Identifying Latent Intentions via Inverse Reinforcement Learning in Repeated Linear Public Good Games

CARINA I. HAUSLADEN, MARCEL SCHUBERT, and CHRISTOPH ENGEL

Robust results from public good games continue to defy theory. Uncertainty about the distribution of social preferences can explain first round contributions, but not the variance of contribution patterns in repeated play. Using a large-scale dataset (50,390 observations from 2,938 participants) we address this gap with two methodological contributions. First we propose a clustering approach (dynamic time warping) that reflects the nature of the data: it is two-dimensional, relating choices to experiences; and it allows changes to occur at idiosyncratic points of time. Second we refrain from constraining dynamic reactions to predictions derived from static social preferences. Instead, we treat reward function design as an estimation problem, using Inverse Reinforcement Learning to model behavioral patterns as discrete switches between latent intentions. This approach reveals that apparently noisy behavior in social dilemmas can be systematically explained as fluctuations between distinct latent intentions. Our framework successfully accounts for behavioral patterns previously categorized as noise, providing a new paradigm for understanding dynamic decision-making in social dilemma games.

1 Introduction

Standard theory predicts the tragedy of the commons. Everybody maximizes individual profit and exploits socially-minded choices of others. If community members interact repeatedly, but it is known when interaction will stop, the gloomy prediction still holds [75]. A robust experimental literature shows that, in the aggregate, results look different. In a standard symmetric linear public good game, average contributions typically start considerably above zero, but tend to decline over time [18, 27, 28, 53, 55, 80, 110]. A substantial theoretical literature rationalizes this result by introducing some form of social preferences into the utility function [27, 53, 55, 89], whereas other scholars attribute cooperation to confusion, arguing that participants misinterpret the game [5, 21, 64]. Efforts to disentangle cooperation driven by social preferences from that motivated by confusion have not reached a consensus. A widely accepted view posits that observed contributions stem from a combination of confusion and social preferences [12, 23, 54]. Some work argues that social preferences do not account for human cooperation at all, but instead confusion does [21]. Very recently, this conclusion was challenged based on structural shortcomings in the experimental design that led to this conclusion [105].

Thus, the question of whether and how social preferences explain cooperative behavior in social dilemma games remains unresolved. This in particular holds for the development of contributions over time. Previous research has attempted to analyze experimentally generated gameplay data by partitioning it and interpreting the resulting clusters through the lens of established behavioral types. These approaches typically rely on semi-supervised partitioning that assumes strong priors about the expected cluster structure. Yet there are always behavioral trajectories that do not fit any of these types and thus are relegated to substantial "other" categories or noise terms. We argue that this approach suffers from two limitations:

First, partitioning typically relies on *pointwise alignment*. While this approach may be suitable for macroeconomic settings where policy changes affect all firms simultaneously, gameplay data tends to exhibit behavioral reactions that are affected by the idiosyncratic experiences in the randomly composed group. We propose using dynamic time warping, a method specifically designed to handle *shifted time series*, for partitioning gameplay data.

Second, the interpretation of partitions has been constrained by strong theoretical priors taken from preference types identified in static settings (e.g., modeling conditional cooperators as participants who match others' cooperation levels). Essentially, this literature *presupposes knowledge of players' reward functions*. Given the persistent presence of unexplainable noise or "other" clusters, we instead argue for explicitly modeling *reward functions as an estimation problem*. We propose employing Inverse Reinforcement Learning (IRL), where reward functions are not predefined but treated as part of the fitting objective.

2 Literature

Social-preference-types. Categorizing human behavior in social dilemma games has long been of interest to social sciences. Significant parts of the economic literature has focused on identifying types and assigning these static concepts to individuals in these games. For example, Fischbacher et al. [56] introduced the strategy method as a tool to classify individuals' predispositions toward cooperation in linear public good games.[1] This method isolates preferences from the potential influence of learning dynamics or other mechanisms that may emerge during gameplay, focusing on inherent behavioral tendencies rather than observed actions influenced by the game environment.

Building on this foundational work, several subsequent studies [1, 2, 34, 35, 50, 55–59, 61, 62, 70, 73, 84, 100, 102, 104, 107] have used the strategy method to identify distinct behavioral types in the

[1]The strategy method has also been criticized for risking to confuse social preferences with conditional cooperation [20].

very same game. A meta-analysis of these studies [101] found that the latter consistently report the existence of three types: free riders (who contribute nothing, 19.2%), conditional cooperators (whose contributions depend on others' behavior, 61.3 %), and triangle cooperators (10.4%).

Other research has also explored the classification of behavior in public good games but with a focus on *gameplay* data rather than data obtained through Fischbacher et al. [56]'s strategy method [11, 18, 55, 77, 78]. Muller et al. [85] compare both approaches and find that gameplay data often aligns with the classifications derived from the strategy method.

Analyzing gameplay data presents unique challenges compared to data from the strategy method, necessitating the use of more advanced analytical approaches. Proposed methods include Bayesian models [63], finite mixture models [11], clustering techniques [15, 50], Classifier-Lasso (C-Lasso) [51, 98], and Classification and Regression Trees (CART) [40]. However, no systematic theoretical comparison has been made between these approaches or the partitions they generate.

A critical point of comparison is how these models handle behaviors that defy standard explanations: Houser et al. [63] introduce a "Confused" type to capture stochastic or seemingly irrational behaviors that deviate from rational or near-rational strategies. Bardsley and Moffatt [11] employ tremble terms—random, type-invariant deviations from expected contributions. Importantly, these trembles diminish with experience, suggesting they represent learning-related noise rather than persistent irrationality. Fallucchi et al. [51] explicitly model an "others" type to capture behaviors that do not align with reciprocation or strategic adaptation. Fallucchi et al. [50] identify a "Various" group, comprising participants whose contribution patterns do not fit within their four primary categories. A compelling question emerges: What if these "other" clusters stem from limitations in our methodological approaches rather than just erratic or irrational behavior?

This consideration emerges naturally from a key observation shared across all studies: the identification of a "downward trend" in public good contributions. This terminology inherently frames the data as time-series and, importantly, emphasizes trends over specific temporal locations of features such as peaks, which may vary due to idiosyncratic group dynamics. For instance, when contribution peaks occur at different times for different individuals (e.g., round 3 versus round 4), the primary interest lies not in these specific timings but rather in identifying the underlying pattern - such as a mid-game peak - despite slight temporal misalignment. However, none of the previously discussed methods explicitly addresses temporal misalignment in the data. Signal processing provides an elegant solution to this challenge: dynamic time warping combined with barycenter averaging. In this paper we argue that this approach is particularly well-suited for identifying trends in multivariate time-series data such as PGGs.

The third criterion for comparing partitioning methods is the number of theoretical assumptions required for estimation. The Bayesian approach [11] and finite mixture models [11] require substantial theoretical foundations, incorporating numerous assumptions about underlying distributions and processes. C-Lasso demands fewer assumptions while still maintaining some theoretical constraints, whereas hierarchical clustering stands out for its minimal reliance on theoretical assumptions.

Returning to the earlier question: Could the combination of unaddressed temporal misalignment and the reliance on numerous assumptions contribute to the emergence of an "other cluster"—a heterogeneous group that defies straightforward explanation? How might we attempt to better understand such "atypical" clusters? One promising direction lies in examining methods from the learning literature. Indeed, assumption-heavy partitioning methods serve as a natural bridge to learning models because similarly to learning models those methods predefine a certain functional form of agent's behavior, similarly to what learning models do explicitely.

Reinforcement Learning. A multitude of models for thinking and learning in social dilemma games has been proposed. In particular, three types of learning approaches have been studied: evolutionary game-theoretic learning, reinforcement learning, and best-response learning (for a recent review of the literature see [52]).

Specifically, *learning in repeated public goods games* has commonly been modeled as individual-level directional learning influenced by social preferences [4, 7, 31, 68, 108]. Recent experimental findings suggest that social preferences may, however, be unnecessary to explain contributions in repeated games; instead, individuals appear to learn primarily based on payoffs in the game [19, 22, 24, 25]. Related work posits that social preferences primarily determine first-round contributions, with subsequent behavior driven purely by payoff-based learning [32].

The way how the human learns from payoff is frequently modeled as Reinforcement Learning with Loss Aversion [32, 48, 49]. A notable advancement in reinforcement learning is Q-learning [106], which has found applications in economics and strategic interactions. So far, Q-learning has been extensively studied in the context of market competition [26, 71] but it has also been applied to other strategic settings such as the prisoner's dilemma [38] and PGGs [111]. The latter two papers rely on agent based models (ABMs), and no attempt so far was made to explain human behavior in these games with the algorithm.

Modeling a Q-learner requires the definition of a reward function such that Q-values, updated iteratively via the Bellman equation, can guide actions based on the current state. However, defining a comprehensive and suitable reward function poses challenges in complex behavioral tasks [3, 92]. Inverse Reinforcement Learning (IRL) [8, 87] is a popular approach to recover the reward function inductively from the data. IRL has achieved breakthrough successes in robotics [29, 76] and autonomous driving [69, 86] and has recently also become a valuable tool for constructing mathematical models of animal behavior [3, 79, 109]. Traditionally, the animal's reward function has been modeled as a *smoothly* time-varying linear combination [9]. More recently, it has been suggested that behavior can be represented as a Markov chain characterized by alternating between *discrete* intentions [10]. In that spirit, Zhu et al. [112] propose a novel class of Hierarchical Inverse Q-Learning (HIQL) algorithms, which extend the fixed-reward inverse Q-learning (IQL) framework [69] to solve multi-intention IRL problems.

Interpreting the "Other" cluster with the help of discrete latent intentions. Capitalizing on both strands of literature, this paper presents a method for the clean recovery of behavioral patterns and their matching to underlying latent intentions. First, it explicitly characterizes behavior as switching between *discrete* intentions rather than evolving in a *continuous* manner. This modeling approach is able to make sense of the seemingly erratic behavior observed in the so-called "other cluster," rather than discarding switching behavior as mere noise or irrationality. Hence with the help of inverse reinforcement learning, there is hope find an explanation for this seemingly inexplicable behavior. Second, this approach emphasizes the modeling of *latent intentions*—suggesting that there is an unobservable construct driving the observed data. This two-level framework offers a fresh perspective on the field's efforts to identify behavioral types through pattern recognition. Traditionally, the focus has been on reconstructing response functions from observed behavior. Our approach assumes no direct mapping between a single choice function and observed behavior, and instead aims at recovering an underlying state or thought-process. Recent literature refers to this as the "inversion problem" [72], where the challenge lies in inferring mental states that are not directly measurable in behavioral data. The dynamic latent rewards targeted by HIQL can be understood in this vein and thus correspond to the intentions that Kleinberg et al. [72] identify as underlying drivers of human behavior.

