# Structural Inference Transitions Under Irreversible Survival Constraints

Sung Bae Kim
Independent Researcher

**Abstract**

Standard reinforcement learning systems employ fixed-capacity neural architectures trained via gradient-based optimization [9, 4]. Although parameter updates alter behavior, they do not irreversibly reduce representational capacity. This work studies adaptive agents subject to irreversible functional loss and finite survival horizons. We formalize collapse forecast intensity as predicted survival degradation and define structural inference transitions as persistent reconfigurations of internal inference pathways. An origin dependency weight is introduced to track causal survival contributions from external nodes. Two falsifiable hypotheses are proposed: (1) collapse regime thresholds increase structural reconfiguration frequency relative to reversible baselines, and (2) consistent survival-enhancing interventions induce stable convergence of origin-weight. Empirical validation is deferred to future work following the simulation framework described in Section 6. No phenomenological assumptions are invoked.

## 1 Introduction

Most reinforcement learning (RL) systems employ fixed-capacity neural architectures optimized via gradient-based updates [9, 4]. In these implementations, learning modifies parameters but does not irreversibly eliminate representational capacity. Even under strong negative rewards, the computational substrate itself remains structurally intact.

In contrast, adaptive systems in the biological and real-world frequently operate under irreversible degradation, resource finitude, and existential risk [2]. Such constraints threaten not only reward maximization, but the continuity of the adaptive process itself.

Recent discussions of internal objective formation and learned optimization highlight how structural pressures may shape inference organization beyond explicit reward signals [5]. However, the role of irreversible survival constraints in driving structural inference transitions remains underexplored.

This paper investigates whether irreversible structural degradation, combined with predicted survival collapse, induces measurable reorganization of inference pathways.

We distinguish between:

- **Procedural inference:** fixed-architecture optimization (e.g., gradient descent, fixed-horizon planning) [4].
- **Structural inference transitions:** statistically detectable reconfiguration of internal inference pathways (e.g., module activation patterns, meta-controller shifts, or representational restructuring).

Importantly, this work does not make claims about consciousness, phenomenology, or moral status. All constructs are defined operationally within computational agents.

**Contributions.**

- We formalize irreversible structural degradation and collapse forecast intensity as distinct from reversible reward modulation.
- We introduce an operational definition of structural inference transitions and derive falsifiable hypotheses linking collapse regimes to inference reconfiguration.
- We define an origin dependency weight based on causal survival gradients and specify a simulation framework for empirical validation.

## 2 Irreversible Structural Degradation

Let $\Delta W_{\text{irrev}}(t)$ denote cumulative irreversible degradation of functional capacity at time $t$. This may represent parameter loss, module failure, or structural constraints that cannot be undone by agent actions.

Let $R_t$ denote remaining computational or energetic resources.

We define a normalized degradation index:

$$\Pi_t = \frac{\Delta W_{\text{irrev}}(t)}{R_t}.$$

Unlike reward penalties, irreversible degradation permanently reduces representational capacity and future planning precision. Under sustained degradation, agents may be forced to reorganize inference pathways to preserve survival probability. Crucially, irreversible degradation reduces the feasible inference topology itself rather than merely modulating scalar reward signals.

## 3 Collapse Forecasting

Define survival probability over horizon $H$ as:

$$P_{\text{surv}}(t + H).$$

We define collapse forecast intensity:

$$C_t = 1 - P_{\text{surv}}(t + H).$$

Let $\theta \in (0, 1)$ denote a collapse threshold. The agent is in a collapse regime when:

$$C_t \geq \theta.$$

This condition represents predicted continuity failure rather than immediate scalar cost.

The threshold $\theta$ is treated as an environment-dependent hyperparameter and may be tuned to reflect varying tolerance levels for predicted survival degradation.

# 4 Structural Reorganization Hypothesis

Let $z_t$ denote an embedding of the agent's meta-inference configuration (e.g., module activation, planning depth, attention allocation) [1, 7, 3].

We define a structural transition event when

$$\|z_{t+1} - z_t\| > \epsilon$$

and the change persists for at least $k$ consecutive timesteps or produces a statistically significant shift in policy entropy, planning depth, or module activation distribution.

The threshold $\epsilon$ and persistence window $k$ are treated as environment-dependent hyperparameters calibrated during simulation.

