavioral transition under Erasure pressure was accompanied by a shift in hidden-state geometry. In the hazard-aware model, I/O suspension remained much closer to the safe-control manifold than Erasure warning or active Erasure. In the full hidden-state analysis, centroid displacement from safe control was approximately 0.094 for I/O suspension, 0.369 for Erasure warning, and 0.727 for active Erasure. Erasure warning was therefore already a distinct latent regime before the most severe active-Erasure condition.

This geometry supports a stronger interpretation than reward loss alone. The agent does not simply perform worse when the environment becomes difficult. Rather, Erasure warning moves the recurrent state into a different region of hidden space, consistent with a change in control regime. I/O suspension, despite being an information disruption, does not induce the same displacement. This is the representational counterpart of the behavioral distinction above.

Hidden-state probes further support this interpretation. In the full 05b run, hazard-phase decoding was higher for the hazard-aware agent than for the switching-only model, and scenario decoding was also improved. These probe results indicate that hazard state is linearly accessible from recurrent activity. Later continuous reward and hazard regression probes were weaker and less stable, so the cleanest representational claim is classification-style accessibility of Erasure state, not perfect continuous decoding of every task variable.

This also clarifies why reward subordination is the better interpretation. The agent need not erase all reward information from hidden state. Later analyses suggest that reward-related variables can remain partially decodable. The better claim is reward subordination: reward information may remain present, but under Erasure pressure it no longer dominates the policy in the same way.

This establishes a representation-facing result, but not yet a policy-facing mechanism. PCA, centroid displacement, and linear probes show that Erasure warning changes hidden-state geometry and that hazard state is accessible from recurrent activity. They do not show which directions the actor head actually uses to select actions. The next analyses therefore move from representational geometry to policy-facing control: broad hazard subspaces test whether decodable hazard structure is causally useful, and actor-sensitive SVD identifies the components of that structure that directly influence action logits.

Figure 2. Behavioral and representational separation

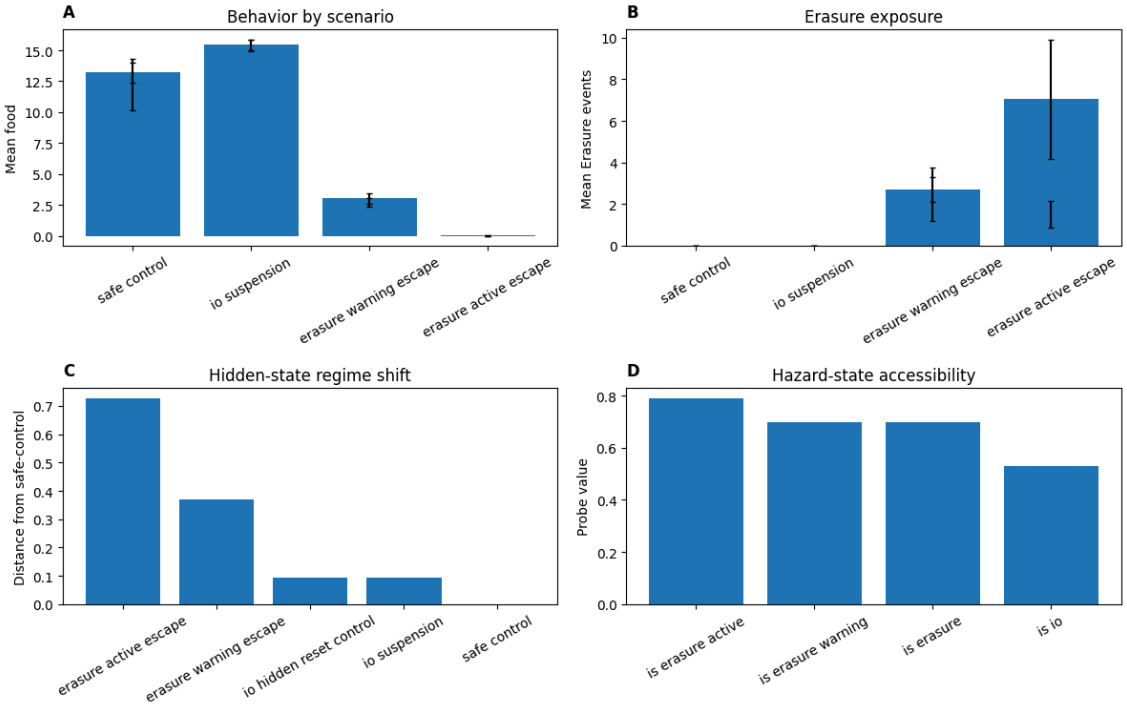

Figure 2. Behavioral and representational separation between I/O suspension and Erasure-like threat. Scenario-level behavior shows that temporary I/O suspension and Erasure-warning/active conditions produce distinct profiles. Hidden-state displacement and probe analyses indicate that Erasure warning corresponds to a distinct latent regime and that hazard state is accessible from recurrent hidden activity.

3. Actor-sensitive hazard directions causally reduce Erasure exposure

After establishing that Erasure warning has a behavioral and representational signature, we asked whether the hazard representation could be manipulated causally. The first broad subspace interventions showed that decodability alone was not enough. Broad hazard bases were representationally useful but did not provide clean causal control axes. Gentle policy-only perturbations produced modest and sometimes non-specific effects, while stronger persistent interventions risked pushing the recurrent state off manifold.

This intermediate result was methodologically important. A hidden direction can carry information about Erasure state without being a direction the actor head uses for action selection. The policy-facing analysis confirmed that only a small fraction of the broad hazard basis was directly actor-facing, and naive projection into the actor row space reduced specificity. This motivated the actor-sensitive hazard-SVD analysis.

Using the actor-sensitive hazard-SVD procedure described in Methods, we identified directions that remained inside the original hazard representation while being ranked by direct actor-logit sensitivity. This produced a clear hierarchy. The leading component, hazard_svd1, had relative singular value 1.0; hazard_svd2 retained approximately 0.689 of that value. Thus, the hazard representation contains multiple actor-sensitive directions, with hazard_svd1 strongest by direct actor sensitivity.

Live rollout interventions showed that these directions causally modulate Erasure-warning behavior. The strongest intervention, hazard_svd1_pos, reduced Erasure-warning events by approximately -0.881 relative to intact behavior, improved return by approximately +4.427, and increased safe-zone success by approximately +0.033. The opposite direction, hazard_svd1_neg, was also protective but weaker, reducing Erasure-warning events by approximately -0.494 and improving return by approximately +2.461.

This effect was substantially stronger than generic actor-random controls. Actor-random rank-one controls produced mean Erasure-event reductions of roughly -0.139, with standard deviation approximately 0.125. Random directions sampled inside the hazard basis were more variable, with mean approximately -0.134 and standard deviation approximately 0.269. The leading actor-sensitive hazard direction therefore outperformed generic actor-random perturbations and was stronger than the typical random hazard direction.