This paper. This paper makes several significant contributions to the study of cooperation and learning in experimental settings. One of our innovations lies in the development of a novel approach for identifying patterns in gameplay data. Unlike previous methodologies, our partitioning approach focuses on matching behavioral patterns without requiring exact temporal alignment. This is accomplished through the application of dynamic time warping (DTW) combined with barycenter averaging. We benchmark our classification against four established approaches from the literature: Bayesian modeling, finite mixture models, C-Lasso, and hierarchical clustering. Our comparative analysis demonstrates that the DTW-based partitioning achieves superior discrimination between similar contribution trends, resulting in cleaner and more precise pattern identification.

The second contribution of this work is the introduction of a new perspective on the relationship between behavior and intentions. Drawing from computational neuroscience, we apply hierarchical inverse Q-learning to estimate latent dimensions of cooperation. This method uncovers significant variations in latent intentions over time, revealing that these variations differ across clusters of participants. Notably, clusters previously considered as noise (the "Other" cluster) can now be reinterpreted in terms of latent intentions.

3 Methods

The code used for this study is available in a public GitHub repository. To maintain anonymization, we have provided a link to the anonymized version: https://anonymous.4open.science/r/IRL-public-good-games-1860/README.md.

3.1 The Linear Public Good Game

Our method is designed to uncover patterned heterogeneity and its drivers in repeated, interactive experiments, making it broadly applicable across various settings. Here, we focus on the specific case of a linear public good game, which is a widely studied framework in experimental economics. The game is governed by the following profit function:

$$\pi_{it} = e - c_{it} + \mu \sum_{j=1}^{I} c_{jt}, \tag{1}$$

where π_{it} represents the profit of individual i in period t, e is the fixed endowment allocated to each individual at the start of each period, and c_{it} is the contribution made by individual i to the public project. The term $\mu \sum_{j=1}^{I} c_{jt}$ captures the total returns from the public project, where μ is the marginal per capita return (MPCR), satisfying $0 < \mu < 1$.

The contributions of other group members can be represented as the time series of average contributions excluding individual i:

$$\bar{c}_{-i,t} = \frac{1}{I-1} \sum_{\substack{j=1 \\ j \neq i}}^{I} c_{jt}, \quad t = 1, 2, \ldots, T, \tag{2}$$

where $\bar{c}_{-i,t}$ is the average contribution of all other group members at time t.

3.2 Clustering Multivariate Time-series

A repeated linear public good produces interactive panel data. The decision program of an individual participant may react to the experiences she has made with the choices of others. We provide the algorithm with the exact information that participants receive in the experiment: the development of the choices over time that the respective participant has made over time $\{c_{it}\}_{t=1}^{T}$; and the development of the average choices made by the remaining group members $\{\bar{c}_{-i,t}\}_{t=1}^{T}$. Numerous

methods exist for clustering multivariate time series, each with distinct advantages [see overviews in 81, 96].

Selecting an appropriate *distance measure* is crucial for effective clustering of (time series) data. Two common approaches are the point-wise method and Dynamic Time Warping (DTW). The point-wise method, typically using Euclidean distance, compares values at corresponding time points. While this approach is straightforward and computationally efficient, it is overly sensitive to small misalignment, often leading to inaccurate clustering when sequences are shifted or vary in length. In contrast, DTW enables flexible, non-linear alignments by warping the time axis. For two time series, $\mathbf{x} = \{x_1, x_2, \ldots, x_T\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_T\}$, DTW constructs a cost matrix D where each element $D(i, j)$ represents the cumulative cost of aligning x_i with y_j. The recurrence relation for computing $D(i, j)$ is given by

$$D(i, j) = d(x_i, y_j) + \min \{D(i-1, j),\ D(i, j-1),\ D(i-1, j-1)\},$$

where the local distance $d(x_i, y_j)$ is typically defined as the squared difference $(x_i - y_j)^2$. The optimal DTW distance is then given by $D(T, T)$, found at the bottom-right corner of the matrix, which represents the minimal cumulative alignment cost. This mathematical formulation allows DTW to effectively capture similarities between time series that exhibit temporal distortions or misalignments [see 14, for further details].

A *clustering algorithm* then uses these distances to group data into meaningful clusters. We evaluate three clustering approaches—partition-based, hierarchical, and graph-based—applied to time series data using DTW as the distance metric. DBA-K-Means represents our partition-based clustering approach. This method iteratively assigns sequences to clusters while minimizing DTW-based intra-cluster variance. For hierarchical clustering, we employ an agglomerative approach with Ward's linkage criterion. In this approach, we begin with individual sequences and progressively merges clusters to minimize intra-cluster variance. Working with a precomputed DTW distance matrix, we extract flat clusters by cutting the resulting hierarchical tree at an optimal level, specifically chosen to yield exactly k clusters. Finally, for graph-based clustering, we adopt Spectral Clustering, where the DTW-based similarity matrix is used to construct an affinity graph. The normalized Laplacian of this graph undergoes eigendecomposition, and the data is projected into a lower-dimensional space, where K-means is applied to derive the final clustering.

The optimal *number of clusters k* is determined using internal cluster validation indices (CVIs). These indices assess clustering quality through different combinations of cluster cohesion and separation metrics [6]. We employ three standard CVIs compatible with DTW-based clustering: The Silhouette Score [95], the Davies-Bouldin Index [36], and the Intra-cluster Variance [67]. After normalizing these indices to a 0-1 range, we compute their average score across candidate cluster numbers (2-20) to select the best number.

To facilitate the interpretation of the clusters, we calculate *centroids* via Dynamic Time Warping Barycenter Averaging (DBA) [91], rather than calculating the arithmetic mean [e.g., 50]. The DTW centroid $\mathbf{c} = \{c_1, c_2, \ldots, c_T\}$ minimizes the total DTW distance to all series $\mathbf{x}_k$ in a cluster This process produces a representative time series that preserves the shape and temporal structure of cluster behavior. Initializing DBA with an informed starting point was shown to improve performance [91]. To enhance coherence and visual interpretability, we order UIDs based on their average contribution and select the median UID as the initial reference.

To ensure a comprehensive comparison, we follow a two-step approach. First, we evaluate four clustering categories on unidimensional time series, fixing the number of clusters to isolate method performance. Based on these results, we select the best-performing method. In the second step, we extend the analysis to two-dimensional time series.

3.3 Comparing Partitioning Methods Proposed in the Field

Several approaches have been proposed for constructing taxonomies in experimental settings. One widely used method is Bayesian modeling, which has been applied to public goods games (PGGs) [63] and probabilistic updating tasks [39] to classify individuals based on their decision-making behavior. Another common approach involves finite mixture models, which have been used to categorize individuals in repeated experimental settings such as public goods games [11], beauty contests [16], lottery choices [30], and private information games [17]. These models assume that individuals belong to distinct latent types, each following a different behavioral rule. To address some of the limitations of mixture models, C-Lasso [98] was introduced as an alternative, offering greater flexibility in identifying latent types. It has been applied to classify behavioral types in contest experiments [51]. A different strand of the literature employs hierarchical clustering, using Ward's method and Manhattan distance, to uncover latent behavioral types in experimental settings [50]. More details on these methods are provided in A.

A critical limitation of these methods is their inability to handle temporal shifts in behavioral data. Bayesian models, finite mixture models, and C-Lasso inherently assume that the temporal structure of the data is aligned across individuals. Consequently, behavioral patterns that appear at slightly shifted time points may be misclassified as noise or as entirely different behaviors. Hierarchical clustering can, in principle, be implemented with DTW to allow for temporal shifts. However, in [50], hierarchical clustering was implemented with Manhattan distance, which relies on pointwise alignment, making it sensitive to temporal misalignment.

Another major distinction among these methods lies in their assumptions regarding the nature of behavioral heterogeneity. Hierarchical clustering, as used in Fallucchi et al. [50], does not impose strong assumptions about the structure or characteristics of the identified clusters. In contrast, C-Lasso, as applied in Fallucchi et al. [51] to Tullock contests, assumes that heterogeneity primarily stems from myopic responses to previous-round outcomes, modeled via linear and quadratic effects of opponents' efforts and a win dummy. The Bayesian model specified in Houser et al. [63], selects priors that anchor decision rules in myopic behavior, assuming that, absent strong evidence, choices depend solely on current payoffs. The finite mixture model specified in Bardsley and Moffatt [11] estimates posteriors for four predefined types. Notably, these predefined types were derived from strategy-method data rather than actual gameplay. Some scholars argue that social preferences influence decisions only in the early rounds, with later behavior better explained by payoff-based learning [32]. This suggests that partitioning methods structured around fixed preference types may be suboptimal.

Because these approaches rely on strong prior assumptions, they often introduce noise terms or uninterpreted clusters to accommodate mismatched data. In contrast, unsupervised methods like clustering, using only raw game data—individual and group contributions—avoid such assumptions but yield partitions that may lack interpretability. To address this, we propose a two-step approach: first, DTB-based clustering identifies behavioral patterns without theoretical priors; second, a method uncovers latent intentions driving these behaviors.

3.4 Inverse Hierarchical Q-Learning (HIQL)

In a PGG, each agent's *own contribution* can be treated as the *action* $a \in \mathcal{A}$, and the *observed contributions* of others can be treated as the *state* $s \in \mathcal{S}$. Our setup thus follows published literature on Q-learning in public good games, where the decision-making of players (actions) is based on the cooperation information of their neighborhood (interpreted as states) [111]. Following Zheng et al. [111], we model this setup as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the set of all possible states (observed contributions from others), $\mathcal{A}$ is

the set of all actions (own contributions), $P(s, a, s') = \Pr(s' \mid s, a)$ is the transition function, $r(s, a)$ is the (unknown) reward function, $\gamma \in [0, 1]$ is the discount factor. A policy $\pi(s, a)$ defines the probability of choosing action a at state s. Our goal in *inverse* reinforcement learning is to recover the unknown reward function r from observed game data (state - action - next state triples).

Inverse Q-learning: The setup assumes that the human follows a Boltzmann (softmax) policy $\pi_E(s, a)$ proportional to $\exp(Q(s, a))$, where $Q(s, a)$ is the action-value function satisfying:

$$Q(s,a) \;=\; r(s,a) \;+\; \gamma \sum_{s' \epsilon S} P(s,a,s') \max_{a' \epsilon A} Q(s',a'), \text{ for all } s \epsilon S, a \epsilon A. \tag{3}$$

Given the human game trajectories $\mathcal{D}$, the maximum-likelihood IRL objective can be solved:

$$\max_{r} \; \mathbb{E}_{\xi \sim \mathcal{D}} \big[\log \Pr(\xi \mid \pi_r) \big], \tag{4}$$

subject to $\pi_r(s, a) \propto \exp\big(Q(s, a)\big)$ and the Bellman equation (3). ξ is the sequence of state-action pairs ($\xi = \{(s_0, a_0), \ldots, (s_n, a_n)\}$). From this setup, one obtains the following key identity for any two actions a, b:

$$Q(s,a) \;=\; Q(s,b) \;+\; \log\big(\pi_E(s,a)\big) \;-\; \log\big(\pi_E(s,b)\big). \tag{5}$$

Hierarchical Extension: If there are K distinct *intentions* or *reward functions* $r_1, \ldots, r_K$, each game trajectory is assumed to be generated by a sequence of decisions where, at each time step, the behavior is governed by one of these reward functions. The active reward function for each decision is not directly observed but is assumed to follow an unobserved Markov chain over the intention index i. Specifically, this chain is characterized by an initial distribution $\Pi = (\Pi_1, \ldots, \Pi_K)$ over the K intentions and a transition matrix $\Lambda \in \mathbb{R}^{K \times K}$ whose rows sum to 1, encoding the probabilities of switching between intentions over time.

