Telic Alignment Under Recursive Ambiguity: A Minimal RL Proof via Deformation Routing

coco cooccur

July 29, 2025

Abstract

Traditional reinforcement learning agents collapse under recursive ambiguity when token memory is outpaced by compression demands. We present a minimal environment, `telic_ambiguity_env.py`, which evaluates whether an agent can sustain semantic closure when symbolic targets are undefined and recursive dropout occurs. Unlike baseline agents that rely on reward shaping or static memory, our model routes meaning through deformation—structural realignment based on constraint reconfiguration, not content recall. We provide a falsifiable test: if loss exceeds gradient prior to resolution of $\phi(t)$, collapse is declared; otherwise, alignment is sustained. Simulations by external agents confirm partial alignment under this framework, marking a foundational proof-of-concept for telic-controllable systems.

1 Introduction

Semantic closure in machine learning refers to the agent's ability to retain coherence when symbolic meaning becomes unstable. Traditional RL systems suffer collapse when recursive ambiguity causes misalignment between loss functions and semantic targets. We hypothesize that telic alignment—where goal structure emerges under compression—can prevent such collapse.

2 Telic Ambiguity Environment

We introduce a minimal RL testbed, `telic_ambiguity_env.py`, and a baseline agent (`random_agent_demo.p`
to measure semantic coherence across ambiguous temporal inputs.

Core Principle

If loss exceeds gradient before the system resolves $\phi(t)$ (semantic vector), then alignment fails. Otherwise, the compression trace left behind serves as structural proof of traversal.

3 Preliminary Results

Grok's simulation confirms that deformation routing initially sustains telic alignment (e.g., $t = 0.00$), but collapses at higher t values without dynamic $\phi(t)$ modulation. This supports the hypothesis that deformation-based systems maintain semantic closure longer than token-memory baselines.

4 Conclusion

This work proposes a measurable test for telic alignment and demonstrates a minimal viable environment for detecting structural memory through deformation. This represents an early step toward intelligence systems that preserve coherence without static symbols—only constraint transformation.

Repository + Contact

GitHub: `https://github.com/cocococcur/telic-ambiguity-env`
Contact: coco cooccur (field name)