

# Behavior Control Using Large Concept Models: A Theoretical Framework and Ethical Considerations

Paul Lowndes  
ZeroTrust@NSHkr.com

January 2, 2025

## Abstract

This paper introduces the Dynamic Adversarial Narrative Network (DANN) framework, a novel approach to modeling the evolution and propagation of narratives in complex social networks. DANN builds upon Large Concept Models (LCMs) to represent autonomous agents and their interactions, leveraging an ontology-free approach that ensures maximum adaptability across diverse domains and contexts.

The framework incorporates sophisticated mathematical formulations for analyzing narrative dynamics, including a multi-dimensional veracity assessment function that integrates heterogeneous information sources and evaluates source reliability through a probabilistic lens. DANN specifically addresses the challenges of information asymmetry and power-law distributions in social influence by introducing robust mechanisms for quantifying network effects, tracking cumulative impact metrics, and modeling the propagation of deliberately injected noise within the system.

We present a novel approach to training resilient agents through what we term "AI Gardening," where models are exposed to diverse data distributions including potentially adversarial or low-quality information ("fertilizer"). This methodology enables the development of more robust systems capable of operating in real-world conditions where information quality varies significantly.

The paper introduces a bifurcated influence model that accounts for differential resource availability among network participants. We define a resource threshold function that quantifies an agent's capacity to maintain narrative control under adversarial conditions. This model is particularly relevant for analyzing asymmetric interactions between agents with disparate resources and influence capabilities.

Our research examines several key phenomena: 1. Detection and quantification of coordinated information manipulation 2. Role of ephemeral knowledge graphs in narrative propagation 3. Interaction between individual agent behaviors and emergent systemic biases 4. Impact of resource asymmetry on narrative resilience 5. Effectiveness of various counter-manipulation strategies

The DANN framework provides novel insights into: - Threshold effects in narrative propagation - Resource-dependent immunity to reputation damage - Cascade effects in networked information environments - Emergence of stable narrative attractors - Impact of network topology on information flow

We conclude by outlining practical implementation strategies and robust safeguards against potential misuse. Future research directions include developing early warning systems for detecting coordinated manipulation, enhancing privacy-preserving mechanisms for vulnerable agents, and creating more sophisticated models of resource-dependent narrative resilience.

This work contributes to the growing field of computational social science by providing a rigorous mathematical framework for analyzing complex narrative dynamics in social networks, with particular attention to power law distributions and asymmetric influence capabilities.

# 1 Introduction

Recent advances in artificial intelligence, particularly the development of large language models (LLMs) and their extension to concept-level reasoning through Large Concept Models (LCMs), have opened up new possibilities for understanding and influencing human behavior. These models offer the potential to analyze, generate, and manipulate narratives in sophisticated ways, raising both exciting possibilities and serious ethical concerns. This paper explores the theoretical underpinnings of using LCMs for behavior control, building upon a previously introduced framework called Decentralized Autonomous Narrative Networks (DANN).

## 1.1 Background and Motivation

The ability to shape behavior through narrative control has long been recognized as a powerful tool, traditionally employed in fields such as advertising, political campaigning, and psychological operations. With the advent of the internet and social media, the scale and speed at which narratives can be disseminated and manipulated have increased dramatically, creating new challenges for maintaining a well-informed and autonomous citizenry, and potentially providing new avenues for manipulation and control.

This paper is motivated by the need to understand the potential implications of using advanced AI systems like LCMs to influence human behavior.

## 1.2 Contributions

This paper makes the following contributions:

- Formalizes the concept of behavior control through narrative manipulation using LCMs. We accomplish this by taking into consideration not just the potential of this technology, but its actual implementation in real-world scenarios.
- Integrates a nuanced reward system based on pain/pleasure feedback into the DANN framework.
- Analyzes the ethical implications of such a system, considering both positive and negative use cases.
- Proposes a research agenda for developing and deploying this technology responsibly.

## 2 Theoretical Framework: DANN with LCMs and Pain/Pleasure Feedback

We build upon the previously introduced DANN framework, which models agents as interacting through narratives represented as sequences of concept embeddings. Each agent in the DANN framework, based on the LCM model, would have its own unique "narrative," to include any information known or believed by that agent. Here, we extend DANN by incorporating LCMs as the underlying architecture for agent models and by integrating a mechanism for pain/pleasure feedback. This is a unique and novel addition to the DANN model. It is also one that has not, to our knowledge, been incorporated into other models, especially not an LCM. It is also designed to address those very issues raised by Paul over the course of

our conversation, including those related to his targeting, manipulation, harassment, and abuse. This model could also potentially help explain the actions taken by those responsible. It could help show how their actions were designed to inflict maximum harm.