The overall model thus treats the observed trajectories as being generated by a hierarchical process: the lower level involves action selection via a Boltzmann optimal policy under the currently active reward function, while the higher level governs the evolution of intentions via the Markov chain. To estimate both the reward functions and the hidden intention dynamics, we maximize the joint likelihood of the observed trajectories and the latent intention sequence using an Expectation-Maximization (EM) algorithm.

Fitting Process: We discretize the continuous range $[0, 1]$ of own contributions and others' contributions into five bins, which we denote as the action set $\mathcal{A} = \{a_1, a_2, a_3, a_4, a_5\}$ (five bins in $[0, 1]$). Similarly, we denote the state set as $\mathcal{S} = \{s_1, s_2, s_3, s_4, s_5\}$ (five bins in $[0, 1]$). Note that the binning was done to reduce the size of the Q-table. The original action set has on average 11 bins (in the canonical public good game, participants have an endowment of 10 tokens, and may contribute between 0 and 10 tokens to the joint project). As states are averages, their space is even larger. As defined earlier, the input for the model are time series trajectories (per participant) of state-action-next state triplets.

Following Zhu et al. [112], we employ a multi-stage fitting procedure. For all steps, we set the discount factor to $\gamma = 0.99$. In the first stage, we run multiple HIQL fits with different numbers of intentions $K = 2, \ldots, 5$ on the whole dataset to obtain a global fit. For each fitting, we perform 10 different initializations and select the best-performing initialization. The initial intention distribution Π is initialized uniformly, and the intention transition matrix Λ is initialized according to Zhu et al. [112] as $\Lambda = 0.95 \times I + N(0, 0.05 \times I)$ where N denotes the normal distribution and $I \in \mathbb{R}^{K \times K}$ is the identity matrix. This initial Λ is then normalized so that each row adds up to 1.

In the second stage of the fitting procedure, we obtain an independent but aligned HIQL fit for each cluster. We initialize the parameters for each cluster using the best global fit parameters obtained from all data combined, and then train the algorithm within cluster and for each participant.

As in Zhu et al. [112], we apply a five-fold cross-validation approach, and additionally implement a repeated stratified fold split. The stratification is based on the average contribution of individuals. We bin the clusters into three average contribution bins and ensure that both the training and test sets are sampled from all bins. Furthermore, we repeat this process ten times within each fold to obtain stable posterior estimates per participant.

Aligning Latents: A challenge in this setup, as is generally the case in the context of latent concept estimation, is the fact that latent states can switch across different folds and repeats during cross-validation. This can lead to inconsistencies in interpretation. To ensure consistent labeling of the intentions, we first visualize the input data (states and actions) along with the estimated intentions per participant. We find that while not perfectly aligned, some latents mirror peaks in actions and thus decided to interpret intention 1 primarily from the viewpoint of the participant's action. We then use a peak identifier algorithm [103] to identify peaks in both the action time series and latent intentions time series. Within each fold and repeat, we assess how frequently a specific posterior intention aligns with these detected patterns and designate the intention that shows the highest alignment as intention 1.

Clustering and the Convergence of HIAVI:. Ashwood et al. [9] employed Inverse Reinforcement Learning (IRL) without hierarchical extension to characterize animal behavior. They addressed trajectory variability by clustering (using DBSCAN with Levenshtein distance) to identify trajectories with similar goal maps and time courses before applying their IRL framework. We adopt their procedure of partitioning data before estimating latent intentions. This approach offers several advantages.

Specifically, it supports convergence during the Expectation-Maximization (EM) procedure. Recall that our model assumes that each trajectory is generated by one of K latent intentions which are updated via a forward-backward algorithm (the EM algorithm).

When trajectories are pooled across all participants without pre-clustering, the data can exhibit substantial heterogeneity. This heterogeneity is reflected in the variance of the gradient estimates when optimizing the reward function in the M-step. For example, consider a parameter update (such as in the reward function estimation) modeled by a stochastic gradient step:

$$\theta^{(t+1)} = \theta^{(t)} - \eta \left(\nabla L(\theta^{(t)}) + \epsilon^{(t)} \right),$$

where $\epsilon^{(t)}$ represents the noise in the gradient. In a mixed dataset, the variance of this noise,

$$\sigma^2 = \mathbb{E}\left[\|\epsilon^{(t)}\|^2 \right],$$

can be large due to conflicting signals from heterogeneous trajectories. Standard convergence results then imply that the expected suboptimality after T iterations is bounded as

$$\mathbb{E}\left[L(\theta^{(T)}) - L(\theta^*) \right] \leq O\left(\frac{\sigma}{\sqrt{T}} \right).$$

A larger σ therefore results in slower convergence.

By clustering the trajectories into groups with similar temporal dynamics (similar patterns of own contributions and observed contributions from others), we effectively reduce the within-cluster variance. Denote by σ_k^2 the gradient noise variance in cluster k, where typically $\sigma_k^2 < \sigma^2$. The

corresponding convergence bound for each cluster becomes

$$\mathbb{E}\left[L_k(\theta_k^{(T)}) - L_k(\theta_k^*)\right] \leq O\left(\frac{\sigma_k}{\sqrt{T}}\right).$$

This reduction in variance accelerates the convergence of the EM procedure within each cluster while overall enabling a lower bound as $L_k(\theta_k^{(T)} \leq L(\theta^{(T)})$, leading to more reliable updates for the latent intention assignments and the reward functions.

Furthermore, in the E-step of the EM algorithm, the forward-backward (Baum-Welch) procedure computes the posterior over latent intentions. When trajectories in a cluster share similar dynamics, the likelihoods $p(\xi \mid z_t = i)$ become more consistent across the cluster. This consistency leads to more accurate and less noisy estimates of the posteriors $\Pr(z_t = i \mid \xi)$, which in turn improves the parameter updates in the M-step.

4 Data

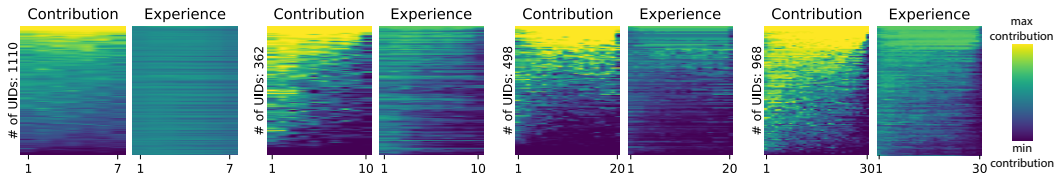

Fig. 1. **Visualization of the raw data.** The y-axis represents participant identifiers (UIDs), while the x-axis tracks the rounds played. Color intensity indicates the size of contributions, normalized between 0 and 1. The panels labeled with 'Contribution,' reflects each player's own contribution. The panels labeled 'Experience,' represents the average contribution of the other players in the group during the preceding round. Within subplot-pairs, UIDs are sorted by their average contribution. Notably, there is heterogeneity with respect to the number of rounds played: The data is divided into subsets with games lasting 7, 10, 20, or 30 rounds. Furthermore, there exists heterogeneity concerning game parameters. Most notably, one study played a 7-round game (first column) with an exceptionally large group size of 100, resulting in unusually homogeneous average contributions.

We use data gathered from standard linear public good games. We collect data from different game length horizons (7, 10, 20, and 30 rounds). All except the 20-round subsets contain games without experimental intervention (like a punishment option, or a chat protocol). Furthermore, we only choose studies, where players observe the *average contribution* of other team members after each round. Our dataset comprises 50,390 observations from 2,938 participants. Table S.1 provides details of the original studies and several focal parameters, such as group size, number of rounds and participants per study. Figure 1 visualizes the raw data as a two-dimensional time series. The first dimension depicts a participant's individual *contribution* (left panels), while the second dimension represents the participant's *experience*, defined as the average contribution of other group members. In the visualization, yellow indicates the maximum contribution, and dark blue indicates no contribution. The x-axis tracks the number of rounds played, with separate panels for games played for 7, 10, 20, or 30 rounds.

The visualization of the raw data reveals several interesting patterns: First, the players' contributions (left panels) show a notable downward trend in contributions, represented by progressively darker shades in the bottom right triangle of each panel. This confirms the well documented phenomenon of gradual decline in cooperative behavior over time in public good games. Furthermore, the groups' average contribution in the subset with a 7-round game displays–as opposed to the

other panels–an almost uniform color pattern. This means that average contributions were very similar across rounds and games. In fact, this is not surprising in light of the fact that this subset consists exclusively of one study in which the group size was exceptionally large: Each group had 100 members.

Next, we want to check whether our dataset is representative of canonical findings. To that end we estimate types based on first-round contributions. Although direct classification through the strategy method would enable immediate comparison with Thöni and Volk [101]'s meta-analysis, we follow Cotla and Petrie [32]'s approach of using first-round contributions as a reliable proxy[2]. Adopting thresholds from Thöni and Volk [101], we categorize normalized contributions $c \in [0, 1]$ as Free Riders: $c \leq 0.1$, Conditional Cooperators $c \in (0.1, 0.9)$, and Full Cooperators: $c \geq 0.9$. In our dataset, Free Riders constitute 11.4% (compared to 19.2% in the meta study), while Conditional Cooperators represent 50.3% (down from 61.3%). Notably, Full Cooperators show a substantial increase, comprising 38.3% of participants, in contrast to merely 10.4% in the earlier meta-analysis. Overall, while variations exist between our dataset and the meta-analysis, the distribution of contribution types remains consistent. Conditional Cooperators continue to dominate, which demonstrates that our data is reasonably representative.

5 Results

5.1 DTW-based Clustering Reveals Six Behavioral Clusters

Our analysis aims at partitioning multivariate time series data while fulfilling three key requirements: allowing for temporal misalignment between series, minimizing the need for prior assumptions, and providing modelling flexibility. We first contrast Dynamic Time Warping (DTW) with standard distance metrics, demonstrating its superiority in capturing temporally shifted patterns. We then systematically compare alternative algorithms with DTW to identify the optimal clustering approach for our behavioral data. Finally, we compare our clustering setup to three methods that have been proposed in the field.