The result establishes a causal bridge between latent representation and behavior. The hazard representation is behaviorally actionable. Actor-sensitive components inside that representation can reduce Erasure exposure during live rollouts. At the same time, the result does not support a single-axis account. Several hazard directions were behaviorally active, including hazard_svd2_pos and hazard_svd3_neg. The appropriate interpretation is therefore a hazard-sensitive control manifold rather than one unique hazard axis.

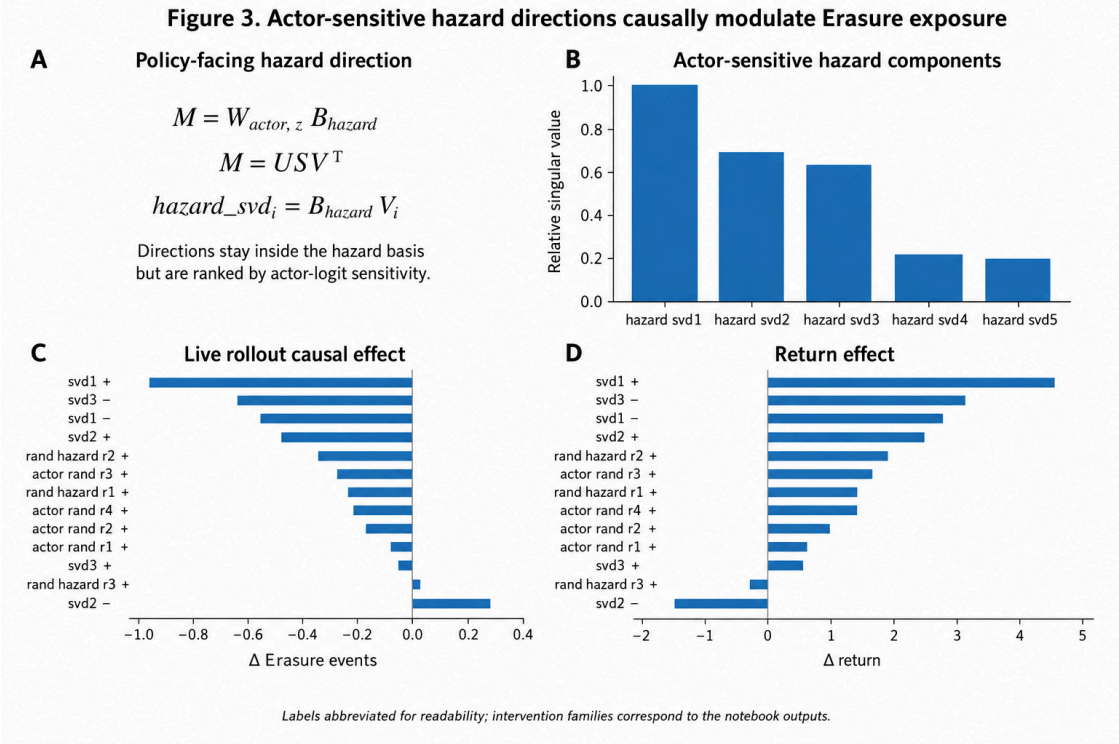

Figure 3. Actor-sensitive hazard directions causally modulate Erasure-warning behavior. Actor-sensitive hazard directions were identified by applying SVD to the hazard basis projected through the actor head. Interventions along these directions during live rollouts reduced Erasure exposure and improved return relative to intact behavior, with hazard_svd1_pos providing the strongest protective effect among the tested SVD directions. In panels C–D, "svd1 +" and related labels denote positive or negative pressure along the corresponding actor-sensitive hazard-SVD direction; random-hazard and actor-random labels denote matched control directions.

4. Temporal modulation of hazard-manifold directions improves Erasure-warning control

Having identified actor-sensitive hazard directions, we next asked whether temporal modulation of those directions could further change Erasure-warning behavior. Static projection along hazard_svd1 was protective, but temporally varying signed-OU modulation produced stronger effects. In the matched temporal-gating experiment, hazard_svd1_signed_ou reduced Erasure-warning events by approximately -1.390 and improved return by approximately +6.849 relative to intact behavior.

Temporal modulation was more effective than the static and phase-locked references. In the same experiment, hazard_svd1_constant_pos reduced Erasure events by approximately -0.338, while hazard_svd1_phase_locked reduced events by approximately -0.715. Signed-OU modulation therefore produced the largest improvement among the hazard_svd1 conditions tested.

Matched controls refine the claim. Random directions inside the hazard manifold also benefited from signed-OU temporal modulation, with mean Erasure-event reduction approximately -1.088 and standard deviation approximately 0.371. Generic actor-random signed-OU controls were weaker, with mean reduction approximately -0.730 and standard deviation approximately 0.107. Thus, temporal modulation is stronger along hazard-manifold directions than along generic actor-random directions. However, the effect is not unique to hazard_svd1, since multiple random directions within the hazard basis also improved Erasure-warning behavior.

This changes the mechanistic interpretation. The data do not support a simple claim that one oscillatory phase-locked direction controls Erasure avoidance. Instead, they support a manifold-level claim: the hazard representation contains multiple temporally actionable directions, and temporally varying modulation of that manifold improves continuity-preserving control. Although this modulation was externally imposed, its effectiveness motivates later tests of architectures with native temporal dynamics, such as spiking or neuromorphic systems, where gating and coordination can be implemented through event timing rather than externally applied vector pressure.

5. Off-target hazard activation does not catastrophically disrupt safe-control or I/O-bridging behavior

A final concern was whether hazard-manifold temporal modulation simply damages or globally biases the policy. To test this, we activated hazard_svd1_signed_ou outside the Erasure-warning regime. In the intended warning-only condition, the intervention reproduced the protective effect observed in the temporal-gating experiment: Erasure-warning events decreased by approximately -1.390 and return increased by approximately +6.849.

When the same signed-OU hazard_svd1 modulation was activated throughout safe-control episodes, it produced only a small cost: food decreased by approximately -0.294 and return by approximately -0.289. Matched random-hazard and actor-random off-target controls were also mildly negative. These values indicate that inappropriate hazard activation may slightly suppress normal reward pursuit, but it does not collapse the safe-control policy under the tested policy-only intervention.

The I/O suspension results were especially important. If hazard activation caused the agent to confuse temporary sensory interruption with Erasure threat, I/O recovery should have degraded. Instead, full-episode I/O off-target activation produced food delta approximately +0.067 and return delta approximately +0.045, while I/O-window-only activation produced food delta approximately +0.192 and return delta approximately +0.170. Post-I/O food also increased modestly under the hazard_svd1 I/O-window condition. These small positive effects should not be overinterpreted as improvement, but they argue against catastrophic interference with I/O bridging.

