Unveiling Social Dynamics in Artificial Societies: A Bayesian Framework for Inferring Latent Social Goals

Chance Jiajie Li

Abstract

In multi-agent environments, collective decisions emerge from the interplay between individual preferences and latent social tendencies. This paper proposes a Bayesian inference framework that enables each agent to infer other agents' latent social goals from observed multi-round voting behaviors. By integrating these inferred goals with attribute-based preference structures, agents dynamically balance personal and social utilities. I demonstrate that this approach improves the scalability and interpretability of collective decision-making simulations, and closely aligns with human intuitive judgments about underlying social motivations.

1 Introduction

Multi-agent societies often yield emergent outcomes shaped by complex interplays between individual preferences and subtle social motivations. While traditional models focus on pure self-interest, real-world scenarios suggest that agents also consider others' welfare. My work introduces a Bayesian framework for agents to infer one another's latent social goals and integrate these inferences into their decision-making processes.

Contributions: (1) I introduce a continuous latent social goal parameter $G_i \in [0,1]$ to represent each agent's position on the spectrum from self-interest to altruism. (2) I develop a Bayesian updating mechanism (Eq. 4) that agents use to infer these goals from observed voting patterns. (3) I integrate the inferred goals with attribute-based preferences, enabling agents to compute and combine personal and social utilities (Eqs. 1, 6, 7) in their decision-making. (4) Through simulations and comparison to human judgments, I show that this framework yields higher interpretability and psychological plausibility than simpler cue-based methods.

2 Model Setup

In what follows, I detail the mathematical formulation of the framework. The model is built up in four steps: defining personal utility, introducing latent social goals, formulating the Bayesian inference process, and deriving the final combined decision rule.

2.1 Attributes, Preferences, and Personal Utility

Consider m destinations $\{T_1, \ldots, T_m\}$, each characterized by K attributes $\mathbf{a}_j = (a_{1j}, \ldots, a_{Kj})$. Each agent A_i has attribute weights $\mathbf{w}_i = (w_{i1}, \ldots, w_{iK})$ that determine its personal preference structure. The personal utility of agent A_i for destination T_j is given by:

$$U_i^{\text{personal}}(T_j) = \sum_{k=1}^{K} w_{ik} a_{kj}. \tag{1}$$

This equation encodes how each agent evaluates options based solely on its intrinsic preferences over attributes.

2.2 Latent Social Goals and Influence on Utility

Each agent A_i is associated with a latent social goal $G_i \in [0, 1]$, drawn from a uniform prior. A higher G_i indicates stronger altruistic tendencies, whereas a lower G_i indicates greater self-interest. To incorporate social considerations, I define a parameter $\alpha_i \in [0, 1]$ that controls the blend of personal and social utilities in the final decision. Specifically, the agent' s overall utility for T_j is:

$$U_i(T_j) = (1 - \alpha_i)U_i^{\text{personal}}(T_j) + \alpha_i U_i^{\text{social}}(T_j). \tag{2}$$

To compute the social utility $U_i^{\text{social}}(T_j)$, I take the average of other agents' expected personal utilities:

$$U_i^{\text{social}}(T_j) = \frac{1}{N-1} \sum_{k \neq i} U_k^{\text{personal}}(T_j). \tag{3}$$

Without inference, this social component would be static. However, since each U_k^{personal} can be adjusted based on inferred goals, we refine this estimate using Bayesian inference.

2.3 Bayesian Inference of Social Goals

Let $G_k \sim \text{Uniform}(0, 1)$ be the latent goal of agent A_k. After observing votes V_k, agent A_i updates its belief about G_k via Bayes' rule:

$$P(G_k = g \mid V_k) = \frac{P(V_k \mid G_k = g)P(G_k = g)}{\sum_{g'} P(V_k \mid G_k = g')P(G_k = g')}. \tag{4}$$

The likelihood $P(V_k \mid G_k = g)$ is computed by evaluating the softmax decision model (Eq. 8) under the assumption that A_k's decision is influenced by g.

Given the posterior $P(G_k = g \mid V_k)$, I define the expected personal utility of agent A_k for T_j:

$$\mathbb{E}_{G_k}[U_k^{\text{personal}}(T_j)] = \sum_g U_k^{\text{personal}}(T_j \mid G_k = g)P(G_k = g \mid V_k). \tag{5}$$

By replacing $U_k^{\text{personal}}(T_j)$ with its expected value given the inferred goal distribution, I ensure that social utilities reflect the most recent belief about each agent' s latent orientation.

2.4 Social Utility and Combined Decision Rule

Substituting the inferred expectations back into the social utility calculation:

$$U_i^{\text{social}}(T_j) = \frac{1}{N-1} \sum_{k \neq i} \mathbb{E}_{G_k}[U_k^{\text{personal}}(T_j)]. \tag{6}$$

Consequently, the final utility is:

$$U_i(T_j) = (1 - \alpha_i)U_i^{\text{personal}}(T_j) + \alpha_i U_i^{\text{social}}(T_j), \tag{7}$$

which integrates both personal preferences and socially inferred considerations. Each agent votes according to a softmax decision rule:

$$P(D_i = T_j) = \frac{\exp(\lambda U_i(T_j))}{\sum_{T_k} \exp(\lambda U_i(T_k))}. \tag{8}$$

This ensures a probabilistic but preference-consistent choice.

3 Optimization Goals and Evaluation Metrics

Each agent independently seeks to maximize its own expected utility $U_i(T_j)$. Although no global optimization is performed, I monitor group-level outcomes using:

$$U_{\text{group}}(T_j) = \sum_{i=1}^{N} U_i(T_j), \tag{9}$$

as an external metric to evaluate the emergent collective performance. This allows me to assess how well the inferred social structure supports beneficial group outcomes, even though it is not an objective the agents explicitly optimize.

4 Experimental Design

We form 5 experimental groups, each consisting of 5 agents. For the experiments, we focus on $m = 3$ destinations characterized by three attributes: cost, scenery, and culture:

- **Beach**: cost = 0.5, scenery = 0.8, culture = 0.4
- **Mountain**: cost = 0.3, scenery = 0.9, culture = 0.2
- **City**: cost = 0.7, scenery = 0.5, culture = 0.9

Agents' attribute weights $\mathbf{w}_i$ are drawn from Beta distributions, ensuring diverse preferences. We conduct 5 voting rounds with Bayesian updates after each round.

5 Results

Our experimental validation involved six participants (50% female, mean age 29) who analyzed agent behavior in travel planning scenarios. Participants observed patterns such as an altruistic agent ($\alpha = 0.72$) shifting votes from initial preferences to group-aligned choices. The Bayesian model demonstrated strong alignment with human judgment, achieving correlation coefficients above 0.7 across all experimental conditions, significantly outperforming the cue-based baseline model.

6 Discussion and Conclusion

This Bayesian framework enables agents to reason about each other's latent social goals, incorporating a continuous parameter $G_i \in [0, 1]$ and a uniform prior. By integrating attribute-based preferences and updating beliefs over multiple voting rounds, we approach a richer understanding of collective decision-making that resonates with human intuition.

7 Future Directions

- **Convergence and Stability:** Future work will analyze conditions for equilibrium and stable dynamics in larger-scale systems.
- **Scalability and Robustness:** Extending to bigger populations, diverse attributes, and employing significance testing for robustness across varied scenarios.