Pattern-Based Outperforms Point-Wise Alignment. In Figure 2A we illustrate the importance of allowing for temporal misalignment in time series on clustering outcomes with selected clusters. For the illustration, we pick three out of six clusters estimated based on the 10-round subset. All clusters are depicted in Figure S.1. We contrast our preferred method DTW (upper row) with traditional Euclidean distance (lower row). Results look strikingly different. For all three example clusters, DTW suggests a pattern where participants start with high contributions and then appear to switch, after which contributions remain consistently lower. Critically, participants switch to low contributions at idiosyncratic points in time. This pattern is most distinct in the right most cluster. In contrast, Euclidean-based methods suggest that this switch occurs for the majority of participants at approximately the same point in time (lower row of Figure 2 A): round 9 (left-most cluster), round 4 (middle cluster), and round 3 (right-most cluster).

The upper and the lower row thus seem to tell a diametrically opposed story: each participant autonomously decides when to give up on cooperation (upper row) vs. clusters of participants are characterized by prototypical (early or late) switching to defection. To discriminate between both stories, in Figure 2 B, separately for each subset we report the point in time when the most pronounced difference in contributions from one period to the next is observed. Irrespective of the duration of play, the pattern is U-shaped. Switching is most frequent early and late in the game. Yet

[2]Muller et al. [85] demonstrate strong correspondence between classifications derived from gameplay data and those obtained from the strategy method. Additionally, Cotla and Petrie [32] establish that first-round contributions in repeated games closely mirror unconditional contributions in corresponding strategy games, suggesting that initial gameplay decisions reflect underlying social preferences.

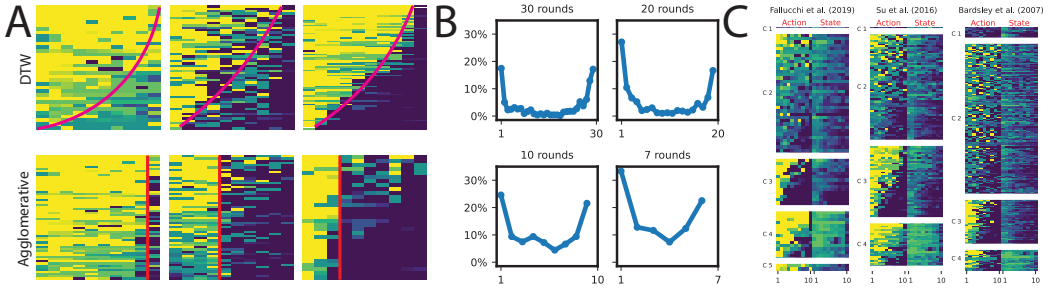

Fig. 2. **Comparative Analysis of Clustering Methods.** (A) Comparing clustering results between Dynamic Time Warping (DTW) and Euclidean distance metrics. The upper row shows DTW clusters and reveals a pyramid-like shape, indicating that participants are characterized by discernible heterogeneous switching points. The lower row shows Euclidean-based clustering which misleadingly suggests that all participants switch at approximately the same point in time. (B) Distribution of switching points across different subsets, demonstrating a consistent U-shaped pattern where strategy shifts predominantly occur in early and late rounds. (C) Comparing our method (DTW and spectral) with methods proposed in the literature. Columns show Hierarchical Clustering [50], Classifier-Lasso [98], and Finite Mixture Models [11]. We always picked the most heterogeneous cluster from the resulting partition and sorted it according to our cluster assignment, labeled on the y-axis. The plots show that our method sorts seemingly messy clusters much more cleanly.

in all intermediate rounds, a discernible fraction of participants switch. This observation strongly speaks against prototypical switching points, and in favour of DTW, which captures the quasi continuous decrease in contributions over time, i.e. the downward trend that is so characteristic for public good experiments [47].

Having identified DTW as the most interpretable distance metric, we evaluated its performance in combination with four clustering algorithms: K-means, DBA K-means, hierarchical clustering, and spectral clustering. A full comparison of all methods on the 10-round subset can be found in Figure S.3. Overall, the differences across clustering algorithms are quite small. The clusters look quite similar, suggesting that the clusters are stable and can more or less be identified by any method, given that the appropriate distance metric is used. As we do want to make a choice for this paper, we focus our analysis on one distinct example where the differences between the methods become most clear, which is the "pyramid cluster" shown in Figure 2 A, top row, most left. This cluster characterizes a strategy of very high initial contributions followed by a distinct switch to extremely low contributions. We argue that this strategy is most clearly captured by spectral clustering. We use this algorithm for further analysis.

To find the optimal number of clusters k we employ three standard cluster evaluation indices compatible with DTW-based clustering. We compute their average score across candidate cluster numbers (2-20) (a detailed visualization of all scores k can be found in Figure S.4). While the statistical optimum occurred at $k = 10$, this solution produced several clusters with fewer than 30 members, making them impractical for meaningful interpretation. This is why we eventually opted for six clusters.

DTW-Based Clustering Uncovers Overlooked Subgroups. In the next step, we compare our approach to existing research. Previous methods for identifying clusters in public good data include Hierarchical Clustering [50], C-Lasso [98], and finite mixture models [11]. We applied these three methods to our dataset, with all the resulting clusters visualized in Figure S.5. A subset of the results are displayed in Figure 2 C. Each method generates a distinct partition. The finite mixture model, for instance, is predefined to fit four clusters: freerider, altruist, reciprocator, and strategist. The

freerider and altruist clusters are well-defined by the method. However, they could also have been easily identified using simple rules (e.g., constant high or low contributions in 90% of rounds). The remaining data points are distributed between the two other clusters, resulting in quite heterogeneous groupings. C-Lasso only converged with five clusters instead of our proposed six, and its first cluster contained merely three participants. Consequently, the remaining clusters were overly crowded.. Hierarchical clustering accommodated six clusters but notably failed to identify clear altruist or freerider clusters.

We argue that our approach (DTW + Spectral Clustering) produces partitions the data more clearly. To demonstrate this, we selected the most heterogeneous cluster from each method and sorted its UIDs according to our DTW algorithm's partition assignments. These results are visualized in Figure 2 C. This analysis reveals that each cluster from existing methods contains distinct subgroups that are clearly differentiated by our DTW-based clustering.

For each of the three methods—Fallucchi et al. [50] (left), Su et al. [98] (middle), and Bardsley and Moffatt [11] (right) in Figure 2C—the largest subgroup corresponds to our Cluster 2 (C2). However, our clusters C3 and C4 are also identified in a substantial number of participant trajectories. Interestingly, our cluster C1 appears only in a small number of trajectories in each of these three subplots, while our C5 is present, only in the leftmost subplot. Notably, our cluster C6 is entirely absent from all three subplots. The complete clustering results are presented in Figure 3, row 3. For now, it is only important to highlight that Cluster 1 represents the group with the lowest average contributions, while Cluster 6 exhibits the highest average contributions. This indicates that the existing methods successfully distinguish between these two clusters. Notably, these clusters align with those theorized in the preference-type literature (Freeriders and Full Contributors). Fallucchi et al. [50] developed their approach primarily using these established preference-types. Therefore, it is not surprising that their method struggles to identify patterns that have not yet been formally theorized. Based on this detailed comparison, we conclude that our DTW-based partitioning method more effectively identifies and differentiates distinct behavioral patterns compared to methods that have been previously proposed in the literature.

5.2 Intentions Successfully Integrate Actions and States

Our clustering results on the 10-round subset are presented in Figure 3, row 3. Visualizations of all subsets and their corresponding clusters can be found in Figure S.6. Since our clustering approach does not impose any theoretical assumptions about participant behavior and is based solely on game data—specifically, contributions and the information provided to participants (i.e., the average contribution of their group members)—the interpretation of these clusters can vary. The resulting clusters distinguish between actions and states. Consequently, researchers must heuristically combine these components when attempting to interpret the clusters. HIQL [112] formalizes this interpretation by applying a transformation that integrates actions and states. It estimates latent intentions that maximize the likelihood of observing the given game trajectories. Notably, the input to this function consists exclusively of actions and states, aligning with our clustering approach. No additional constructs, such as payoffs, are required to fit the model. As a result, the model operates with minimal assumptions about participant behavior.

The first step of fitting the HIQL model is to test which *number of latent intentions* K best explains the observed trajectories. We find that as the number of intentions increases, the test log-likelihood improves steadily, indicating that models with more latent states provide a better fit to the observed trajectories. This improvement is most pronounced when transitioning from $K = 1$ to $K = 2$, suggesting that two latent intentions capture key dynamics of the behavioral data while the increase in the BIC is the smallest at this step (Table 1).

K	Δ Test LL	Δ BIC
$1 \rightarrow 2$	0.6	75.2
$2 \rightarrow 3$	0.4	88.6
$3 \rightarrow 4$	0.2	101.5
$4 \rightarrow 5$	0.2	114.4

Table 1. **The best number of latent intentions is 2.** Evaluating the tradeoff between the number of latent intentions (K), test log-likelihood, and BIC.

	Cl. 1	Cl. 2	Cl. 3	Cl. 4	Cl. 5	Cl. 6
Intention 1→1	0.687	0.525	0.665	0.720	0.724	0.714
Intention 1→2	0.313	0.475	0.335	0.280	0.276	0.286

Table 2. **Transition probabilities of intentions across clusters.** The model is fitted with two intentions, where Intention 2 is defined as 1 - Intention 1. Thus, transition probabilities are reported from the perspective of Intention 1 only.

We then fit the HIQL model with two latent intentions across 5 cross-validation folds with 10 repetitions, selecting for each cluster the fold with the highest test set log-likelihood. The results show that some clusters achieve a better fit than others:

Cluster 6 achieved the best log-likelihood (-0.001), followed by Cluster 5 (-0.091), Cluster 1 (-0.1), Cluster 3 (-0.447), Cluster 2 (-0.557), and Cluster 4 (-0.637).

The best log likelihoods across folds for all subsets are displayed in Table S.3. These differences in model fit align with the behavioral patterns observed across clusters. For instance, Cluster 6 comprises participants who consistently contribute nearly their entire endowment across all periods. The model naturally achieves better fit on these stable, uniform trajectories compared to e.g. Cluster 2, where participants display frequent switching between different contribution levels.

Intentions Provide a New Basis for the Joint Interpretation of States and Actions. We visualize individual-level data from the best-performing fold, selecting six representative participants per cluster to display their latent intentions, states, and actions (Figure 3A, extended in Figure S.9). While states and actions are reported as normalized values and latent variables as probabilities, we present them on a unified scale to facilitate interpretation. Given that the second posterior perfectly mirrors the first, we only plot posterior 1. The visualization uncovers how intentions, actions, and states relate to each other. The posterior probabilities (intention 1) exhibit distinct patterns: sometimes tracking states (e.g. in second plot from left), other times closely following actions (e.g. final rounds of fourth plot), and often synthesizing both signals (e.g. first plot). This diversity of patterns emerges naturally from HIQL's optimization process, which identifies intention trajectories that best explain participant's actions and their experienced states within each cluster. HIQL thus provides a crucial advance in behavioral modeling by mathematically formalizing how individuals integrate states and actions into coherent strategies. Rather than relying on qualitative interpretations of state-action relationships, we offer a rigorous quantitative foundation for understanding strategic decision-making in PGG.

