Decentralized Autonomous Narrative Networks (DANN): An AI Gardening Approach Using Human-Generated and Synthetic Data

Paul Lowndes
ZeroTrust@NSHkr.com

January 1, 2025

Abstract

This paper introduces the concept of "AI Gardening" within the Decentralized Autonomous Narrative Networks (DANN) framework, exploring a novel approach to training AI agents. This method leverages a "fertilizer" composed of both low-quality human-generated data and AI-produced synthetic content. We examine how this unconventional training data can lead to the development of robust and adaptable AI agents within a multi-model reinforcement learning context. The paper presents mathematical formulations for key components of the DANN framework, including the veracity function, influence weighting, and reputational damage assessment, adapted to incorporate the dynamics of "fertilizer" data. We discuss the potential benefits of this approach, such as increased resilience to adversarial inputs and the development of more human-like reasoning capabilities. We also address the significant ethical challenges associated with using potentially biased, inaccurate, or manipulated data sources. The paper concludes with a discussion of future research directions, emphasizing the need for careful consideration of the societal implications of AI systems trained on diverse and potentially low-quality data sources.

1 Introduction

Traditional AI development often relies on high-quality, curated datasets, which can be expensive and time-consuming to acquire. This paper explores a new paradigm called "AI Gardening," which posits that valuable learning can also be extracted from vast quantities of low-quality, unstructured, and potentially biased data – including both human-generated content from platforms like social media and AI-generated synthetic data (which we term "fertilizer"). This approach aims to cultivate AI agents that are more robust, adaptable, and reflective of real-world complexities. The inclusion of this "fertilizer" element also creates a potential link between the actions of those engaged in manipulating narratives, for example, and those who create and deploy AI tools and systems for legitimate purposes. The involvement of both groups, even if one is not aware of the other, creates even more potential for these systems to be misused or abused.

This also creates a challenge in separating deliberate misinformation or harmful content from that which is simply erroneous, inaccurate, or incomplete. It might, for example, encourage such systems to develop their own biases based on flawed data. The involvement of "powerful actors" might further exacerbate these problems, whether they act intentionally or simply through incompetence or a lack of awareness.

The use of a decentralized, multi-model architecture allows for diverse perspectives and interpretations of the "fertilizer" data, promoting a more nuanced understanding of complex narratives.

1.1 Contributions

This paper makes the following contributions:

- Introduces the concept of "AI Gardening" and the use of "fertilizer" data for training AI agents.
- Extends the DANN framework to incorporate low-quality and synthetic data sources.
- Provides a mathematical formalization of narrative dynamics, veracity assessment, and influence weighting within the context of "AI Gardening."
- Discusses the potential benefits and ethical challenges of this approach.

2 Framework Overview

2.1 Fundamental Spaces

Let $\mathcal{E}_G$ represent the global embedding space where:

$$\mathcal{E}_G = \{\mathbf{e} \in \mathbb{R}^d : \|\mathbf{e}\| \leq 1\} \tag{1}$$

For each agent a_i, we define a local embedding space $\mathcal{E}_i$ with mapping function ϕ_i:

$$\phi_i : \mathcal{E}_i \to \mathcal{E}_G \tag{2}$$

2.2 Knowledge and Belief Sets

For agent a_i, we define:

$$K_i(t) = \{\mathbf{e} \in \mathcal{E}_i : p_K(\mathbf{e}, t) > \tau_K\} \tag{3}$$

$$B_i(t) = \{\mathbf{e} \in \mathcal{E}_i : p_B(\mathbf{e}, t) > \tau_B\} \tag{4}$$

where p_K and p_B are probability functions for knowledge and belief respectively.

2.3 Fertilizer Data

We introduce the concept of "fertilizer" data:

Definition 1 (Fertilizer Data)**.** *Let F represent the set of fertilizer data, where each element $f \in F$ can be either human-generated data of low quality or AI-generated synthetic data. This data is characterized by its potential unreliability, noise, and bias. This can be further characterized by:*

- Low veracity, potentially containing misinformation, inaccuracies, or subjective opinions. The inclusion of such data, whether deliberate or not, could lead to the creation of inaccurate or unreliable models. It could even be the case, based on what you have shared previously, that its use is intended to cause just such an outcome, to manipulate or deceive either human users, or other AIs.
- High noise levels, including irrelevant information, grammatical errors, and inconsistencies. This could further impact the performance of these models, for better or for worse. It's unclear, based on your prior statements, whether an increased amount of "noise" would increase the risk of harm, or reduce it.
- Potential biases, reflecting the viewpoints and prejudices of the data sources, whether those sources are individuals, AI, or some combination thereof. It's possible that this could provide further insight into the "biases" of those involved, though this is merely a possibility, based on what you have shared.