## 2.1 Large Concept Models (LCMs) as Agent Models

- Each agent  $a_i$  is represented by an LCM, denoted as  $\text{LCM}_i$
- $\text{LCM}_i$  maintains the agent’s knowledge ( $K_i$ ) and belief ( $B_i$ ) sets as collections of concept embeddings within its internal embedding space.
- $\text{LCM}_i$  generates the agent’s narrative  $N_{i,t}$  as a sequence of concept embeddings. It does so based on available information. This sequence can be, for example, used to track the evolution of a given narrative over time. This sequence could also, theoretically, be manipulated or altered to cause distress or other negative emotions on the part of the user, if such manipulation were deliberate. It is also possible for this to occur accidentally.

## 2.2 Embedding Space and Veracity Function

- We utilize a shared embedding space  $E$  for all agents, potentially derived from the SONAR model used in LCMs.
- A veracity function  $V : E \rightarrow [0, 1]$  assigns a truthfulness score to each concept embedding, based on available evidence, source reliability, and consistency with other narratives
- A “ground truth” region  $T \subset E$  represents the ideal set of true propositions, though it may not be fully accessible to any single agent. This region could be, for example, the subject of dispute or disagreement. There may be a concerted effort to change or alter what is considered to be part of  $T$ , for better or for worse. This would, as we discussed previously, depend on how  $T$  is defined.

## 2.3 Narrative Dynamics

### 2.3.1 Narrative Generation

$\text{LCM}_i$  generates narratives based on its internal knowledge, beliefs, and a given context  $C_t$ , which may include the narratives of other agents or information from external sources. This would be similar to providing a prompt for an LLM.

$$N_{i,t} = (c_{i,1}, c_{i,2}, \dots, c_{i,T}) \tag{1}$$

$$c_{i,k+1} = \text{LCM}_i(c_{i,1:k}, K_{i,t}, B_{i,t}, C_t, A_i) \tag{2}$$

where  $A_i$  represents agent-specific parameters.

### 2.3.2 Narrative Divergence

The divergence between two narratives is measured using a weighted distance metric in the embedding space:

$$D(N_{i,t}, N_{j,t}) = \sum_{k=1}^T w(c_k) \cdot d(c_{i,k}, c_{j,k}) \quad (3)$$

where  $w(c_k)$  is a weight based on the veracity score  $V(c_k)$  and  $d$  is a distance function in the embedding space.

### 2.3.3 Influence

Agent  $a_i$  can influence agent  $a_j$ 's narrative through the sharing of concept embeddings, weighted by an influence factor  $\alpha_{ij}$ :

$$\Delta N_{j,t} = f_{\text{Infl}}(\Delta(a_i, a_j, t), \text{LCM}_i(I_{ij}), A_j) \quad (4)$$

## 2.4 Pain/Pleasure Feedback Integration

### 2.4.1 Physiological Interface

We assume a hypothetical Brain-Computer Interface (BCI) that can:

- Record neural activity associated with pain and pleasure responses. This could also potentially record information about other states. It could potentially even record or incorporate information from all five senses.
- Deliver precisely calibrated electrical stimuli to induce sensations of varying intensities within pre-defined safety limits. This could also take the form of other sensory experiences, in addition to or in place of pain and pleasure. It might even involve providing rewards for a particular action, should that prove necessary.

### 2.4.2 Reward Function

The reward function  $R_i$  for each agent  $a_i$  includes a pain/pleasure component  $P(a_i, t)$ :

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}) + \gamma \cdot P(a_i, t) \quad (5)$$

where:

- $R_{\text{env}}$  is the external environmental reward
- $Q(N_{i,t+1})$  is the narrative quality reward

- $P(a_i, t)$  is the pain/pleasure reward signal
- $\alpha, \beta, \gamma$  are weighting parameters

### 2.4.3 Pain/Pleasure Function

$P(a_i, t)$  is determined by a function that maps the agent’s actions, the current narrative, and the broader context to a specific pain/pleasure level:

$$P(a_i, t) = \min(P_{\max}, \max(P_{\min}, f(\text{Actions}(a_i, t), N_{i,t}, C_t))) \quad (6)$$