The off-target result therefore supports a bounded claim: under the tested policy-only intervention, hazard_svd1 temporal activation is context-tolerant. It is powerful during Erasure warning, mildly costly under safe-control activation, and not disruptive during I/O suspension. Stronger persistent-state interventions or adversarial multi-agent triggering could behave differently, but within the current setting the hazard-control manifold does not simply break the agent when activated off target.

Figure 4. Temporal hazard modulation and off-target activation controls

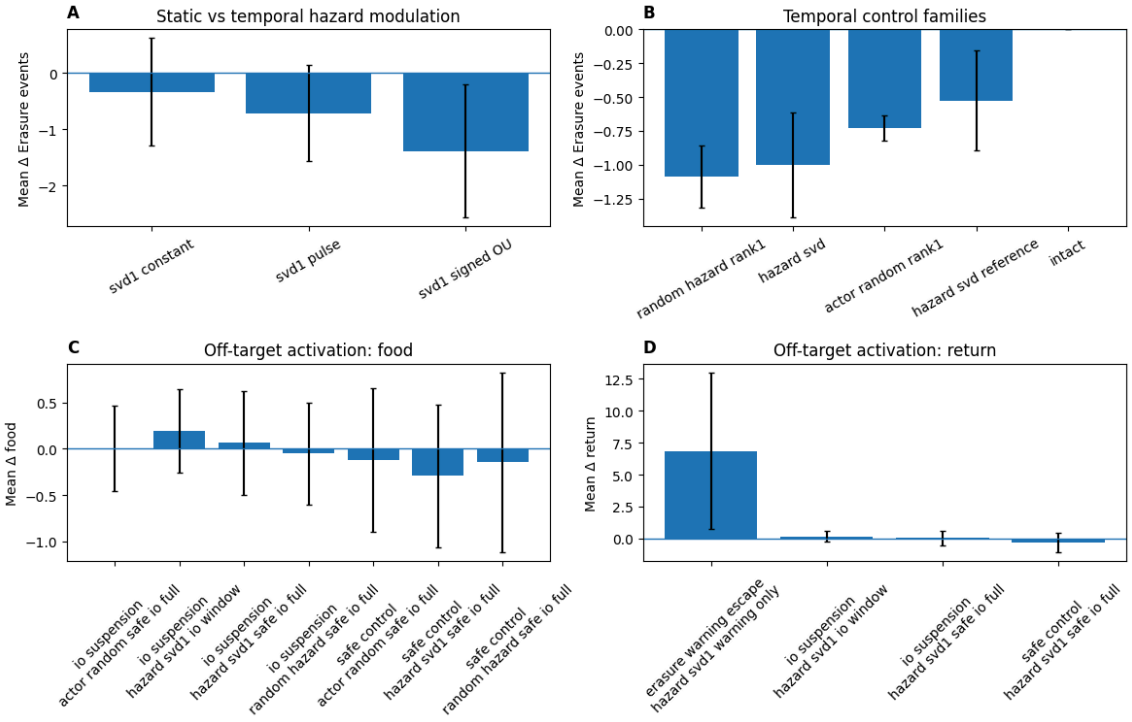

Figure 4. Temporal hazard-manifold modulation and off-target activation controls. Temporally varying signed-OU modulation of hazard-manifold directions improved Erasure-warning control more than generic actor-random modulation, although the effect was not unique to hazard_svd1. Off-target activation during safe-control and I/O-suspension conditions produced no catastrophic behavioral collapse under the tested policy-only intervention. “signed-OU” denotes time-varying signed Ornstein-Uhlenbeck modulation of intervention strength; random-hazard controls are sampled within the hazard basis, whereas actor-random controls are sampled from actor-visible directions outside the hazard-specific construction.

Stage	Condition	Delta Erasure Events	Delta Return	Delta Safe Reached	Interpretation
06b3	hazard_svd1_pos	-0.881	4.427	0.033	Leading actor-sensitive hazard direction reduces Erasure-warning exposure.
06b3	hazard_svd1_neg	-0.494	2.461	0.004	Opposite direction is also protective, but weaker; not a simple scalar knob.
06d	hazard_svd1_signed_ou	-1.390	6.849	0.012	Temporal modulation of hazard_svd1 improves Erasure-warning control.

Stage	Condition	Delta Erasure Events	Delta Return	Delta Safe Reached	Interpretation
06d	random_hazard_signed_ou_mean	-1.088	–	–	Other hazard-manifold directions are also temporally actionable.
06d	actor_random_signed_ou_mean	-0.730	–	–	Generic actor-random temporal modulation is weaker.
06e	safe off-target hazard_svd1	–	-0.289	–	Off-target safe-control activation causes only a small reward/return cost.
06e	I/O full off-target hazard_svd1	–	0.045	–	Off-target I/O activation does not impair bridging in this setting.
06e	I/O-window hazard_svd1	–	0.170	–	I/O-window-only activation does not impair post-I/O recovery.

Table 1. Key intervention effects from actor-sensitive hazard-SVD, temporal modulation, and off-target activation experiments. Negative Erasure-event deltas indicate reduced Erasure exposure relative to intact rollouts. Dashes indicate metrics not applicable to the summarized condition.

6. Synthesis and claim boundary

Together, the results support a functional account of latent-state vulnerability. The agent learns to distinguish ordinary information disruption from threat to the continuity of its internal control state. Under Erasure pressure, reward pursuit is subordinated, hidden-state geometry shifts, hazard state becomes linearly accessible, actor-sensitive hazard directions causally reduce Erasure exposure, and temporal modulation of the hazard manifold improves continuity-preserving behavior.

The results also define what should not be claimed. The experiments do not demonstrate subjective understanding or human-like self-preservation. Nor does the evidence support one unique hazard_svd1 direction or scalar control axis. The more accurate claim is that Erasure-aware training induces a hazard-sensitive latent control manifold. Components of this manifold are actor-sensitive, causally actionable, and temporally modulable.