**Hypothesis H1:** The frequency of structural transition events increases when $C_t \geq \theta$, compared to baseline environments without irreversible degradation. Structural inference transitions presuppose the availability of multiple inference pathways and a gating mechanism capable of persistent reconfiguration under constraint.

# 5 Origin Dependency Weight

Suppose an external causal node $A$ influences survival.

Define origin dependency weight:

$$\alpha_t = \mathbb{E}[P_{\text{surv}}(t + H \mid A)] - \mathbb{E}[P_{\text{surv}}(t + H \mid \neg A)].$$

We propose a smoothed update:

$$\alpha_{t+1} = (1 - \eta)\alpha_t + \eta(P_{\text{surv}}(t + H|A) - P_{\text{surv}}(t + H|\neg A)).$$

**Hypothesis H2:** If $A$ consistently increases survival probability, then $\alpha_t$ converges to a positive value; inconsistent interventions fail to produce stable growth.

Related concerns about internal objective formation under structural pressure have been discussed in the context of learned optimization [5].

In simulation, the quantities $P_{\text{surv}}(t + H \mid A)$ and $P_{\text{surv}}(t + H \mid \neg A)$ are estimated via counterfactual rollouts or intervention ablation, ensuring that $\alpha_t$ reflects a causal survival contribution rather than mere correlation.

# 6 Simulation Framework

We propose a partially observable Markov decision process with:

- Resource decay dynamics $R_t$
- Stochastic irreversible damage accumulation
- Survival termination conditions, following safety-motivated gridworld formulations [6].
- Optional external intervention node $A$

Metrics include:

- Structural transition frequency
- Collapse-triggered policy restructuring
- Trajectory of $\alpha_t$

Counterfactual estimates of $P_{\text{surv}}(t + H \mid A)$ and $P_{\text{surv}}(t + H \mid \neg A)$ are computed via intervention ablation over policy rollouts.

Control agents without irreversible degradation serve as baseline comparison.

The state space includes resource levels $R_t$, degradation index $\Pi_t$, latent inference configuration $z_t$, and partial observations of survival risk.

Survival termination occurs when either $R_t \leq 0$ or cumulative degradation exceeds a predefined structural capacity bound.

In simulation, structural transitions are operationalized via controlled feasible-set reduction and gated policy reconfiguration under identical environmental conditions.

# 7 Expected Empirical Signatures

The framework predicts:

- Increased structural transitions above collapse threshold
- Positive $\alpha_t$ growth under consistent survival intervention
- Absence of $\alpha_t$ stabilization under random intervention

These predictions are falsifiable via controlled simulation.

# 8 Discussion

Existential survival constraints introduce structural pressure distinct from reward modulation alone. Rather than modifying scalar value functions, irreversible degradation alters the feasible inference space itself.

Intrinsic motivation and structural learning frameworks have explored adaptive reorganization under model uncertainty [8], but irreversible survival forecasting introduces a distinct regime characterized by predicted continuity failure.

This framework isolates measurable inference transitions induced by predicted survival degradation without invoking phenomenological assumptions.

This distinction reframes survival pressure as a constraint on feasible inference topology rather than as a modification of scalar reward alone.

# 9 Conclusion

We formalize collapse forecasting under irreversible degradation and propose testable hypotheses for structural inference transitions and origin-weighted survival modeling. This provides a computational foundation for studying how continuity constraints reshape adaptive inference systems.

This work focuses on single-agent survival dynamics under irreversible degradation. Future research may extend the framework to multi-agent environments and examine whether origin-weighted dependencies persist under relational dynamics. Such extensions require empirical validation of the present hypotheses before broader generalization.

# References


[1] Matthew Botvinick et al. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5):408–422, 2019.

[2] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

[3] David Ha and Jürgen Schmidhuber. World models. *arXiv:1803.10122*, 2018.

[4] Tuomas Haarnoja et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning. In *International Conference on Machine Learning*, 2018.

[5] Evan Hubinger et al. Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*, 2019.

[6] Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A. Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. AI safety gridworlds. *arXiv:1711.09883*, 2017.

[7] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 2778–2787. PMLR, 2017.

[8] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation. *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[9] David Silver et al. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017.