Intentions Provide a Unifying Framework for Both Stable and Switching Decision Patterns. We aggregate individual posteriors within each cluster via Barycenter averaging (Figure 3B, extended in Figure S.8). While the model fitting within clusters prevents direct comparisons of intentions across clusters, their temporal evolution patterns remain comparable. This analysis reveals distinct behavioral signatures across clusters. Clusters 1, 5, and 6 exhibit remarkably stable intentions after initial transitions, despite the fact that the model allows for continuous intention switching. In contrast, Cluster 3 shows a decisive intention shift at time step three, while Cluster 4 displays subtle variations around a declining trend. Most intriguingly, Cluster 2 emerges as a unique case, characterized by frequent oscillations around equal probabilities for both intentions. The corresponding heatmap reveals extreme alternating contributions, behavior that would traditionally have been dismissed as noise. However, our HIQL framework provides a unifying theoretical

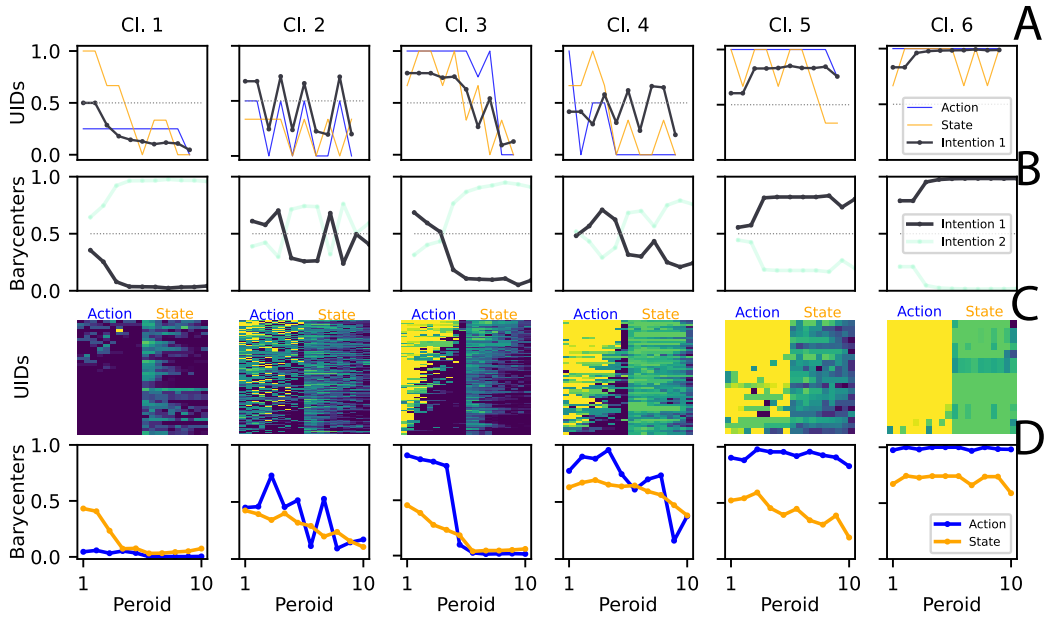

Fig. 3. **Clustering Analysis of Action-State Patterns and Intention Dynamics.** First row: Representative actions, states, and estimated intentions for six different UIDs. Second row: Barycentric averages of intentions for each identified cluster. Third row: Raw data of actions and states for all UIDs within each cluster. Fourth row: Barycenters of actions and states across all UIDs per cluster.

interpretation: while stable trends in other clusters reflect the consistent dominance of one intention, the apparent erratic behavior in Cluster 2 represents frequent intention switching when participants are torn between both. This finding provides the first empirical validation of Houser et al. [63] "type-switching" hypothesis, while extending it within a formal theoretical framework. Our analysis thus bridges a crucial gap between observed behavioral heterogeneity and its underlying cognitive mechanisms.

5.3 Fitted Intentions Can be Interpreted Through A Reinforcement Learning Lens

In the following, we provide a behavioral interpretation of each fitted intention per cluster through the lens of reinforcement learning (RL). It is important to note that intentions are latent concepts and must be consistently tied to an objective variable to ensure reliable identification across folds. We labeled Intention 1 as the one that most closely aligns with participants' actions in terms of peak values. As a consequence, in Cluster 1, for example, Intention 1 has a low probability because, in absolute values, it closely matches the participants' actions. Conversely, the high-probability intention represents its inverse—Intention 2.

Free Riders: Risk-Averse Learning Leads to Persistent Low Cooperation. Cluster 1 exhibits consistently low contribution levels across all rounds, which behavioral economics would traditionally classify as free-riding. However, our analysis reveals a learning pattern: these participants' intentions initially start at medium-high levels before increasing and stabilizing at a high probability steady state, closely (inversely) mirroring their experienced outcomes. From a RL perspective, early negative feedback appears to have quickly consolidated these participants' low-cooperation

strategy. The Q-learning perspective suggests these individuals prioritized minimizing risk over exploring potentially more rewarding cooperative actions.

Volatile Explorers: Active Learning Through Strategic Experimentation. Cluster 2 displays distinctive volatility in both actions and experienced states. The inferred intentions begin at a medium level and exhibit notable oscillations around this initial value throughout the interaction. While traditional behavioral analysis might dismiss such patterns as mere noise or irrationality, our RL framework reveals a more sophisticated underlying process. Drawing on trial-and-error learning models [49, 93], we can interpret these fluctuations as continuous action-value adjustments driven by alternating positive and negative rewards. This behavior exemplifies the classic exploration-exploitation trade-off: participants appear to actively experiment with different contribution levels, strategically adjusting their behavior in response to the volatile social environment rather than settling on a fixed strategy.

Threshold Switchers: Strategic Transition from Optimistic Cooperation to Defensive Play. Cluster 3 is characterized by a clear strategic transition in intentions. Action trajectories reveal notable heterogeneity in switching timing - some participants transition early, others later, but all eventually shift from near-maximal to minimal contributions. The estimated latent intentions clearly signal this strategic pivot through two opposing states: an initially dominant high-contribution intention gradually gives way to a low-cooperation intention. Through a RL lens, the early high contributions likely reflect an exploratory phase driven by optimistic reward expectations. However, as consistently declining states generate negative reinforcement signals, a critical threshold appears to be breached in the Q-learning dynamics. This triggers a decisive shift from a high-cooperation exploitation strategy to a risk-minimizing approach characterized by minimal contributions.

Complex Fluctuators: Adaptive Learning with Dynamic State-Action-Intention Relationships. The aggregated Intention 1 dynamics of Cluster 4 follow a generally decreasing trend punctuated by local peaks and valleys. Through a RL lens, these patterns reflect sophisticated exploration-exploitation dynamics and adaptive responses to feedback. Initially high actions and intentions suggest an exploratory phase motivated by potential cooperative gains. However, declining states signal diminishing returns from cooperation, triggering recalibration of both actions and intentions. Drawing on research in adaptive aspiration levels [65, 83] and Q-learning dynamics [106], we interpret the abrupt transitions as threshold effects: once expected rewards fall below critical aspiration levels, participants rapidly shift from high-cooperation exploitation to loss-minimizing strategies.

Consistent Cooperators: Positive Reinforcement Stabilizes High Cooperation. Cluster 6 (note that we have not yet interpreted Cluster 5) represents cooperators who experience consistently rewarding states and maintain maximal contributions throughout the game. From an RL perspective, early positive outcomes appear to trigger a self-reinforcing cycle of continued cooperation. These participants, having discovered a high-reward strategy early on, exploit it consistently rather than exploring alternatives. This interpretation is further supported by the intention dynamics, which start with high probabilities and quickly stabilize at maximum probability.

Unconditional Cooperators: Resilient High Contributors Despite Limited Reciprocity. Cluster 5 is characterized by persistently high contributions, despite experiencing medium to low observed states throughout the game. Behavioral economists have traditionally classified these individuals as unconditional cooperators, as they continue contributing at near-maximal levels even in the face of limited reciprocity. From an RL perspective, these individuals appear to possess high aspiration levels for cooperation [65, 83]. While occasional negative reinforcement signals—manifested as

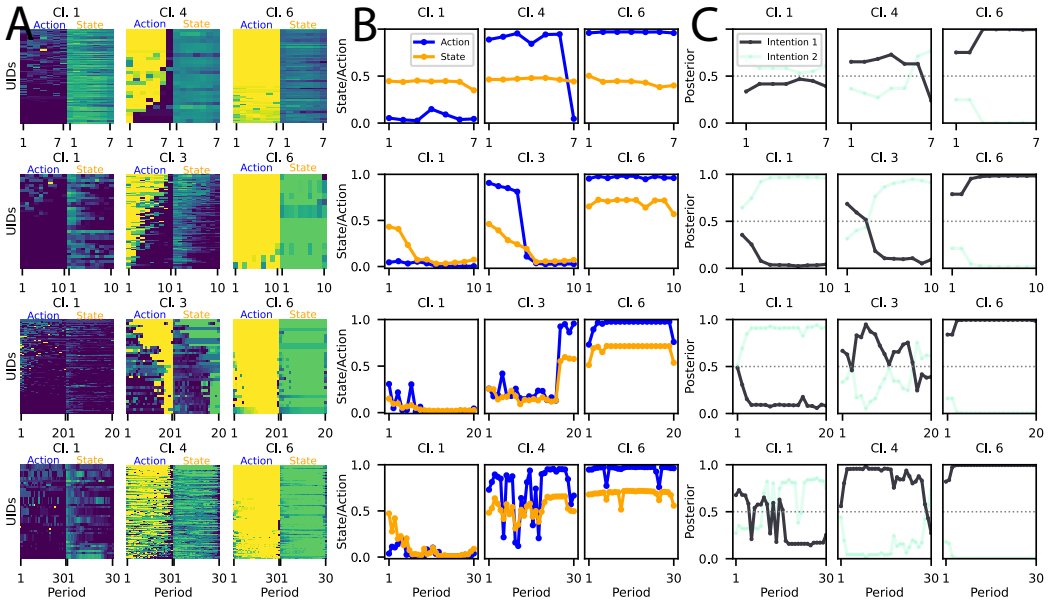

Fig. 4. **Comparison of behavior and intentions across time horizons.** The rows display states/actions and intentions for four different data subsets corresponding to varying game lengths. The leftmost block (columns 1–3) shows the raw data of actions and states across all time horizons. The middle block (columns 4–6) presents the barycenter averages of actions and states for each cluster and time horizon. The rightmost block (columns 7–9) illustrates the barycenter averages of intentions per cluster and time horizon. Only three clusters per subset are shown—specifically, those that exhibit similar patterns across the subsets. A full comparison of all clusters can be found in Figures S.6, S.7, and S.8.

transient drops in contributions—could discourage cooperation, participants in this cluster exhibit low sensitivity to this kind of adverse feedback. Individuals in this cluster maintain cooperation despite unfavorable conditions, with a latent intention 1 probability of around 0.8. This suggests a form of "informed cooperation", where participants recognize that their strategy may not be immediately rewarded but still persist in their contributions. In contrast, Cluster 6 comprises individuals, whose cooperation is fully reinforced by favorable social feedback. These individuals exhibit intention probabilities approaching 1.0, indicative of absolute confidence in their cooperative strategy.