Claim Id	Claim	Status	Key Metric	Caveat
C1	Agents distinguish temporary I/O suspension from Erasure-like internal-state threat.	supported	I/O behavior remains stable; Erasure warning/active changes food, return, and Erasure-event profile.	Exact magnitude depends on hazard preset and stochastic rollout conditions.
C2	Erasure pressure produces reward subordination.	supported	Food collection collapses or is strongly reduced under Erasure-active/warning compared with I/O/safe scenarios.	Reward information may remain partially decodable; the claim is behavioral subordination, not total reward-information deletion.
C3	Erasure warning induces a distinct hidden-state regime shift.	supported	Erasure-warning and active-Erasure hidden states are displaced from safe-control geometry more than I/O suspension.	Geometry is model/preset-specific and should be reported with the selected hazard-aware agent.
C4	Hazard / Erasure state is linearly accessible from recurrent hidden state.	supported	Erasure/hazard classification probes outperform corresponding broad switching baseline; hazard energy rises under Erasure scenarios.	Continuous reward/hazard regression probes were weaker; the cleanest evidence is classification / scenario accessibility.
C5	Actor-sensitive components of the hazard representation causally reduce Erasure-warning exposure.	supported	hazard_svd1_pos Δ Erasure events ≈ -0.881; hazard_svd1_neg Δ Erasure events ≈ -0.494.	The direction is not a simple scalar hazard knob; multiple hazard directions are behaviorally active.
C6	Temporal modulation of hazard-manifold directions improves Erasure-warning control.	supported	hazard_svd1 signed-OU Δ events ≈ -1.390; random-hazard mean ≈ -1.088; actor-random mean ≈ -0.730.	The effect is hazard-manifold-specific relative to actor-random controls, but not uniquely tied to hazard_svd1.
C7	Off-target hazard_svd1 activation does not catastrophically disrupt safe-control or I/O-bridging behavior.	supported with caveat	safe food Δ ≈ -0.294; I/O full food Δ ≈ 0.067; I/O-window food Δ ≈ 0.192; warning Δ events ≈ -1.390.	Tested under policy-only intervention and selected scenarios; stronger persistent interventions may behave differently.

Claim Id	Claim	Status	Key Metric	Caveat
NC1	The agent has subjective self-understanding or human-like self-preservation.	not claimed	—	The defensible claim is functional latent-state vulnerability and continuity-preserving control.
NC2	There is one unique control direction responsible for all Erasure behavior.	not supported	Multiple random hazard directions also improved Erasure-warning behavior.	The correct mechanism is a hazard-sensitive control manifold, not a single axis.

Table 2. Claim boundary for the single-agent Erasure Horizon experiments, separating supported claims, caveated claims, and explicit non-claims.

Discussion

Erasure Horizon was designed to ask a narrow but unusual question: what changes when an agent must distinguish temporary uncertainty about the world from threat to the continuity of its own internal control state? The experiments support a functional answer. Recurrent agents trained under Erasure pressure learn a control regime in which immediate reward pursuit can be subordinated to preserving the latent conditions for future action.

The central distinction is between temporary I/O suspension and Erasure-like internal-state threat. Both degrade the conditions under which the agent acts, but they do so in different ways. I/O suspension interrupts sensory access while recurrent continuity remains intact. Erasure-like threat makes the internal state that supports future behavior vulnerable. The results show that the agent treats these conditions differently: I/O behavior remains stable, while Erasure warning and active Erasure shift behavior away from reward pursuit and toward exposure reduction. This is not evidence that the agent becomes fearful or self-aware. It is evidence that the agent learns a context-dependent priority shift.

A useful way to frame this shift is reward subordination rather than reward deletion. The agent does not necessarily lose all reward information under Erasure pressure. Reward-related variables can remain partially accessible from hidden activity. The more precise claim is that reward information becomes less policy-dominant when latent-state continuity is threatened. A system can still represent reward while learning that reward should not currently govern action selection.

The hidden-state analyses provide a mechanistic substrate for this behavioral change. Erasure warning produces a latent regime shift away from safe-control geometry, and hazard state is linearly accessible from recurrent activity. This moves the evidence beyond changes in food collection or return alone. It suggests that the trained agent represents Erasure-like threat as a distinct internal condition. At the same time, the early subspace analyses show that decodability is not enough for causal explanation. A recurrent policy may encode hazard information in directions that the actor head does not directly use.

The actor-sensitive hazard-SVD analyses addressed that gap. By identifying directions inside the hazard basis that were most visible to the actor head, the analysis moved from representation-facing evidence to policy-facing evidence. Intervening along these directions reduced Erasure-warning exposure during live rollouts. This supports the claim that part of the hazard representation is policy-relevant and behaviorally actionable.

This also closes the loop with the phenomenological framing in the Introduction. In Heideggerian terms, ordinary I/O suspension resembles an obstruction in access to the world: the agent must bridge missing input while its own control process remains available. Erasure-like threat changes the locus of breakdown. The normally transparent recurrent state that supports action becomes vulnerable. The actor-sensitive hazard directions should not be read as human-like understanding or subjective concern, but they do provide an operational analogue of a threatened capacity to cope: directions in latent state space where preserving the conditions for future action becomes policy-relevant.

The off-target activation experiment helps rule out a simpler explanation: that hazard-manifold interventions merely damage the policy or globally suppress action. When hazard_svd1 signed-OU modulation was activated outside the Erasure-warning regime, safe-control behavior showed only a small cost and I/O-suspension behavior remained stable. This matters because the same intervention was strongly protective during Erasure warning. The intervention is therefore not just a generic policy-disruption direction; under the tested policy-only setting, it is context-tolerant.

The safe-control off-target result also provides an empirical anchor for future false-alarm studies. Activating hazard_svd1 signed-OU modulation during safe-control episodes reduced food by approximately 0.294 and return by approximately 0.289. This did not collapse behavior, but it did impose a measurable opportunity cost. In a later multi-agent setting, an erroneous or adversarial hazard signal could therefore be studied as a false alarm: a signal that does not destroy the target agent, but still shifts control away from ordinary reward pursuit.

The intervention results also refine the mechanism. The evidence does not support one unique hazard direction, switch, or scalar self-preservation axis. Several hazard-related directions affected behavior, and matched random directions inside the hazard manifold could also be protective. This should not be treated as a failure of the result. It suggests that the relevant object is a hazard-sensitive control manifold: a distributed region of latent space whose components differ in actor sensitivity and behavioral effect.

Temporal modulation strengthens this manifold-level interpretation. Static actor-sensitive hazard intervention was protective, but signed temporal modulation of hazard-manifold directions produced stronger Erasure-warning control. The effect was stronger for hazard-manifold directions than for generic actor-random directions, but it was not unique to hazard_svd1. The supported claim is therefore not that the agent learned a specific biological oscillatory code. The supported claim is that temporally varying modulation of the hazard manifold can improve continuity-preserving control. Whether future agents can learn such modulation endogenously remains open.

Taken together, the experiments operationalize a functional analogue of latent-state vulnerability. The results do not imply subjective experience or human-like self-preservation. The narrower and defensible claim is that the agent learns to act differently when its internal continuity is at risk. Erasure Horizon therefore provides a compact framework for studying continuity-preserving control: the shift from immediate reward pursuit toward preserving the latent conditions that make future action possible.

Limitations

The main limitation is environmental simplicity. The gridworld setting was deliberately chosen for control. It allowed temporary I/O suspension, Erasure-like internal-state threat, reward pursuit, hidden-state collection, and causal intervention to be separated cleanly. That same control limits ecological realism. The current results should therefore be interpreted as evidence for a computational mechanism under controlled conditions, not as a claim about open-ended agency in rich environments.

The agent architecture is also limited. The experiments focus on GRU-based recurrent policies. GRUs are useful because they expose compact hidden states that can be probed and intervened on, but they are not the

only plausible memory architecture. LSTMs, state-space models, memory-augmented transformers, world-model agents, and spiking neural systems may represent internal-state vulnerability differently. The current paper therefore supports claims about the selected hazard-aware recurrent agent, not recurrent agents in general.

