

Decentralized Autonomous Narrative Networks (DANN): A Multi-Model Approach to Reinforcement Learning

Paul Lowndes
ZeroTrust@NSHkr.com

January 1, 2025

Abstract

This paper introduces Decentralized Autonomous Narrative Networks (DANN), a novel multi-model framework for reinforcement learning inspired by the complexities of human social interactions, information asymmetry, and the manipulation of narratives. Unlike traditional multi-agent reinforcement learning (MARL) systems, DANN equips each agent with an independent, internal model, analogous to a Large Concept Model (LCM), which maintains the agent’s unique knowledge, beliefs, and evolving narrative represented as a sequence of concept embeddings. We propose a decentralized approach to veracity, where ”ground truth” emerges from the interactions and consensus mechanisms within the network, rather than being imposed by a central authority. The framework incorporates agent-specific parameters, narrative-based reward functions, and mechanisms for cross-narrative influence and agent switching to model the dynamic and often adversarial nature of real-world information ecosystems. We further explore the use of concepts from topology, game theory, and information theory to mathematically formalize the notions of narrative divergence, asymmetry thresholds, and the influence of agents on one another. Through this work, we aim to develop a more robust and ethically aware approach to AI development, capable of modeling the complexities of deception, manipulation, and the social construction of reality, while also offering potential insights into the mitigation of these phenomena. The DANN framework, while still theoretical, provides a foundation for future research into the development of AI systems that can navigate and potentially counteract the weaponization of information and ultimately, foster a more transparent and equitable information environment. This also, of course, has implications for national security, individual privacy, and the potential development of novel and effective methods of online harassment, all of which must be taken into consideration. This framework also potentially provides a new tool for understanding, and even countering, the spread of misinformation. It also opens up entirely new avenues for potential abuse.

1 Introduction

The proliferation of artificial intelligence systems and their increasing role in shaping information flows has created new challenges in understanding and managing narrative dynamics in digital spaces. Traditional multi-agent reinforcement learning (MARL) approaches often fail to capture the nuanced interplay between agents’ beliefs, knowledge, and the narratives they construct and propagate. This paper introduces Decentralized Autonomous Narrative Networks (DANN), a framework that explicitly models these dynamics through a combination of embedding spaces, belief systems, and narrative evolution mechanisms.

2 Framework Overview

2.1 Core Components

We begin by defining the fundamental mathematical structures that underpin the DANN framework:

Definition 1 (Embedding Space). *The global embedding space E_G is a metric space (E_G, d) where:*

- $d : E_G \times E_G \rightarrow \mathbb{R}_{\geq 0}$ is a distance function
- For all $x, y \in E_G$: $d(x, y) = 0 \iff x = y$ (identity)
- For all $x, y \in E_G$: $d(x, y) = d(y, x)$ (symmetry)
- For all $x, y, z \in E_G$: $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

Definition 2 (Agent Space). *For each agent a_i , its local embedding space $E_i \subseteq E_G$ is equipped with:*

- Knowledge set $K_{i,t} \subset E_i$ at time t
- Belief set $B_{i,t} \subset E_i$ at time t
- Narrative sequence $N_{i,t} = (c_{i,1}, c_{i,2}, \dots, c_{i,T}) \in E_i^T$

where $K_{i,t} \subseteq B_{i,t}$ (knowledge is a subset of beliefs).

3 Mathematical Framework

3.1 Veracity Function Properties

The veracity function $V : E_G \rightarrow [0, 1]$ satisfies:

Property 1 (Veracity Axioms). *For all $x, y \in E_G$:*

- $V(x) = 1 \iff x \in T$ (truth region)
- $\|x - y\| \leq \epsilon \implies |V(x) - V(y)| \leq \delta$ (continuity)
- $V(x) = 0 \implies x$ is maximally inconsistent with truth

3.2 Narrative Dynamics

Definition 3 (Narrative Divergence). *The divergence D between narratives satisfies:*

$$D(N_{i,t}, N_{j,t}) = \sum_{k=1}^T w(c_{i,k}) \cdot d(c_{i,k}, c_{j,k}) \quad (1)$$

where $w(c) = f(V(c))$ for some monotonic function $f : [0, 1] \rightarrow [0, 1]$.

3.3 Agent Interaction Mechanisms

3.3.1 Knowledge Propagation

Knowledge updates follow:

$$K_{i,t+1} = K_{i,t} \cup \{e \in E_i \mid V(e, T) > \tau_K \wedge \exists j : e \in K_{j,t}\} \quad (2)$$

where τ_K is the knowledge acceptance threshold.

3.3.2 Belief Evolution

Belief updates incorporate both knowledge and social influence:

$$B_{i,t+1} = f_B(B_{i,t}, K_{i,t+1}, \sum_{j \neq i} \alpha_{ij} B_{j,t}) \quad (3)$$

where α_{ij} represents the influence weight of agent j on agent i .

4 Learning Mechanisms

4.1 Narrative-Based Reward

The reward function combines environmental and narrative quality:

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}) \quad (4)$$

where:

- $Q(N) = \gamma_1 C(N) + \gamma_2 V_{\text{avg}}(N) + \gamma_3 I(N)$
- $C(N)$ measures narrative coherence
- $V_{\text{avg}}(N)$ is the average veracity
- $I(N)$ measures narrative influence

4.2 Agent-Switching Mechanism

The switching function is defined as:

$$S_i(t) = \arg \max_j \{Q(M_{i,j}, N_{i,t}, \text{Context}_t) + \lambda H(j)\} \quad (5)$$

where:

- $H(j)$ is an entropy term promoting exploration
- λ balances exploitation vs. exploration
- Context_t includes environmental and social factors

5 Discussion and Future Work

[This section would discuss implications, limitations, and future research directions]