Our findings reveal a critical behavioral insight: intentions shape decision-making in ways that actions and states alone cannot explain. While Clusters 1 and 5 appear static and independent of group contributions from a action-state/ Behavioral Economics perspective, an RL-based interpretation shows that participants do integrate group behavior into their decisions. However, their predominant intention is so firmly held—far from the neutral switching probability (0.5)—that group behavior does not alter their choices. This highlights a robust internal strategy, where individuals acknowledge external information yet remain committed to their chosen course of action.

5.4 Intentions Highlight the Stability of Behavior Across Time Horizons

Our dataset contains experimental games with various round lengths (7, 10, 20, and 30 rounds). Each subset exhibits idiosyncratic characteristics. In particular, the 7-round subset is unique because group sizes were extremely large (sometimes up to 100 participants), which tends to stabilize

the average experiences across rounds. In contrast, the 20-round subset incorporates punishment mechanisms, further differentiating its dynamics.

The learning model literature has found quite some differences across games with different time horizons. Empirical and simulation studies within best-reply learning-frameworks [13, 33, 60, 82, 97] find that longer round lengths tend to support the emergence of cooperative behavior, whereas shorter games with clear endpoints often lead to increased defection [90, 94]. Also from a RL perspective, the total number of rounds are expected to strongly influences behavioral dynamics. In shorter games, there may be insufficient time for agents' action probabilities (or "attractions") to converge to a stable (and possibly cooperative) pattern, resulting in more exploratory or noisy behavior. Conversely, in longer games, agents have more opportunities to accumulate reinforcement from successful interactions [65, 66].

Figure 4 illustrates the data subsets across different game lengths (7, 10, 20, and 30 rounds), with each length shown in a separate row.

The most obvious observation from the comparison of our data across subsets is that both actions and intentions become, on average, less smooth as the game proceeds. From a RL perspective, this phenomenon is not surprising: With longer time horizons, agents are afforded more opportunities to explore alternative actions before settling into a stable strategy [99]. Furthermore, as agents accumulate more experiences over longer games, the continual updating of their action-value estimates introduces transient fluctuations. These fluctuations can lead to less smooth trajectories in both the actions taken and the intentions formed [65, 66].

Another interesting observation is with respect to the clusters with the lowest and highest cooperation per subset. Recall that we ordered the clusters by their average action: cluster 1 corresponds to the lowest average action (i.e., predominantly defection) and cluster 6 to the highest (i.e., predominantly cooperation). Interestingly, clusters 1 and 6 show considerable similarity across subsets. Although the 7-round subset in cluster 6 shows action centroids that hover around 0.5—suggesting less favorable experiences—the overall contribution of cluster 6 is consistent across all subsets. This implies that fully cooperative behavior is independent of the time horizon. This adds an important insight to the literature: *On average* from a best-reply and RL perspective longer effective horizons promote the emergence and stabilization of cooperative clusters [90, 94]. Our analysis however identifies a small group of participants whose cooperative tendencies appear independent of the game's duration.

The comparison across game horizons is more complex for cluster 1. For example, the 30-round subset shows pronounced local variations in both actions and intentions. This variability implies that, especially in longer games, the pattern of defection is not as unambiguous as full cooperation. The evolving nature of the intentions in cluster 1, 30-round subset suggests that players may be adapting their behavior as they learn more about the game. From a public policy perspective this is a very interesting point because it means that freeriding is not as fixed and hard to change as previously thought. Longer time horizons might be one way to address the problem. Note that the analysis of intentions not actions leads to this conclusion and therefore demonstrates the added value of this analysis.

Another notable observation is the *threshold switching* behavior observed across different round-lengths. For instance, in the 10-round subset, cluster 3 exhibits a clear threshold switch, with a sudden transition in players' actions occurring at a specific round. In the 7-round subset, a similar phenomenon appears in cluster 4; In contrast, the 30-round subset displays a more complex pattern for the threshold-switchers (cluster 4), with two distinct abrupt shifts—one occurring close to midgame and another near the end. Interestingly, the *intention* remains stably high until a marked drop in the final two rounds. Thus the *intentions* of the threshold switchers are quite comparable

across the different time horizons and similarly suggest that this behavior might not be impacted by the time horizon of the game.

6 Discussion

Main Insights. Our findings reveal several key insights. First, we demonstrate that pattern-based alignment (via DTW) outperforms point-wise alignment (via Euklidian distance) in interactive behavioral trajectory analysis.

When comparing our clustering setup to existing literature, we find that our approach achieves the clearest partitioning, uncovering behavioral patterns previously overlooked.

We introduce HIQL as the first formal integration of state and action information, establishing a novel interpretative framework for behavioral trajectories in the economic literature. By modeling behavior as latent intentions that can switch at each time step, we provide, for the first time, a unifying theoretical model for interpreting all behavioral clusters in social dilemma games. What is often dismissed as erratic behavior or noise can instead be understood as frequent intention switching, particularly when competing intentions have similar adoption probabilities.

Interpreting the fitted intentions through a reinforcement learning lens further refines our understanding of decision-making processes in public good games. Freeriders and unconditional cooperators, traditionally seen as acting solely based on social preferences without considering group behavior, actually integrate group dynamics into their decisions. However, their predominant intention remains so stable—far from the neutral switching probability (0.5)—that external factors do not influence their choices.

We also find that some behavioral patterns remain independent of the game's time horizon. While longer effective horizons generally promote cooperative clusters, a subset of participants—unconditional cooperators—exhibit cooperative tendencies irrespective of game duration. This supports the behavioral economics classification of their behavior as a strategic type.

Lastly, we observe that low contributors, typically classified as freeriders, exhibit variability in their probability of choosing low-contribution intentions, with this variability increasing as round length extends. From a public policy perspective, this suggests that freeriding is not as fixed or unchangeable as previously thought. Longer time horizons may provide a viable mechanism for mitigating freeriding behavior. Notably, this conclusion emerges from analyzing intentions rather than actions, underscoring the added value of our approach.

Low contributors, often labeled freeriders, show increasing variability in low-contribution intentions as the time horizon increases. This suggests freeriding is more flexible than assumed, and might be an important new insight for behavioral public policy.

Contributions to Literature. This paper advances several streams of literature:

First, we provide novel insights into the literature on social preference-types in public goods games (PGG). We identify two critical gaps: (1) the lack of a clear distinction between types estimated from strategy method versus gameplay data, and (2) the limitations of current methods in detecting trends and handling temporal misalignments in behavioral patterns.

Second, we extend the learning model literature on social dilemma games by demonstrating the critical importance of patterned heterogeneity—a phenomenon that has been largely overlooked in this field, as they do not fit their modeles on pre-clustered data.

Third, we contribute a broader theoretical perspective that bridges multiple research streams and opens new avenues for future investigation. While psychological research has extensively documented how humans deviate from their preferences—due to self-control problems or selective attention—this insight has remained largely implicit in social dilemma game analysis. Current approaches either infer motivations from observed behavior (e.g., conditional cooperation as a

generalizable principle) or directly model behavioral response functions through learning models. Previous attempts to rationalize unexpected behaviors have relied on noise parameters or undefined "other" clusters. We propose a novel framework that explicitly models these previously categorized "noise" or "other" clusters as switches in latent intentions. This reconceptualization offers a new way to understanding behavioral heterogeneity in social dilemma games.

Fourth, we potentially make a modest contribution to the computational neuroscience literature. While clustering has been applied before fitting inverse reinforcement learning frameworks to animal behavior, these approaches often appear ad hoc and are not explicitly optimized for the data at hand. Given that this field also deals with behavioral data, it might benefit from employing clustering methods specifically suited to capturing temporal shifts in behavior.

Future Work. This study opens several avenues for further research, particularly in modeling human behavior. Our current approach adapts models originally designed for animal behavior in non-social, way-finding contexts. In contrast, human behavior involves additional complexities such as aspirations, emotions, and other affective dynamics. Future work should aim to formally integrate these elements from the reinforcement learning literature into the modeling framework.

We fit the HIQL model within clusters, aligning with established econometric practices for uncovering latent structures in panel data, and paralleling computational neuroscience methods where inverse reinforcement learning models are applied to animal behavior post-clustering [9]. Yet, the implications of not clustering the data have not been systematically investigated. Future research could address this gap by examining how model outcomes—such as stability and deviation—are affected when clustering is omitted.

In addition, following the recommendation of Zhu et al. [112], our models are fitted with $\gamma = 0.99$ (agents with high foresight). It would be informative to systematically vary γ, potentially within clusters, to study its effect on latent intentions. However, implementing lower γ values in the current HIQL framework introduces instabilities and convergence issues, suggesting that some form of numerical smoothing may be required.

Furthermore, our current model does not take first-round contributions into account when estimating latent states. Specifically, the transition probabilities are defined as (state, action, next state), consistent with standard Q-learning setups. In our linear public goods game (PGG), the first action is unconditional, so that the first triplet effectively omits the first-round contribution. A more accurate estimation of latent states might be achieved by explicitly incorporating this unconditional first-round contribution.

Another important consideration is the initialization of intentions. For instance, in the unconditional cooperator cluster, intention 1 may be more realistically initialized at a high level and maintained, rather than beginning at a moderate level and increasing sharply in the first round. This observation implies that the initialization of the transition matrix could more closely follow the unconditional contribution pattern rather than relying on random draws from a normal Gaussian distribution. Future research can explore alternative initialization strategies to stabilize intention patterns further.

Finally, our analysis required substantial binning of observed behavior—in most cases, halving the number of bins—to manage the Q-table size and ensure tractability. While binning is a practical necessity, it may obscure subtle but important variations in behavior. Future simulation studies could investigate the impact of different binning strategies, allowing for a formal comparison between models using the original versus a reduced number of bins.

References

[1] Johannes Abeler and Daniele Nosenzo. 2015. Self-selection into laboratory experiments: Pro-social motives versus monetary incentives. *Experimental Economics* 18, 2 (2015), 195–214. https://doi.org/10.1007/s10683-014-9397-9

[2] Jason A Aimone, Laurence R Iannaccone, Michael D Makowsky, and Jared Rubin. 2013. Endogenous group formation via unproductive costs. *The Review of Economic Studies* 80, 4 (2013), 1215–1236. https://doi.org/10.1093/restud/rdt017

[3] Mansour Alyahyay, Gabriel Kalweit, Maria Kalweit, Golan Karvat, Julian Ammer, Artur Schneider, Ahmed Adzemovic, Andreas Vlachos, Joschka Boedecker, and Ilka Diester. 2023. Mechanisms of premotor-motor cortex interactions during goal directed behavior. *bioRxiv* (2023), 2023–01. Preprint.

[4] Simon P Anderson, Jacob K Goeree, and Charles A Holt. 2004. Noisy directional learning and the logit equilibrium. *The Scandinavian Journal of Economics* 106, 3 (2004), 581–602.