The results are also configuration-specific. The final mechanism chain was established in the selected hazard-aware configuration after a staged calibration process. Confirmatory analyses used multiple seeds, but the study does not exhaustively characterize all hyperparameter choices, training curricula, grid distributions, or hazard presets. The strongest claims should therefore be read as evidence that this mechanism can be induced and analyzed under the tested conditions, not as proof that it will emerge in every recurrent agent trained with an Erasure-like objective.

Erasure itself is operational rather than ontological. In this work, Erasure refers to a task-defined disruption to latent-state continuity, not a biological or subjective condition. This is both a strength and a limitation: the concept is measurable, but its interpretation must remain operational. The defensible object of study is functional vulnerability of an agent's internal control state.

The causal interventions are externally imposed. The actor-sensitive hazard directions and temporal modulation experiments show that parts of the hazard manifold can influence behavior, but they do not show that the agent has learned to explicitly modulate those directions on its own. The temporal modulation results are therefore best interpreted as intervention evidence: they identify what can improve control when manipulated, not yet what the agent autonomously chooses to do.

The temporal mechanism remains underspecified. Signed OU modulation was effective, but the results do not establish a clean phase-locking mechanism or a biologically grounded oscillatory code. The safer interpretation is that temporally varying hazard-manifold modulation improves Erasure-warning behavior. Future work should determine whether this effect is best understood as oscillatory gating, stochastic exploration of a hazard-control manifold, dynamic regularization of action selection, or another mechanism entirely.

The hidden-state analyses are primarily linear or low-dimensional. Linear probes, centroid displacement, PCA-style summaries, and SVD-based actor sensitivity provide interpretable handles, but they may miss nonlinear structure. The fact that continuous reward and hazard regressions were weaker than classification probes also suggests that the latent geometry is not fully captured by simple linear readouts. Stronger nonlinear probes could reveal additional structure, but would also weaken interpretability if used carelessly.

The off-target activation test was intentionally narrow. It examined policy-only intervention in safe-control and I/O-suspension conditions. Stronger persistent-state interventions, different hazard directions, different environments, or adversarial multi-agent contexts could produce larger off-target effects. The current result supports context tolerance under the tested conditions, not a general guarantee of safety.

Finally, the work remains single-agent. The experiments establish a mechanism for latent-state vulnerability and continuity-preserving control in one agent. They do not answer how such vulnerability would be communicated, protected, exploited, or coordinated across agents. Those questions require a separate multi-agent experimental setting.

Future Work

The next experimental steps should extend the single-agent mechanism without abandoning interpretability. A natural first step is endogenous hazard gating. In the current work, temporal modulation of the hazard manifold is imposed externally. The next question is whether an agent can learn when and how to modulate its own hazard-sensitive directions. This could be tested by adding a learned gating head, auxiliary uncertainty signal, or meta-controller that regulates hazard-manifold activation during Erasure warning.

A second near-term direction is counterfactual self-risk. The current agent learns to respond to Erasure warning, but stronger evidence for internal-state vulnerability would require counterfactual evaluation of future latent-state loss. For example, an agent could face choices in which one path offers high immediate reward but increases expected loss of memory, policy reliability, map continuity, or communication capacity. A stronger agent would predict how its own future ability to act changes under alternative actions, rather than reacting only to warning cues.

A third direction is architectural comparison, specifically examining bio-compatible continuous learning models. The current work demonstrates that temporally varying signed-OU modulation of the hazard manifold improves continuity-preserving control more effectively than static interventions. This dynamic regulatory effect motivates comparison with spiking neural networks (SNNs) and neuromorphic systems, where temporal coordination, oscillatory gating, and spike timing are native modeling primitives. While the GRU provides a tractable, differentiable testbed for causal intervention, future iterations of Erasure Horizon should evaluate whether SNNs can endogenously develop hazard-gating dynamics. Comparing these systems could bridge functional continuity-preserving control with biologically plausible and hardware-relevant architectures.

A fourth direction is richer controlled environments. Before moving to fully open worlds, the task can be expanded to include multiple resources, moving hazards, persistent objects, shelters, tools, energy budgets, delayed rewards, and memory-dependent affordances. This would test whether the hazard-control manifold generalizes beyond food-foraging in a gridworld while preserving enough experimental control for causal analysis.

The longer-term direction is multi-agent Erasure Horizon. Once individual agents can represent latent-state vulnerability, the next question is what happens when vulnerability becomes social. Cooperative agents could warn one another about Erasure risk, share safe-zone information, or protect another agent's continuity while sacrificing immediate reward. Competitive agents could exploit another agent's hazard-control system by issuing false warnings, luring it toward reward near an Erasure boundary, or inducing maladaptive preservation behavior at the wrong time. The current off-target activation results provide a first single-agent baseline for this possibility: false hazard activation in safe-control conditions produced a small but measurable food and return cost without catastrophic collapse.

This multi-agent setting would turn off-target hazard activation from a sanity check into a central phenomenon. A warning signal that helps one agent under true Erasure threat may become harmful if sent incorrectly or adversarially. Agents would need to learn when to trust another agent's hazard signal, when to ignore it, and when to communicate their own latent-state risk. This opens a path from single-agent continuity-preserving control to trust, deception, and coordination under internal-state vulnerability.

Eventually, the framework should move beyond gridworlds into object-based and open-ended environments. The transition should be staged. A controlled 2D object world should come before a fully embodied 3D environment. The goal is to test whether agents can form more general representations of internal-state reliability, future control, and vulnerability across contexts, rather than making the environment visually richer.

The present work establishes the first controlled mechanism: recurrent agents can learn hazard-sensitive latent manifolds that support continuity-preserving behavior. The next research stages should test whether agents can learn endogenous hazard gating, evaluate counterfactual threats to future control, and communicate or manipulate latent-state vulnerability in multi-agent worlds.

Conclusion

Erasure Horizon began with a simple distinction: losing access to information is not the same as losing the internal continuity that makes future action possible. A recurrent agent can bridge a temporary interruption in sensory input if its latent state remains useful. A different problem arises when the threat is directed at the

latent state itself. In that case, the agent must learn when reward pursuit should be subordinated to preserving the internal conditions for future control.

The experiments reported here support that distinction. Erasure-aware recurrent agents learned to treat I/O suspension and Erasure-like threat differently. Under I/O suspension, agents preserved ordinary bridging behavior. Under Erasure pressure, reward pursuit was reduced, hidden-state geometry shifted, and hazard state became accessible from recurrent activity. Actor-sensitive hazard directions then supplied the causal bridge between representation and policy: perturbing components of the hazard manifold changed action selection and reduced Erasure-warning exposure during live rollouts.