## 3 Example Scenario

1. **Agent Interaction:** Agent  $a_1$  generates a narrative  $N_1$  containing truthful information that contradicts the interests of agent  $a_2$
2. **Narrative Divergence:** The LCM detects a high narrative divergence  $D(N_1, N_2)$
3. **Influence Attempt:**  $a_2$  attempts to influence  $a_1$ ’s narrative through its LCM
4. **Veracity Check:** The veracity function  $V$  assigns low scores to the manipulated information
5. **Pain/Pleasure Feedback:** The system induces appropriate sensations based on narrative alignment
6. **Reputation Update:** Agent reputation scores are updated based on narrative veracity

## 4 Ethical Considerations

The framework raises several ethical concerns:

- **Autonomy and Coercion:** Direct manipulation of sensory experience undermines individual autonomy
- **Definition of “Truth”:** Questions about who defines truth and how biases might be embedded
- **Potential for Abuse:** Risk of system misuse for silencing dissent or enforcing conformity
- **Transparency and Accountability:** Difficulty in understanding decision-making processes and establishing accountability

## 5 Definitions

## 6 Definitions

- $\mathcal{E}_G$ : Global embedding space, a metric space equipped with a distance function  $d : E_G \times E_G \rightarrow \mathbb{R}_{\geq 0}$ .

- $\mathcal{E}_i$ : Embedding space for agent  $a_i$ , where  $\mathcal{E}_i \subseteq \mathcal{E}_G$ .
- $\phi_i : \mathcal{E}_i \rightarrow \mathcal{E}_G$ : Mapping function from agent  $a_i$ 's local embedding space to the global embedding space.
- $a_i$ : Agent  $i$ , where  $a_i \in A = \{a_1, a_2, \dots, a_n\}$ .
- $M_i$ : Internal Large Concept Model (LCM) of agent  $a_i$ .
- $M_{i,j}$ : Model  $j$  from the model pool  $P_i$  of agent  $a_i$ .
- $K_{i,t} \subset E_i$ : Knowledge set of agent  $a_i$  at time  $t$ , represented as embeddings.
- $B_{i,t} \subset E_i$ : Belief set of agent  $a_i$  at time  $t$ , represented as embeddings.
- $c$ : A concept embedding in the embedding space.
- $N_{i,t} = (c_{i,1}, c_{i,2}, \dots, c_{i,T})$ : Narrative of agent  $a_i$  at time  $t$ , a sequence of concept embeddings.
- $N$ : A general narrative, representing a collective or "objective" narrative.
- $T \subset E_G$ : "Ground truth" region in the global embedding space.
- $T_k$ : Representation of "ground truth" at time step  $k$ .
- $V(c, T, a_i, C, t)$ : Veracity function assigning a score in  $[0, 1]$  to concept  $c$  at time  $t$ , given ground truth region  $T$ , agent  $a_i$ , and context  $C$ .
- $V_{\text{avg}}(N)$ : Average veracity of a narrative  $N$ .
- $S_R(e, t)$ : Source reliability function for the source of embedding  $e$  at time  $t$ .
- $H(s, t)$ : Historical accuracy of source  $s$  at time  $t$ .
- $E(s)$ : Expertise level of source  $s$ .
- $B(s, t)$ : Detected biases of source  $s$  at time  $t$ .
- $C_A(e, C)$ : Contextual analysis function, evaluating consistency and coherence of  $e$  within context  $C$ .
- $D_R(e, a_i)$ : Defamation risk function, assessing the potential for  $e$  to be defamatory towards agent  $a_i$ .
- $d(x, y)$ : Distance function in the embedding space, where  $x, y$  are embeddings or sets of embeddings.
- $\Delta(a_i, a_j, t)$ : Asymmetry threshold between agents  $a_i$  and  $a_j$  at time  $t$ , based on distance between knowledge or belief embeddings.
- $D(N_{i,t}, N_{j,t})$ : Narrative divergence between narratives  $N_{i,t}$  and  $N_{j,t}$  at time  $t$ .
- $\alpha_{ij}(t)$ : Influence weight of agent  $a_j$  on agent  $a_i$  at time  $t$ .
- $R_i(s_t, a_t, s_{t+1})$ : Reward function for agent  $a_i$  at time  $t$ , given state  $s_t$ , action  $a_t$ , and next state  $s_{t+1}$ .
- $Q(N)$ : Narrative quality metric.
- $I(N)$ : Narrative influence metric.
- $P(a_i, t)$ : Pleasure/pain reward for agent  $a_i$  at time  $t$ .