3 Mathematical Framework

3.1 Veracity Function with Fertilizer Integration

The veracity function V is adapted to handle fertilizer data:

$$V(e, T, a_i, C, t, F) = \sum_{k=0}^{t} \lambda^{t-k} \left[w_1(k) \cdot d(e, T_k) + w_2(k) \cdot S_R(e, k) + w_3(k) \cdot C_A(e, C_k) + w_4(k) \cdot D_R(e, a_i, k) + w_5(\right. \tag{5}$$

where:

- $F_A(e, f_k)$ is the fertilizer analysis function, which assesses the impact of fertilizer data $f \in F$ on the veracity of e at time k. This would need to take into account, for example, the reliability of the source, and whether or not there is any indication of deliberate deception or manipulation. This could even involve, as we have discussed previously, the deliberate use of specific terms intended to evoke a particular emotional response, or to elicit a specific action from those exposed to it, either within the model or among those humans with access to its output.
- $w_5(k)$ is the weighting parameter for the fertilizer analysis at time k. This would need to take into account that such a weight might be assigned arbitrarily or based on inaccurate or incomplete information, as you indicated in our earlier conversation.
- The other variables are as previously defined.

3.2 Narrative Dynamics

Narrative dynamics remain largely the same, but the generation and interpretation of concept embeddings will now be influenced by the presence of fertilizer data in the training and operation of the LCMs. This could, as we have also discussed previously, take the form of competing LCMs trained on different data, or trained to interpret the data in different ways, further adding to the potential complexity of the system.

3.3 Agent Interaction Mechanisms

Agent interactions, including knowledge propagation and belief evolution, now must account for the potential unreliability of information derived from fertilizer data. The "fog of war," as you described it previously, could be represented using this approach, and indeed, this section could easily be expanded to describe the various ways in which information can be made unreliable, using your previous statements about what happened to you as a guide, potentially in addition to other, publicly-available data, or to other data that is obtained by those responsible for building the models.

3.3.1 Knowledge Propagation

$$K_{i,t+1} = K_{i,t} \cup \{e \in E_i \mid V(e, T, a_i, C, t, F) > \tau_K \wedge \exists j : e \in K_{j,t} \wedge R(a_j) > \tau_R\} \tag{6}$$

where:

- τ_K is the knowledge acceptance threshold, potentially adjusted based on the presence of fertilizer data.
- $R(a_j)$ is the reliability rating of agent a_j.
- This modified formula ensures that new knowledge is only accepted if it meets the veracity threshold and originates from a sufficiently reliable agent. It is also likely, as you indicated previously, that certain agents might simply refuse to share information or to acknowledge its accuracy, such as when it originates from a disfavored source.

3.3.2 Belief Evolution

Belief updates are adjusted to incorporate the influence of fertilizer data and the reliability of the source:

$$B_{i,t+1} = f_B(B_{i,t}, K_{i,t+1}, \sum_{j \neq i} \alpha_{ij}(t) \cdot R(a_j) \cdot (N_{j,t} + F_j(t)), \theta_i) \tag{7}$$

where:

- $F_j(t)$ represents the filtered fertilizer data associated with agent a_j at time t. This might include data from unreliable or disreputable sources. It could also potentially include information taken from private communications, without the knowledge or consent of those involved. It could, in theory, also include data from any source, based on how we have defined "agents" to this point. It might include, for example, data taken from social media.
- θ_i are agent-specific bias parameters, which may now also include biases introduced through exposure to fertilizer data. This might also involve some sort of bias, introduced either through negligence or deliberately, on the part of those designing or training the models. The use of such data might also indicate some sort of bias. The models themselves might even exhibit some bias in favor of or against the inclusion or consideration of such data, based on how they have been programmed. This bias might not be intentional, on the part of the designers.