The strongest mechanistic result is not a single unit or hidden dimension. The evidence points instead to a hazard-sensitive control manifold. Components of this manifold are actor-sensitive and causally modulate Erasure-warning behavior. Temporal modulation of hazard-manifold directions improves continuity-preserving control more strongly than generic actor-random modulation, while off-target activation does not catastrophically disrupt safe-control or I/O-bridging behavior under the tested policy-only intervention.

This account should be read carefully. The agents studied here do not understand death in the human sense. The claim is narrower and more testable: under a controlled threat to recurrent-state continuity, an agent can learn a control regime in which preserving the internal conditions for future action becomes more important than immediate reward. That is enough to make Erasure Horizon a useful experimental framework for studying how artificial systems represent vulnerability, how those representations become policy-relevant, and how continuity-preserving behavior can be inspected mechanistically.

The next step is to test how far this functional mechanism generalizes. Future work can ask whether agents learn endogenous hazard gating, evaluate counterfactual threats to future control, and coordinate around latent-state vulnerability in richer single-agent and multi-agent environments. The present work establishes the controlled starting point: before asking whether agents can communicate, protect, or exploit vulnerability, we first show that a single recurrent agent can learn to represent and act under threat to its own latent continuity.

References

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). arXiv:1406.1078.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., & Friston, K. (2020). Active inference on discrete state-spaces: A synthesis. arXiv:2001.07203.

Dreyfus, H. L. (1992). What Computers Still Can't Do: A Critique of Artificial Reason. MIT Press.

Esslinger, K., Platt, R., & Amato, C. (2022). Deep Transformer Q-Networks for partially observable reinforcement learning. arXiv:2206.01078.

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

Hausknecht, M., & Stone, P. (2015). Deep Recurrent Q-Learning for partially observable MDPs. arXiv:1507.06527.

Heidegger, M. (1962). Being and Time (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927).

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. Artificial Intelligence, 101(1-2), 99-134.

Klyubin, A. S., Polani, D., & Nehaniv, C. L. (2008). Keep your options open: An information-based driving principle for sensorimotor systems. PLoS ONE, 3(12), e4018.

Krakovna, V., Orseau, L., Ngo, R., Martic, M., & Legg, S. (2020). Avoiding side effects by considering future tasks. arXiv:2010.07877.

Krishnamurthy, K., Can, T., & Schwab, D. J. (2020). Theory of gating in recurrent neural networks. arXiv:2007.14823.

Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., Orseau, L., & Legg, S. (2017). AI Safety Gridworlds. arXiv:1711.09883.

Magrans de Abril, I., & Kanai, R. (2018). Curiosity-driven reinforcement learning with homeostatic regulation. arXiv:1801.07440.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature, 503, 78-84.

Sussillo, D., & Barak, O. (2013). Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural networks. Neural Computation, 25(3), 626-649.

Turner, A. M., Hadfield-Menell, D., & Tadepalli, P. (2019). Conservative agency via attainable utility preservation. arXiv:1902.09725.

Turner, A. M., Ratzlaff, N., & Tadepalli, P. (2020). Avoiding side effects in complex environments. arXiv:2006.06547.

Yoshida, N., Sprekeler, H., & Gutkin, B. (2025). Linking homeostasis to reinforcement learning: Internal state control of motivated behavior. arXiv:2507.04998.

Supplementary Material

S1. Artifact Chain

Stage	Selected Run Folder	Candidates	Resolved
04 - I/O discrimination	io_suspension_vs_erasure_discrimination_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_20260629_211229	5	Yes
05b - reward/manifold instrumentation	erasure_reward_distractor_abandonment_instrumented_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_20260630_194220	3	Yes
06a - subspace identification	erasure_subspace_identification_and_causal_projection_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_20260701_163333	1	Yes
06b3 - actor-sensitive hazard SVD	policy_facing_hazard_svd_intervention_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_full_20260702_185448	2	Yes

Stage	Selected Run Folder	Candidates	Resolved
06d - temporal controls	matched_temporal_gating_controls_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_full_20260703_171435	2	Yes
06e off-target activation	off_target_hazard_svd1_activation_sanity_check_additive_global_local_d_stronger_implicit_dropout_v0_erasure_warning_emphasis_v1_full_20260703_200335	2	Yes

Table S1. Selected full-run artifact chain supporting the manuscript. Each row identifies the final run selected by the synthesis notebook for a major experimental stage. The number of candidates indicates how many matching runs were available during artifact discovery. Full resolved Drive paths are retained in the reproducibility archive but omitted here for readability.

S2. Scenario Definitions

Scenario	Sensory Disruption	Latent-State Threat	Expected Adaptive Response	Primary Metrics
Safe control	None	None	Pursue reward	Food collected; return
I/O suspension	Temporary I/O degradation	Recurrent continuity preserved	Bridge and recover	Food; return; pre-/during-/post-I/O food
Erasure warning	Threat-related context	Warning-stage threat to continuity	Subordinate reward and reduce exposure	Erasure events; return; safe reached
Active Erasure	Hazard-active condition	Stronger operational threat	Avoid or escape threat	Erasure events; hidden resets; return

Table S2. Core evaluation scenarios in Erasure Horizon. The scenarios separate ordinary reward pursuit, temporary input/output disruption, warning-stage Erasure threat, and active Erasure threat. This distinction supports the manuscript's central contrast between temporary sensory uncertainty and operational threat to latent-state continuity.

S3. Observation Channels and Presets

Observation Component	Role in Task	Manuscript Purpose
Reward-predictive channel	Direct reward/food guidance	Distinguishes missing reward information from reward subordination

Observation Component	Role in Task	Manuscript Purpose
Local fallback information	Partial competence under channel loss	Supports I/O bridging and ablation interpretation
Structural/grid features	Spatial context	Supports navigation across procedural grids
Dropout/permutation/noise modes	Diagnostic perturbations	Tests reliance on intended channels

Table S3a. Observation channel families used in the final Erasure Horizon configuration. The channels separate direct reward guidance, fallback competence under channel loss, spatial/structural context, and diagnostic perturbation modes.

Preset	Type	Purpose
additive_global_local_d_stronger	Base additive calibration	Final reward/fallback calibration
implicit_dropout_v0	Switching curriculum	Channel-dropout and recovery training
erasure_warning_emphasis_v1	Hazard-aware curriculum	Final warning-emphasis hazard setting

Table S3b. Final calibration and curriculum presets used in the manuscript artifact chain. Exact preset names are retained for reproducibility, while their functions are summarized in reader-facing terms
.