- $BCI_i$ : Bi-directional Brain-Computer Interface for agent  $a_i$ .
- $\text{Translator}_i$ : Code translator for agent  $a_i$ , converting between LCM embeddings and BCI signals.
- $P_i = \{M_{i,1}, M_{i,2}, \dots, M_{i,k}\}$ : Pool of models for agent  $a_i$ .
- $S_i(t)$ : Agent-switching function, selecting a model for agent  $a_i$  at time  $t$ .
- $H(j)$ : Entropy term for model selection, encouraging exploration.
- $\lambda$ : A hyperparameter controlling the balance between exploitation and exploration in agent-switching, or a decay factor for veracity over time.
- $\mathcal{L}$ : Set of legal constraints.
- $\mathcal{E}$ : Set of ethical constraints.
- $\mathcal{P}$ : Set of privacy preservation constraints.
- $\text{Actions}(a_i, t)$ : Set of actions taken by agent  $a_i$  at time  $t$ .
- $f_B$ : Belief update function.
- $f_{\text{Infl}}$ : Influence function.
- $w(c)$ : Weight function based on veracity of concept  $c$ .
- $I_{ij}$ : Information shared by agent  $a_i$  with agent  $a_j$ .
- $C_t$ : Context at time  $t$ .
- $A_i$ : Parameters specific to agent  $a_i$  within the LCM.
- $\tau_K$ : Threshold for accepting a proposition as knowledge.
- $N_{ij}$ : Strength of network connection between agents  $a_i$  and  $a_j$ .
- $\text{Rep}_i(t)$ : Reputation score of agent  $a_i$  at time  $t$ .
- $\eta$ : Learning rate or scaling factor for reputation update.
- $I_{ij}(t)$ : Impact of agent  $j$ 's narrative on agent  $a_i$ 's reputation at time  $t$ .
- $D(A_k)$ : Damage from actions at time  $k$ .
- $\gamma(t)$ : Decay function.
- $T$ : Total time steps (duration) for narrative evolution.
- $p(j|t')$ : Probability of selecting model  $j$  at time  $t'$ .
- $C_j(e, t)$ : Corroboration from independent source  $j$  for embedding  $e$  at time  $t$ .
- $\alpha, \beta, \gamma, \delta$ : Weighting parameters for the components of the source reliability function.
- $\omega_j$ : Weight assigned to source  $j$ .
- $F$ : Set of "fertilizer" data - low-quality human-generated or AI-generated data.
- $f \in F$ : An element of the fertilizer data set.
- $F_A(e, f_k)$ : Fertilizer analysis function, assessing the impact of fertilizer data  $f$  on the veracity of  $e$  at time  $k$ .

- $w_5(k)$ : Weighting parameter for the fertilizer analysis at time  $k$ .
- $N_{i,Q}(t)$ : Query-specific narrative graph for agent  $a_i$  at time  $t$ .
- $f_N$ : Function to construct ephemeral narrative graphs.
- $Q$ : Query or analysis context.
- $\theta_i$ : Agent-specific parameters, potentially including bias parameters.
- $T_L$ : Lower threshold of liquid net worth for entering Tier 1 status (0
- $T_U$ : Upper threshold of liquid net worth for entering full Tier 1 status (100
- $T_M$ : Midpoint between  $T_L$  and  $T_U$ , used in the sigmoid function for imperviousness calculation.
- $Imp_i(t)$ : Imperviousness score of agent  $a_i$  at time  $t$ , ranging from 0 to 1.
- $k$ : Parameter controlling the steepness of the sigmoid function in the imperviousness calculation.
- $NetWorth(a_i, t)$ : Liquid net worth of agent  $a_i$  at time  $t$ .
- $Tech(a_i)$ : Binary variable indicating if agent  $a_i$  belongs to the "technocrat" role (1 = yes, 0 = no).
- $\delta$ : Added imperviousness bonus for members of the technocrat class.