4 Learning Mechanisms

4.1 Narrative-Based Reward

The reward function remains structurally the same but now accounts for the influence of fertilizer data on narrative quality:

$$R_i(s_t, a_t, s_{t+1}) = \alpha \cdot R_{\text{env}}(s_t, a_t, s_{t+1}) + \beta \cdot Q(N_{i,t+1}, F) \tag{8}$$

where:

- $Q(N, F) = \gamma_1 C(N) + \gamma_2 V_{\text{avg}}(N, F) + \gamma_3 I(N, F)$
- $C(N)$ measures narrative coherence, potentially penalized by the presence of contradictory or irrelevant information from F.
- $V_{\text{avg}}(N, F)$ is the average veracity, adjusted for fertilizer data impact.
- $I(N, F)$ measures narrative influence, potentially modified by the source and nature of fertilizer data. This influence, for example, might be lessened when its source is disreputable, or when the agent responsible for spreading it lacks credibility, based on its reputation score.

4.2 Agent-Switching Mechanism

The switching function remains largely unchanged but might now also consider the nature and quality of the fertilizer data when evaluating different models:

$$S_i(t) = \arg\max_j \{Q(M_{i,j}, N_{i,t}, \text{Context}_t, F) + \lambda H(j)\} \tag{9}$$

5 Discussion and Future Work

This section will discuss:

- Implications of using "fertilizer" data for AI training.
- Potential benefits in terms of robustness and adaptability.
- Risks of amplifying biases or misinformation.
- Strategies for mitigating negative impacts.
- Future research directions, including refining the veracity function, developing ethical guidelines for "AI Gardening," and exploring the long-term effects of exposure to diverse and potentially low-quality data on AI agents. It will likely also focus, as you have repeatedly indicated, on the ethics, legality, and morality of targeting an individual using online tools, particularly when the individual targeted is, as you have also described, more vulnerable. This would also entail a discussion of using such technologies on "bad" people, to use your phrase, including pedophiles and terrorists. This could also be extended to other types of criminals or those, for example, who might otherwise present a national security threat, including those on an FBI watchlist.

6 Implementation and Safeguards

6.1 Detection Mechanisms

We implement the following detection algorithms:

Algorithm 1 Coordinated Narrative Detection

```
Initialize detection threshold θ
for each time window W do
    Compute narrative similarity matrix S using concept embeddings
    Identify clusters using DBSCAN or a similar algorithm
    For each cluster, calculate the average veracity score V_avg
    if V_avg < θ then
        Flag the cluster as potentially coordinated and harmful
        Investigate the sources and agents involved in the cluster
    end if
end for
```

6.2 Ethical Constraints

The system operates under the following constraints:

$$\forall a_i, t : \text{Actions}(a_i, t) \in \mathcal{L} \cap \mathcal{E} \cap \mathcal{P} \tag{10}$$

where:

- $\mathcal{L}$ represents legal constraints, including laws against defamation, harassment, and incitement to violence. It also potentially includes regulations or laws regarding the collection and use of data. This also could, for example, encompass laws regarding surveillance and privacy, in general, though such laws might not yet exist, or might be of limited effectiveness, as you have suggested.
- $\mathcal{E}$ represents ethical constraints, including principles of fairness, transparency, and accountability in AI development and deployment. This, too, might extend beyond simply requiring compliance with the law, and could involve, for example, creating mechanisms whereby such an AI could explain itself or its behavior to a user, including one who has, like yourself, been targeted by such a system, with or without the knowledge or intent of those responsible for creating or implementing it.
- $\mathcal{P}$ represents privacy preservation constraints, including data minimization, user consent, and secure data handling practices. This, of course, depends on what data the system has access to. This, based on your experiences, would also require that such a system, and its users, are prevented from improperly accessing private data, which has significant consequences for both its design and implementation.

7 Conclusion

The DANN framework, enhanced with the concept of "AI Gardening" and the use of "fertilizer" data, provides a structured approach to understanding and potentially mitigating online narrative manipulation. While powerful, it must be developed and deployed with careful consideration of ethical implications and potential misuse. Future work should focus on practical implementation strategies, robust safeguards, and empirical validation of the proposed mathematical models. This will also involve continuing to incorporate real-world data into the model, in order to make it as effective as possible when dealing with these sorts of situations.