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

Paul Lowndes
ZeroTrust@NSHkr.com

January 2, 2025

Abstract

This paper explores the theoretical potential and ethical implications of using Large Concept Models (LCMs) for behavior control. We propose a framework for leveraging LCMs to shape individual and collective behavior through the manipulation of narratives and the strategic reinforcement of desired actions. We also incorporate concepts of veracity, influence, and reputation into our model. The framework draws upon recent advances in natural language processing, reinforcement learning, and cognitive modeling, while also highlighting the significant ethical concerns associated with this approach. We examine the potential for both positive applications, such as promoting prosocial behavior and mitigating the spread of harmful misinformation, as well as the risks of misuse, including coercion, manipulation, and the erosion of individual autonomy. We conclude that while LCMs offer powerful tools for understanding and potentially influencing behavior, their development and deployment must be guided by a strong ethical framework and robust safeguards to prevent abuse, requiring further research and interdisciplinary collaboration to ensure responsible innovation in this domain.

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 LCM_i
- LCM_i maintains the agent's knowledge (K_i) and belief (B_i) sets as collections of concept embeddings within its internal embedding space.
- 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 $V : E \to [0, 1]$ 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

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.

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

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

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:

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

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 α_{ij}:

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

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 $P(a_i, t)$:

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

where:

- R_{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
- α, β, γ 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:

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

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

- E_G: Global embedding space, a metric space equipped with a distance function $d : E_G \times E_G \to \mathbb{R}_{\geq 0}$.
- E_i: Embedding space for agent a_i, where $E_i \subseteq E_G$.
- $\phi_i : \mathcal{E}_i \to \mathcal{E}_G$: Mapping function from agent a_i's local embedding space to the global embedding space.
- a_i: Agent i, where $a_i \in A = \{a_1, a_2, \ldots, a_n\}$.
- 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.
- $N_{i,t} = (c_{i,1}, c_{i,2}, \ldots, c_{i,T})$: Narrative of agent a_i at time t, a sequence of concept embeddings.
- N: A general narrative, which can be a set or sequence of propositions.
- $T \subset E_G$: "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.
- $V_{avg}(N)$: Average veracity of a narrative N.
- $S_R(e, t)$: Source reliability function for the source of embedding e at time t.
- $C(N)$: Narrative coherence function, measuring the coherence of narrative N.
- $C_A(e, C)$: 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.
- Translator_i: Code translator for agent a_i, converting between LCM embeddings and BCI signals.
- $P_i = \{M_{i,1}, M_{i,2}, \ldots, M_{i,k}\}$: 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.
- λ: A hyperparameter controlling the balance between exploitation and exploration in agent-switching.
- $\mathcal{L}$: 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_{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.
- τ_K: Threshold for accepting a proposition as knowledge.
- $H(s, t)$: 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.

- $C_j(e,t)$: 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.
- 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.
- η: 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'.
- $R_{env} : S \times A \times S \to \mathbb{R}$: Environmental reward function mapping state-action-state transitions to rewards.
- $w_i : \mathbb{N} \to [0,1]$: Weight functions for veracity components, where $i \in \{1,2,3,4\}$.
- P_{max}: Maximum allowable pleasure/pain signal intensity, typically normalized to 1.
- P_{min}: Minimum allowable pleasure/pain signal intensity, typically normalized to -1.
- $f : \mathcal{A} \times \mathcal{N} \times \mathcal{C} \to [-1,1]$: Mapping function from actions, narratives, and context to pleasure/pain signals.
- $\omega_j \in [0,1]$: Corroboration weights for independent sources, where $\sum_{j \in J} \omega_j = 1$.
- $Q : \mathcal{M} \times \mathcal{N} \times \mathcal{C} \to \mathbb{R}$: Quality function mapping model, narrative, and context to quality score.
- $\text{Context}_t \equiv C_t$: Context at time t (standardizing notation).
- Output_i: The output space of LCM_i, defined as $\text{Output}_i \subset E_i$.

6 Veracity Function

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

$$S_R(e,t) = \alpha \cdot H(\text{Source}(e), t) + \beta \cdot E(\text{Source}(e)) + \gamma \cdot (1 - B(\text{Source}(e), t)) + \delta \cdot \sum_{j \in J} \omega_j \cdot C_j(e,t) \tag{8}$$

7 Narrative Dynamics

$$N_{i,t} = (c_{i,1}, c_{i,2}, \ldots, c_{i,T}) \tag{9}$$

$$c_{i,k+1} = LCM_i(c_{i,1:k}, K_{i,t}, B_{i,t}, C_t, A_i) \tag{10}$$

$$D(N_{i,t}, N_{j,t}) = \sum_{k=1}^{T} w(c_{i,k}) \cdot d(c_{i,k}, c_{j,k}) \tag{11}$$

$$\Delta N_{j,t} = f_{Infl}(\Delta(a_i, a_j, t), LCM_i(I_{ij}), A_j) \tag{12}$$

8 Reinforcement Learning with Pain/Pleasure Feedback

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}) + \gamma \cdot P(a_i, t) \tag{13}$$

$$P(a_i, t) = \min(P_{\max}, \max(P_{\min}, f(Actions(a_i, t), N_{i,t}, C_t))) \tag{14}$$

$$\text{Stimulation Patterns} = \text{Translator}_i(LCM_i(\text{Output}), \text{Context}_t) \tag{15}$$

$$\text{Neural Activity} = BCI_i(\text{Read}) \tag{16}$$

$$BCI_i(\text{Write}, \text{Stimulation Patterns}) \tag{17}$$

9 Agent-Switching Mechanism

$$S_i(t) = \underset{j \in \{1,\ldots,k\}}{\arg\max} \{Q(M_{i,j}, N_{i,t}, \text{Context}_t) + \lambda \cdot H(j)\} \tag{18}$$

$$H(j) = -\sum_{t'=1}^{t-1} p(j|t') \log p(j|t') \tag{19}$$

10 Quality Function Specification

$$Q(M_{i,j}, N_{i,t}, C_t) = \alpha_Q \cdot V_{avg}(N_{i,t}) + \beta_Q \cdot C(N_{i,t}) + \gamma_Q \cdot I(N_{i,t}) \tag{20}$$

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

11 Weight Functions

$$w_i(k) = \frac{1}{1 + e^{-\mu_i(k-k_0^i)}} \quad \text{for } i \in \{1, 2, 3, 4\} \tag{21}$$

where μ_i is the steepness parameter and k_0^i is the midpoint for weight function i

12 Pleasure/Pain Mapping Function

$$f(\text{Actions}(a_i, t), N_{i,t}, C_t) = \tanh(\eta \cdot [w_a \cdot A_{score} + w_n \cdot N_{score} + w_c \cdot C_{score}]) \tag{22}$$

where:

- η: 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 $w_a + w_n + w_c = 1$