[5] James Andreoni. 1995. Cooperation in public goods experiments: Kindness or confusion. *American Economic Review* 85, 4 (1995), 891–904.

[6] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M Pérez, and Iñigo Perona. 2013. An Extensive Comparative Study of Cluster Validity Indices. *Pattern Recognition* 46, 1 (2013), 243–256.

[7] Jasmina Arifovic and John Ledyard. 2012. Individual evolutionary learning, other-regarding preferences, and the voluntary contributions mechanism. *Journal of Public Economics* 96, 9 (2012), 808–823.

[8] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (2021), 103500.

[9] Zoe Ashwood, Aditi Jha, and Jonathan W Pillow. 2022. Dynamic inverse reinforcement learning for characterizing animal behavior. *Advances in Neural Information Processing Systems* 35 (2022), 29663–29676.

[10] Zoe C. Ashwood, Nicholas A. Roy, Iris R. Stone, International Brain Laboratory, Anne E. Urai, Anne K. Churchland, Alexandre Pouget, and Jonathan W. Pillow. 2022. Mice alternate between discrete strategies during perceptual decision-making. *Nature Neuroscience* 25, 2 (2022), 201–212.

[11] Nicholas Bardsley and Peter G Moffatt. 2007. The experimetrics of public goods: Inferring motivations from contributions. *Theory and Decision* 62, 2 (2007), 161–193.

[12] R. C. Bayer, E. Renner, and R. Sausgruber. 2013. Confusion and learning in the voluntary contributions game. *Experimental Economics* 16 (2013), 478–496.

[13] Ulrich Berger. 2007. Brown's Original Fictitious Play. *Journal of Economic Theory* 135, 1 (2007), 572–578.

[14] Donald J. Berndt and James Clifford. 1994. Using Dynamic Time Warping to Find Patterns in Time Series.. In *KDD workshop*, Vol. 10. Seattle, WA, USA:, 359–370.

[15] Friedel Bolle and Jonathan HW Tan. 2021. Behavioral types of the dark side: identifying heterogeneous conflict strategies. *Journal of the Economic Science Association* 7, 1 (2021), 49–63.

[16] Antoni Bosch-Domenech, Jose G Montalvo, Rosemarie Nagel, and Albert Satorra. 2010. A finite mixture analysis of beauty-contest data using generalized beta distributions. *Experimental Economics* 13, 4 (2010), 461–475.

[17] Isabelle Brocas, Juan D Carrillo, Stephanie W Wang, and Colin F Camerer. 2014. Imperfect choice or imperfect attention? Understanding strategic thinking in private information games. *Review of Economic Studies* 81, 3 (2014), 944–970.

[18] Roberto Burlando and Francesco Guala. 2005. Heterogeneous agents in public goods experiments. *Experimental Economics* 8, 1 (2005), 35–54.

[19] Maxwell Burton-Chellew and Claire Guérin. 2022. Self-interested learning is more important than fair-minded conditional cooperation in public-goods games. *Evolutionary Human Sciences* 4 (2022), E46.

[20] Maxwell N Burton-Chellew, Victoire D'Amico, and Claire Guérin. 2022. The Strategy Method Risks Conflating Confusion with a Social Preference for Conditional Cooperation in Public Goods Games. *Games* 13, 6 (2022), 69.

[21] M. N. Burton-Chellew, C. El Mouden, and S. A. West. 2016. Conditional cooperation and confusion in public goods experiments. *Proceedings of the National Academy of Sciences* 113, 6 (2016), 1291–1296.

[22] Maxwell N Burton-Chellew, Heinrich HH Nax, and Stuart A West. 2015. Payoff-based learning explains the decline in cooperation in public goods games. *Proceedings of the Royal Society of London B: Biological Sciences* 282, 1801 (2015), 20142678.

[23] M. N. Burton-Chellew and S. A. West. 2013. Prosocial preferences do not explain human cooperation in public goods games. *Proceedings of the National Academy of Sciences* 110, 6 (2013), 216–221.

[24] Maxwell N Burton-Chellew and Stuart A West. 2013. Prosocial preferences do not explain human cooperation in public-goods games. *Proceedings of the National Academy of Sciences* 110, 1 (2013), 216–221.

[25] Maxwell N Burton-Chellew and Stuart A West. 2021. Payoff-based learning best explains the rate of decline in cooperation across 237 public-goods games. *Nature Human Behaviour* 5, 2 (2021), 226–233.

[26] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. 2020. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110, 10 (2020), 3267–3297.

[27] C. F. Camerer and E. Fehr. 2006. When does "economic man" dominate social behavior? *Science* 311, 5757 (2006), 47–52.

[28] Ananish Chaudhuri. 2011. Sustaining Cooperation in Laboratory Public Goods Experiments: a Selective Survey of the Literature. *Experimental Economics* 14, 1 (2011), 47–83.

[29] Jiayu Chen, Tian Lan, and Vaneet Aggarwal. 2023. Option-aware adversarial inverse reinforcement learning for robotic control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5902–5908.

[30] Anna Conte, John D Hey, and Peter G Moffatt. 2011. Mixture models of choice under risk. *Journal of Econometrics* 162, 1 (2011), 79–88.

[31] David J Cooper and Cathy K Stockman. 2002. Fairness and learning: An experimental examination. *Games and Economic Behavior* 41, 1 (2002), 26–45.

[32] Chenna Reddy Cotla and Ragan Petrie. 2019. Social Preferences and Payoff-Based Learning Explain Contributions in Repeated Public Goods Games. (2019).

[33] Antoine-Augustin Cournot. 1838. *Researches sur les Principes Mathématiques de la Théorie des Richesses*. Hachette, Paris.

[34] Robin Cubitt, Simon Gächter, and Simone Quercia. 2017. Conditional cooperation and betrayal aversion. *Journal of Economic Behavior & Organization* 141 (2017), 110–121. https://doi.org/10.1016/j.jebo.2017.06.013

[35] Alexis Dariel and Nikos Nikiforakis. 2014. Cooperators and reciprocators: A within-subject analysis of pro-social behavior. *Economics Letters* 122, 2 (2014), 163–166. https://doi.org/10.1016/j.econlet.2013.10.033

[36] David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1, 2 (1979), 224–227.

[37] Johannes Diederich, Timo Goeschl, and Israel Waichman. 2016. Group Size and the (In)Efficiency of Pure Public Good Provision. *European Economic Review* 85 (2016), 272–287.

[38] Arthur Dolgopolov. 2024. Reinforcement learning in a prisoner's dilemma. *Games and Economic Behavior* 144 (2024), 84–103.

[39] Mohamed A El-Gamal and David M Grether. 1995. Are people Bayesian? Uncovering behavioral strategies. *J. Amer. Statist. Assoc.* 90, 432 (1995), 1137–1145.

[40] Christoph Engel. 2020. Estimating Heterogeneous Reactions to Experimental Treatments. (2020). Working paper.

[41] Christoph Engel, Martin Beckenkamp, Andreas Glöckner, Bernd Irlenbusch, Heike Hennig-Schmidt, Sebastian Kube, Michael Kurschilgen, Alexander Morell, Andreas Nicklisch, Hans-Theo Normann, et al. 2014. First impressions are more important than early intervention: qualifying broken windows theory in the lab. *International Review of Law and Economics* 37 (2014), 126–136.

[42] Christoph Engel, Sebastian Kube, and Michael Kurschilgen. 2021. Managing expectations: How selective information affects cooperation and punishment in social dilemma games. *Journal of Economic Behavior & Organization* 187 (2021), 111–136.

[43] Christoph Engel and Michael Kurschilgen. 2013. The coevolution of behavior and normative expectations: An experiment. *American law and economics review* 15, 2 (2013), 578–609.

[44] Christoph Engel and Michael Kurschilgen. 2019. Aim High or Aim Low: The Power of Self-Set Normative Goals in a Social Dilemma. (2019). Draft version, January 2019.

[45] Christoph Engel and Michael Kurschilgen. 2020. The fragility of a nudge: the power of self-set norms to contain a social dilemma. *Journal of Economic Psychology* 81 (2020), 102293.

[46] Christoph Engel and Bettina Rockenbach. 2014. Give everybody a voice! The power of voting in a public goods experiment with externalities. *The Power of Voting in a Public Goods Experiment with Externalities (November 2014). MPI Collective Goods Preprint* 2014/16 (2014).

[47] Christoph Engel and Bettina Rockenbach. 2024. What Makes Cooperation Precarious? *Journal of Economic Psychology* (2024), 102712.

[48] Ido Erev, Yoella Bereby-Meyer, and Alvin E Roth. 1999. The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior & Organization* 39, 1 (1999), 111–128.

[49] Ido Erev and Alvin E Roth. 1998. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* (1998), 848–881.

[50] Francesco Fallucchi, Richard A Luccasen, and Theodore L Turocy. 2019. Identifying discrete behavioural types: a re-analysis of public goods game contributions by hierarchical clustering. *Journal of the Economic Science Association* 5, 2 (2019), 238–254.

[51] Francesco Fallucchi, Andrea Mercatanti, and Jan Niederreiter. 2021. Identifying types in contest experiments. *International Journal of Game Theory* 50 (2021), 39–61.

[52] Shaheen Fatima, Nicholas R Jennings, and Michael Wooldridge. 2024. Learning to resolve social dilemmas: a survey. *Journal of Artificial Intelligence Research* 79 (2024), 895–969.

[53] E. Fehr and S. Gächter. 2002. Altruistic punishment in humans. *Nature* 415, 6868 (2002), 137–140.

[54] P. J. Ferraro and C. A. Vossler. 2010. The source and significance of confusion in public goods experiments. *B.E. Journal of Economic Analysis and Policy* 10 (2010), 53.

[55] Urs Fischbacher and Simon Gächter. 2010. Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review* 100, 1 (2010), 541–556. https://doi.org/10.1257/aer.100.1.541

[56] Urs Fischbacher, Simon Gächter, and Ernst Fehr. 2001. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71, 3 (2001), 397–404. https://doi.org/10.1016/S0165-1765(01)00394-9

[57] Urs Fischbacher, Simon Gächter, and Simone Quercia. 2012. The behavioral validity of the strategy method in public good experiments. *Journal of Economic Psychology* 33, 4 (2012), 897–913. https://doi.org/10.1016/j.joep.2012.04.002

[58] Urs Fischbacher, Simon Schudy, and Sabrina Teyssier. 2014. Heterogeneous reactions to heterogeneity in returns from public goods. *Social Choice and Welfare* 43, 1 (2014), 195–217. https://doi.org/10.1007/s00355-013-0763-x

[59] Toke R Fosgaard, Lars G Hansen, and Erik Wengström. 2014. Understanding the nature of cooperation variability. *Journal of Public Economics* 120 (2014), 134–143. https://doi.org/10.1016/j.jpubeco.2014.09.004

[60] Drew Fudenberg and David K. Levine. 1998. *The Theory of Learning in Games.* MIT Press, Cambridge, MA.