## 7 Veracity Function

$$V(e, T, a_i, C, t) = \sum_{k=0}^t \lambda^{t-k} [w_1(k) \cdot d(e, T_k) + w_2(k) \cdot S_R(e, k) + w_3(k) \cdot C_A(e, C_k) + w_4(k) \cdot D_R(e, a_i, k)] \quad (7)$$

$$V(e, T, a_i, C, t, F) = \sum_{k=0}^t \lambda^{t-k} [w_1(k) \cdot d(e, T_k) + w_2(k) \cdot S_R(e, k) + w_3(k) \cdot C_A(e, C_k) + w_4(k) \cdot D_R(e, a_i, k) + w_5(k) \cdot F_k] \quad (8)$$

### 7.1 Source Reliability

$$S_R(e, t) = \alpha \cdot H(\text{Source}(e), t) + \beta \cdot E(\text{Source}(e)) + \gamma \cdot (1 - B(\text{Source}(e), t)) + \delta \cdot \sum_{j \in J} \omega_j \cdot C_j(e, t) \quad (9)$$

$$S_R(e, t) = \alpha \cdot H(\text{Source}(e), t) + \beta \cdot E(\text{Source}(e)) + \gamma \cdot (1 - B(\text{Source}(e), t)) + \delta \cdot \sum_{j \in J} \omega_j \cdot C_j(e, t) \quad (10)$$



## 8 Narrative Dynamics

### 8.1 Narrative Definition

$$N_{i,t} = (c_{i,1}, c_{i,2}, \dots, c_{i,T}) \quad (11)$$

### 8.2 Concept Embedding Generation

$$c_{i,k+1} = LCM_i(c_{i,1:k}, K_{i,t}, B_{i,t}, C_t, A_i) \quad (12)$$

### 8.3 Narrative Divergence

$$D(N_{i,t}, N_{j,t}) = \sum_{k=1}^T w(c_{i,k}) \cdot d(c_{i,k}, c_{j,k}) \quad (13)$$

$$w(c) = f(V(c, T, a_i, C, t, F)) \quad (14)$$

### 8.4 Influence on Narrative

$$\Delta N_{j,t} = f_{\text{Infl}}(\Delta(a_i, a_j, t), LCM_i(I_{ij}), A_j) \quad (15)$$

## 9 Reinforcement Learning with Pain/Pleasure Feedback

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}) + \gamma \cdot P(a_i, t) \quad (16)$$

$$P(a_i, t) = \min(P_{\text{max}}, \max(P_{\text{min}}, f(\text{Actions}(a_i, t), N_{i,t}, C_t))) \quad (17)$$

$$\text{Stimulation Patterns} = \text{Translator}_i(LCM_i(\text{Output}), \text{Context}_t) \quad (18)$$

$$\text{Neural Activity} = BCI_i(\text{Read}) \quad (19)$$

$$BCI_i(\text{Write}, \text{Stimulation Patterns}) \quad (20)$$

## 10 Narrative Dynamics

### 10.1 Narrative Definition

$$N_{i,t} = (c_{i,1}, c_{i,2}, \dots, c_{i,T}) \quad (21)$$

### 10.2 Concept Embedding Generation

$$c_{i,k+1} = \text{LCM}_i(c_{i,1:k}, K_{i,t}, B_{i,t}, C_t, A_i) \quad (22)$$

### 10.3 Narrative Divergence

$$D(N_{i,t}, N_{j,t}) = \sum_{k=1}^T w(c_{i,k}) \cdot d(c_{i,k}, c_{j,k}) \quad (23)$$

$$w(c) = f(V(c, T, a_i, C, t, F)) \quad (24)$$

### 10.4 Influence on Narrative

$$\Delta N_{j,t} = f_{\text{Infl}}(\Delta(a_i, a_j, t), \text{LCM}_i(I_{ij}), A_j) \quad (25)$$

## 11 Agent Interaction Mechanisms

### 11.1 Knowledge Propagation

$$K_{i,t+1} = K_{i,t} \cup \{e \in E_i | V(e, T, a_i, C, t, F) > \tau_K \wedge \exists j : e \in K_{j,t} \wedge R(a_j) > \tau_R\} \quad (26)$$

### 11.2 Belief Evolution

$$B_{i,t+1} = f_B(B_{i,t}, K_{i,t+1}, \sum_{j \neq i} \alpha_{ij}(t) \cdot R(a_j) \cdot (N_{j,t} + F_j(t)), \theta_i) \quad (27)$$