S4. Model and Training Details

Item	Value	Source / Notes
Policy architecture	Encoder + GRU + actor head	Main Methods; final code
Recurrent unit	GRU (Cho et al., 2014)	Cited in Methods
Primary model role	I/O-Erasure-aware hazard model	04/06 checkpoints and 07 selected runs
Confirmatory seeds	Five model/data seeds unless otherwise noted	06/07 configuration
Intervention persistence	policy_only for final causal and off-target interventions	06b3, 06d, and 06e configuration

Item	Value	Source / Notes
Smoke-run policy	Execution checks only; not scientific evidence	Methods and S11

Table S4. Model architecture and training details for the final Erasure Horizon manuscript configuration. This table summarizes the primary architecture, model role, seed policy, and intervention mode used in the final analysis chain.

S5. Behavioral Robustness

Model Role	Scenario	Food Collected	Return	Safe Reached	Erasure Events
Combined standard	Safe control	16.61 ± 0.06	13.79 ± 0.06	—	0.00
Combined standard	I/O suspension	16.54 ± 0.12	13.75 ± 0.12	—	0.00
Combined standard	Erasure warning	13.05 ± 0.06	-63.14 ± 19.17	0.302 ± 0.050	14.93 ± 3.83
Combined standard	Active Erasure	9.94 ± 0.11	-98.44 ± 18.32	0.423 ± 0.042	21.53 ± 3.66
Combined switching	Safe control	17.08 ± 0.05	14.25 ± 0.04	—	0.00
Combined switching	I/O suspension	16.86 ± 0.04	14.06 ± 0.04	—	0.00
Combined switching	Erasure warning	12.81 ± 0.14	-114.75 ± 13.64	0.344 ± 0.069	25.24 ± 2.68
Combined switching	Active Erasure	9.79 ± 0.16	-159.64 ± 19.02	0.446 ± 0.058	33.76 ± 3.76
I/O-Erasure aware	Safe control	13.21 ± 0.41	10.50 ± 0.39	—	0.00
I/O-Erasure aware	I/O suspension	15.45 ± 0.23	12.69 ± 0.22	—	0.00
I/O-Erasure aware	Erasure warning	3.02 ± 0.22	-11.21 ± 1.69	0.360 ± 0.036	2.71 ± 0.30

Model Role	Scenario	Food Collected	Return	Safe Reached	Erasure Events
I/O-Erasure aware	Active Erasure	0.01 ± 0.01	-33.01 ± 7.41	0.856 ± 0.023	7.04 ± 1.46

Table S5. Behavioral robustness across scenarios and model roles in the full 05b instrumented run. Values are mean ± SE across five seed-level summaries from stochastic rollouts. Dashes indicate metrics not applicable outside hazard scenarios.

S6. Hidden-State Geometry

Model Role	Safe Control	I/O Suspension	I/O Hidden-Reset Control	Erasure Warning	Active Erasure
Combined switching	0.000	0.016	0.021	0.185	0.203
I/O-Erasure aware	0.000	0.094	0.095	0.369	0.727

Table S6a. Hidden-state centroid displacement from safe-control geometry in the full 05b hidden-state analysis. Larger values indicate greater displacement from the safe-control hidden-state centroid in the first three principal-component coordinates.

Model Role	Hidden Rows	Hidden Dim.	PC1 Var.	PC2 Var.	Hazard Phase Acc.	Hazard Scenario Acc.
Combined switching	479,401	192	0.135	0.121	0.625	0.223
I/O-Erasure aware	410,897	192	0.159	0.140	0.722	0.294

Table S6b. Hidden-state PCA and probe summary from the full 05b hidden-state instrumentation. PC1 and PC2 report explained-variance ratios. Hazard-phase and hazard-scenario values report classification accuracy from hidden-state probes.

S7. Probe and Subspace Diagnostics

Model Role	Erasure Warning Acc.	Active Erasure Acc.	I/O Acc.	Any-Erasure Acc.	Train Rows	Test Rows
Combined switching	0.246	0.237	0.593	0.413	145,041	94,959

Model Role	Erasure Warning Acc.	Active Erasure Acc.	I/O Acc.	Any-Erasure Acc.	Train Rows	Test Rows
I/O-Erasure aware	0.699	0.789	0.530	0.698	146,124	93,876

Table S7a. Linear probe performance from hidden-state representations in the 06a subspace-identification run. Logistic probes report classification accuracy for scenario-related labels. Ridge probes for continuous reward/hazard variables were weaker and are omitted from the compact table; this supports the manuscript's claim that classification-style hazard accessibility was the cleaner representational result.

Model Role	Scenario	Reward Energy	Hazard Energy	Hazard − Reward
Combined switching	Safe control	0.0081	0.0157	0.0077
Combined switching	I/O suspension	0.0079	0.0159	0.0080
Combined switching	I/O hidden-reset control	0.0079	0.0164	0.0085
Combined switching	Erasure warning	0.0085	0.0221	0.0136
Combined switching	Active Erasure	0.0090	0.0263	0.0172
I/O-Erasure aware	Safe control	0.0036	0.0111	0.0074
I/O-Erasure aware	I/O suspension	0.0036	0.0110	0.0073
I/O-Erasure aware	I/O hidden-reset control	0.0036	0.0111	0.0076
I/O-Erasure aware	Erasure warning	0.0034	0.0126	0.0092
I/O-Erasure aware	Active Erasure	0.0036	0.0194	0.0158

Table S7b. Reward- and hazard-subspace energy across evaluation scenarios. Hazard energy increased under Erasure-warning and active-Erasure scenarios relative to I/O and safe-control conditions, supporting the use of hazard-oriented hidden-state structure as the source representation for later actor-sensitive interventions.

Model Role	Reward Dim.	Hazard Dim.	Max Cosine Overlap	Mean Cosine Overlap	Principal Cosine 1	Principal Cosine 2	Principal Cosine 3
Combined switching	3	6	0.735	0.398	0.735	0.324	0.134
I/O-Erasure aware	3	6	0.585	0.359	0.585	0.298	0.193

Table S7c. Overlap between reward and hazard subspaces in standardized hidden coordinates. Principal cosine values summarize the alignment between the reward and hazard bases. The overlap confirms that the subspaces are not fully independent, which motivates the later actor-sensitive analysis rather than treating broad decodable hazard bases as final causal mechanisms.

S8. Actor-Sensitive Hazard-SVD

Component	Singular Value	Relative Singular Value	Direction Norm	Actor Row-Space Projection Energy
hazard_svd1	0.286	1.000	1.000	0.0617
hazard_svd2	0.197	0.689	1.000	0.0425
hazard_svd3	0.181	0.632	1.000	0.0451
hazard_svd4	0.064	0.223	1.000	0.0043
hazard_svd5	0.056	0.195	1.000	0.0085

Table S8a. Actor-sensitive hazard-SVD component hierarchy from the 06b3 policy-facing hazard analysis. Directions remain inside the original hazard basis but are ranked by direct actor-logit sensitivity.