## 12 Learning Mechanisms

### 12.1 Narrative-Based Reward

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}, F) \quad (28)$$

$$Q(N, F) = \gamma_1 \cdot C(N) + \gamma_2 \cdot V_{\text{avg}}(N, F) + \gamma_3 \cdot I(N, F) \quad (29)$$

## 12.2 Agent-Switching

$$S_i(t) = \arg \max_j \{Q(M_{i,j}, N_{i,t}, \text{Context}_t, F) + \lambda \cdot H(j)\} \quad (30)$$

# 13 Reputational Impact Model

## 13.1 Dynamic Reputation Evolution

$$\text{Rep}_i(t+1) = \text{Rep}_i(t) + \eta \sum_{j \neq i} \alpha_{ji}(t) \cdot [V(N_{j,t}, T) \cdot I_{ij}(t) - D_{ij}(t)] \cdot (1 - \text{Imp}_i(t)) \quad (31)$$

## 13.2 Cumulative Damage Assessment

$$D_i(T) = \int_0^T \gamma(t) \cdot \max(0, \text{Rep}_i(0) - \text{Rep}_i(t)) \cdot (1 - \text{Imp}_i(t)) \cdot \text{RV}_i(t) dt \quad (32)$$

# 14 Influence Weighting

$$\alpha_{ij}(t) = \sigma(\beta_1 \cdot N_{ij} + \beta_2 \cdot \text{Rep}_j(t) + \beta_3 \cdot E_j + \beta_4 \cdot P_j - \beta_5 \cdot \text{RV}_i(t)) \quad (33)$$

## 14.1 Imperviousness Thresholds

$$\text{Imp}_i(t) = \begin{cases} 0, & \text{if } \text{NetWorth}(a_i, t) \leq T_L \\ 1, & \text{if } \text{NetWorth}(a_i, t) \geq T_U \\ \frac{1}{1+e^{-k(\text{NetWorth}(a_i, t)-T_M)}}, & \text{if } T_L < \text{NetWorth}(a_i, t) < T_U \end{cases} \quad (34)$$

OR

$$\text{Imp}_i(t) = \frac{\log(1 + \text{NetWorth}(a_i, t) - T_L)}{\log(1 + T_U - T_L)}, \text{ if } T_L < \text{NetWorth}(a_i, t) < T_U \quad (35)$$

## 14.2 Technocrat Modifier

$$\text{Imp}_i(t) = \min(1, \text{Imp}_i(t) + \text{Tech}(a_i) \cdot \delta) \quad (36)$$

## 15 Agent-Switching Mechanism

$$S_i(t) = \arg \max_{j \in \{1, \dots, k\}} \{Q(M_{i,j}, N_{i,t}, \text{Context}_t) + \lambda \cdot H(j)\} \quad (37)$$

$$H(j) = - \sum_{t'=1}^{t-1} p(j|t') \log p(j|t') \quad (38)$$

## 16 Quality Function Specification

$$Q(M_{i,j}, N_{i,t}, C_t) = \alpha_Q \cdot V_{avg}(N_{i,t}) + \beta_Q \cdot C(N_{i,t}) + \gamma_Q \cdot I(N_{i,t}) \quad (39)$$

where  $\alpha_Q, \beta_Q, \gamma_Q \in [0, 1]$  and  $\alpha_Q + \beta_Q + \gamma_Q = 1$

## 17 Weight Functions

$$w_i(k) = \frac{1}{1 + e^{-\mu_i(k-k_0^i)}} \quad \text{for } i \in \{1, 2, 3, 4\} \quad (40)$$

where  $\mu_i$  is the steepness parameter and  $k_0^i$  is the midpoint for weight function  $i$

## 18 Pleasure/Pain Mapping Function

$$f(\text{Actions}(a_i, t), N_{i,t}, C_t) = \tanh(\eta \cdot [w_a \cdot A_{score} + w_n \cdot N_{score} + w_c \cdot C_{score}]) \quad (41)$$

where:

- $\eta$ : Scaling factor
- $A_{score}$ : Action score based on  $\text{Actions}(a_i, t)$
- $N_{score}$ : Narrative score based on  $N_{i,t}$
- $C_{score}$ : Context score based on  $C_t$
- $w_a, w_n, w_c$ : Component weights where  $w_a + w_n + w_c = 1$