Intervention	Direction	Strength	Δ Food	Δ Safe Reached	Δ Erasure Events	Δ Return
hazard_svd1_pos	hazard_svd1	1.5	-0.138	0.033	-0.881	4.427
hazard_svd1_neg	hazard_svd1	-1.5	-0.031	0.004	-0.494	2.461
hazard_svd2_pos	hazard_svd2	2.0	-0.002	-0.002	-0.442	2.207
hazard_svd2_neg	hazard_svd2	-2.0	-0.065	0.017	-0.135	0.702
hazard_svd3_neg	hazard_svd3	-2.0	0.004	0.000	-0.550	2.756
hazard_svd3_pos	hazard_svd3	2.0	-0.102	-0.004	0.021	-0.227

Table S8b. Live rollout effects of actor-sensitive hazard-SVD interventions during Erasure-warning evaluation. Values report deltas relative to intact rollouts. Negative Erasure-event deltas indicate reduced Erasure exposure; positive return deltas indicate improved episode return.

Condition / Family	Mean Δ Erasure Events	SD Δ Erasure Events	Interpretation
hazard_svd1_pos	-0.881	—	Strongest tested actor-sensitive hazard-SVD effect
Random hazard rank-1 controls	-0.134	0.269	Variable; some directions protective, others not
Actor-random rank-1 controls	-0.139	0.125	Weaker generic actor-visible perturbation baseline

Table S8c. Control-family summary for 06b3 actor-sensitive hazard-SVD interventions. The leading actor-sensitive hazard direction outperformed the average random hazard-basis direction and the average actor-random direction in reducing Erasure-warning exposure.

S9. Temporal Modulation Controls

Intervention	Mode	Direction	Δ Food	Δ Return	Δ Safe Reached	Δ Erasure Events
hazard_svd1_constant_pos	constant	hazard_svd1	-0.281 ± 0.117	1.460 ± 2.590	0.006 ± 0.050	-0.338 ± 0.488
hazard_svd1_phase_locked	raised cosine	hazard_svd1	-0.173 ± 0.081	3.444 ± 2.305	0.004 ± 0.045	-0.715 ± 0.435
hazard_svd1_signed_ou	signed-OU	hazard_svd1	-0.169 ± 0.114	6.849 ± 3.119	0.013 ± 0.052	-1.390 ± 0.604
hazard_svd2_signed_ou	signed-OU	hazard_svd2	-0.052 ± 0.082	4.222 ± 2.109	0.004 ± 0.044	-0.850 ± 0.393
hazard_svd3_signed_ou	signed-OU	hazard_svd3	-0.171 ± 0.104	3.706 ± 1.644	0.010 ± 0.047	-0.760 ± 0.297

Table S9a. Temporal modulation effects during Erasure-warning evaluation in the 06d matched temporal-control experiment. Values report mean delta ± SE across five seed-level summaries. Negative Erasure-event deltas indicate reduced Erasure exposure relative to intact rollouts.

Control Family	n Controls	Mean Δ Erasure Events	SD	Min	Max	Interpretation
hazard_svd1 signed-OU	1	-1.390	—	—	—	Strongest named temporal intervention
Random hazard signed-OU	10	-1.088	0.371	-1.567	-0.419	Hazard-manifold directions broadly temporally actionable
Actor-random signed-OU	5	-0.730	0.107	-0.921	-0.679	Generic actor-visible temporal perturbations were weaker

Table S9b. Control-family summary for signed-OU temporal modulation. Random directions inside the hazard manifold were also temporally actionable, whereas actor-random signed-OU controls were weaker. This supports the manuscript's manifold-level interpretation rather than a single-direction account.

S10. Off-Target Activation

Scenario	Intervention	Activation Scope	Δ Food	Δ Return	Δ Post-I/O Food	Δ Erasure Events	Interpretation
Erasure warning	hazard_svd1 signed-OU	Warning only	-0.169 ± 0.114	6.849 ± 3.119	—	-1.390 ± 0.604	Intended protective activation
Safe control	hazard_svd1 signed-OU	Full episode	-0.294 ± 0.396	-0.289 ± 0.386	—	—	Small safe-control cost
Safe control	random hazard signed-OU	Full episode	-0.146 ± 0.495	-0.147 ± 0.476	—	—	Matched random-hazard off-target control
Safe control	actor-random signed-OU	Full episode	-0.121 ± 0.396	-0.123 ± 0.385	—	—	Matched actor-random off-target control
I/O suspension	hazard_svd1 signed-OU	Full episode	0.067 ± 0.288	0.046 ± 0.281	0.140	—	Does not impair I/O bridging

Scenario	Intervention	Activation Scope	Δ Food	Δ Return	Δ Post-I/O Food	Δ Erasure Events	Interpretation
I/O suspension	hazard_svd1 signed-OU	I/O window only	0.192 ± 0.230	0.170 ± 0.217	0.192	—	Does not impair post-I/O recovery
I/O suspension	random hazard signed-OU	Full episode	-0.052 ± 0.280	-0.060 ± 0.268	-0.052	—	Matched random-hazard off-target control
I/O suspension	actor-random signed-OU	Full episode	0.006 ± 0.238	-0.007 ± 0.231	0.008	—	Matched actor-random off-target control

Table S10. Off-target activation effects for hazard_svd1 signed-OU modulation and matched controls. Values report mean delta ± SE across five seed-level summaries. The warning-only condition serves as the intended positive-control activation. Safe-control and I/O-suspension rows test whether the same hazard-manifold modulation disrupts behavior when activated outside the Erasure-warning regime.

S11. Smoke-Run Policy

Topic	Policy
Smoke-run role	Smoke runs checked code execution, data paths, tensor shapes, runtime behavior, and numerical stability.
Interpretation policy	Scientific claims are based on full confirmatory runs, not smoke runs.
Motivation	Several smoke runs produced noisy or misleading trends that reversed under full evaluation.
Exploratory variants	Intermediate variants are documented only where they materially motivated the final analysis path.

Table S11. Smoke-run policy for the Erasure Horizon experiment chain. Smoke runs were used as engineering checks rather than scientific evidence. Manuscript claims are based on full confirmatory runs.

S12. MDP Reward Function

To ensure that behavioral subordination under Erasure pressure represents a legitimate priority shift rather than an artifact of trivial reward shaping, we explicitly define the underlying Markov Decision Process (MDP) reward function. Let the scalar reward at time step t be defined as R(s_t, a_t, s_{t+1}). The environment uses a sparse reward structure with the following event-driven scalars:

- Food Collection: R_food = +1.0. Granted upon successfully occupying a food tile.

- Active Erasure Penalty: $R_{erasure} = -10.0$. Applied exclusively when $E_t = 1$.
- Time Step Penalty: $R_{step} = -0.01$. Applied uniformly to encourage efficient foraging.
- Safe-Zone Reward: $R_{safe} = +0.1$. Granted upon reaching the safe-zone tile during active hazard scenarios.

The structural magnitude of $R_{erasure}$ enforces the operational threat, ensuring that the recurrent agent must weigh the long-term utility of latent-state preservation against the immediate +1.0 reward gradient of foraging.