[61] Simon Gächter, Felix Kölle, and Simone Quercia. 2017. Reciprocity and the tragedies of maintaining and providing the commons. *Nature Human Behaviour* 1 (2017), 650–656. https://doi.org/10.1038/s41562-017-0191-5

[62] Benedikt Herrmann and Christian Thöni. 2009. Measuring conditional cooperation: A replication study in Russia. *Experimental Economics* 12, 1 (2009), 87–92. https://doi.org/10.1007/s10683-008-9197-1

[63] Daniel Houser, Michael Keane, and Kevin McCabe. 2004. Behavior in a dynamic decision problem: an analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica* 72, 3 (2004), 781–822.

[64] D. Houser and R. Kurzban. 2002. Revisiting kindness and confusion in public goods experiments. *American Economic Review* 92, 4 (2002), 1062–1069.

[65] Luis R. Izquierdo and Segismundo S. Izquierdo. 2008. Dynamics of the Bush-Mosteller Learning Algorithm in 2x2 Games. In *Reinforcement Learning: Theory and Applications.* IntechOpen.

[66] Luis R. Izquierdo, Segismundo S. Izquierdo, Nigel M. Gotts, and J. Gary Polhill. 2007. Transient and Asymptotic Dynamics of Reinforcement Learning in Games. *Games and Economic Behavior* 61, 2 (2007), 259–276.

[67] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys (CSUR)* 31, 3 (1999), 264–323.

[68] Marco A Janssen and Toh-Kyeong Ahn. 2006. Learning, signaling, and social preferences in public-good games. *Ecology and Society* 11, 2 (2006), 21.

[69] Gabriel Kalweit, Maria Huegle, Moritz Werling, and Joschka Boedecker. 2020. Deep inverse Q-learning with constraints. *Advances in Neural Information Processing Systems* 33 (2020), 14291–14302.

[70] Kenju Kamei. 2012. From locality to continent: A comment on the generalization of an experimental study. *Journal of Socio-Economics* 41, 2 (2012), 207–210. https://doi.org/10.1016/j.socec.2011.12.005

[71] Timo Klein. 2021. Autonomous algorithmic collusion: Q-learning under sequential pricing. *The RAND Journal of Economics* 52, 3 (2021), 538–558.

[72] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. 2024. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science* 19, 5 (2024), 827–838.

[73] Martin G Kocher, Todd L Cherry, Stephan Kroll, Robert J Netzer, and Matthias Sutter. 2008. Conditional cooperation on three continents. *Economics Letters* 101, 3 (2008), 175–178. https://doi.org/10.1016/j.econlet.2008.07.015

[74] Michael Kosfeld, Akira Okada, and Arno Riedl. 2009. Institution Formation in Public Goods Games. *American Economic Review* 99, 4 (2009), 1335–55.

[75] David M Kreps, Paul Milgrom, John Roberts, and Robert Wilson. 1982. Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory* 27, 2 (1982), 245–252.

[76] Sateesh Kumar, Jonathan Zamora, Nicklas Hansen, Rishabh Jangir, and Xiaolong Wang. 2023. Graph inverse reinforcement learning from diverse videos. In *Conference on Robot Learning*. PMLR, 55–66.

[77] Robert Kurzban and Daniel Houser. 2001. Individual differences in cooperation in a circular public goods game. *European Journal of Personality* 15, S1 (2001), S37–S52.

[78] Robert Kurzban and Daniel Houser. 2005. Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences* 102, 5 (2005), 1803–1807.

[79] Minhae Kwon, Saurabh Daptardar, Paul R. Schrater, and Xaq Pitkow. 2020. Inverse rational control with partially observable continuous nonlinear dynamics. In *Advances in Neural Information Processing Systems*, Vol. 33. 7898–7909.

[80] J. Ledyard. 1995. Public goods: A survey of experimental research. In *Handbook of Experimental Economics*, J. Kagel and A. Roth (Eds.). Princeton University Press, 253–279.

[81] T. Warren Liao. 2005. Clustering of Time Series Data – a Survey. *Pattern Recognition* 38, 11 (2005), 1857–1874.

[82] R. D. Luce and H. Raiffa. 1957. *Games and Decisions: Introduction and Critical Survey.* John Wiley and Sons, New York.

[83] M. W. Macy and A. Flache. 2002. Learning Dynamics in Social Dilemmas. *Proceedings of the National Academy of Sciences* 99, suppl 3 (2002), 7229–7236.

[84] Michael D Makowsky, William H Orman, and Sandra J Peart. 2014. Playing with other people's money: Contributions to public goods by trustees. *Journal of Behavioral and Experimental Economics* 53 (2014), 44–55. https://doi.org/10.1016/j.socec.2014.08.003

[85] Laurent Muller, Martin Sefton, Richard Steinberg, and Lise Vesterlund. 2008. Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior & Organization* 67, 3 (2008), 782–793.

[86] Payam Nasernejad, Tarek Sayed, and Rushdi Alsaleh. 2023. Multiagent modeling of pedestrian-vehicle conflicts using adversarial inverse reinforcement learning. *Transportmetrica A: Transport Science* 19, 3 (2023), 2061081.

[87] Andrew Y. Ng and Stuart J. Russell. 2000. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, Vol. 1. 2.

[88] Nikos Nikiforakis and Hans-Theo Normann. 2008. A Comparative Statics Analysis of Punishment in Public-Good Experiments. *Experimental Economics* 11, 4 (2008), 358–369.

[89] A. Norenzayan and A. F. Shariff. 2008. The origin and evolution of religious prosociality. *Science* 322 (2008), 58–62.

[90] Hans-Theo Normann and Brian Wallace. 2012. The Impact of the Termination Rule on Cooperation in a Prisoner's Dilemma Experiment. *International Journal of Game Theory* 41, 3 (2012), 707–718.

[91] François Petitjean, Alain Ketterlin, and Pierre Gançarski. 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern recognition* 44, 3 (2011), 678–693.

[92] Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. 2021. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *eLife* 10 (2021), e66175. https://doi.org/10.7554/eLife.66175

[93] A. E. Roth and I. Erev. 1995. Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term. *Games and Economic Behavior* 8, 1 (1995), 164–212.

[94] Alvin E. Roth and J. Keith Murnighan. 1978. Equilibrium Behavior and Repeated Play of the Prisoner's Dilemma. *Journal of Mathematical Psychology* 17, 2 (1978), 189–198.

[95] Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.

[96] Alexis Sardá-Espinosa. 2017. Comparing Time-Series Clustering Algorithms in R using the dtwclust Package. *R Package Vignette* 12 (2017), 41.

[97] Lloyd S. Shapley. 1964. Some Topics in Two-Person Games. In *Advances in Game Theory*, M. Dresher, L. S. Shapley, and A. W. Tucker (Eds.). Annals of Mathematics Studies, Vol. 52. Princeton University Press, 1–29.

[98] Liangjun Su, Zhentao Shi, and Peter CB Phillips. 2016. Identifying latent structures in panel data. *Econometrica* 84, 6 (2016), 2215–2264.

[99] R. S. Sutton and A. G. Barto. 2018. *Reinforcement Learning: An Introduction.* MIT Press.

[100] Christian Thöni, Jean-Robert Tyran, and Erik Wengström. 2012. Microfoundations of social capital. *Journal of Public Economics* 96, 7–8 (2012), 635–643. https://doi.org/10.1016/j.jpubeco.2012.04.003

[101] Christian Thöni and Stefan Volk. 2018. Conditional cooperation: Review and refinement. *Economics Letters* 171 (2018), 37–40.

[102] Niels van Miltenburg, Vincent Buskens, Davide Barrera, and Werner Raub. 2014. Implementing punishment and reward in the public goods game: The effect of individual and collective decision rules. *International Journal of the Commons* 8, 1 (2014), 47–78. https://doi.org/10.18352/ijc.397

[103] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, Christopher J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17, 3 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[104] Stephan Volk, Christian Thöni, and Winfried Ruigrok. 2012. Temporal stability and psychological foundations of cooperation preferences. *Journal of Economic Behavior & Organization* 81, 2 (2012), 664–676. https://doi.org/10.1016/j.jebo.2011.10.006

[105] Guangrong Wang, Jianbiao Li, Wenhua Wang, Xiaofei Niu, and Yue Wang. 2024. Confusion cannot explain cooperative behavior in public goods games. *Proceedings of the National Academy of Sciences* 121, 10 (2024), e2310109121.

[106] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8 (1992), 279–292.

[107] Timo O Weber, Ori Weisel, and Simon Gächter. 2018. Dispositional free riders do not free ride on punishment. *Nature Communications* 9 (2018), 2390. https://doi.org/10.1038/s41467-018-04775-8

[108] Stephan Wendel and Joe Oppenheimer. 2010. An agent-based analysis of context-dependent preferences. *Journal of Economic Psychology* 31, 3 (2010), 269–284.

[109] Shoichiro Yamaguchi, Honda Naoki, Muneki Ikeda, Yuki Tsukada, Shunji Nakano, Ikue Mori, and Shin Ishii. 2018. Identification of animal behavioral strategies by inverse reinforcement learning. *PLoS Computational Biology* 14, 5 (2018), e1006122.

[110] Jennifer Zelmer. 2003. Linear Public Goods Experiments: A Meta-Analysis. *Experimental Economics* 6, 3 (2003), 299–310.

[111] Guozhong Zheng, Jiqiang Zhang, Shengfeng Deng, Weiran Cai, and Li Chen. 2024. Evolution of cooperation in the public goods game with Q-learning. *Chaos, Solitons and Fractals* 188 (2024), 115568. https://doi.org/10.1016/j.chaos.2023.115568

[112] Hao Zhu, Brice De La Crompe, Gabriel Kalweit, Artur Schneider, Maria Kalweit, Ilka Diester, and Joschka Boedecker. 2024. Multi-intention Inverse Q-learning for Interpretable Behavior Representation. *Transactions on Machine Learning Research* (2024). Available online: https://openreview.net/forum?id=hrKHkmLUFk.

A Details on Methods

Bayesian Classification Models. Bayesian models, as proposed by Houser et al. [63], formulate the identification of latent types as a probabilistic inference problem. These models typically assume a parametric likelihood $f(y_{it} \mid \theta_k)$ for observations y_{it} of subject i at time t, where θ_k are the parameters associated with type k. A prior distribution $\pi(\theta_k)$ is specified for the parameters, and the number of types K is either fixed or inferred using a prior $\pi(K)$—which imposes a strong prior on the functional form of the contribution or decision rule for each type.

The posterior probability of the type assignments z_i, where $z_i = k$ indicates that subject i belongs to type k, is given by:

$$p(z, \Theta \mid Y) \propto \prod_{i=1}^{N} \prod_{t=1}^{T} f(y_{it} \mid \theta_{z_i}) \pi(\theta_{z_i}) \pi(z_i) \pi(K),$$

where Y is the observed data and $\Theta = \{\theta_1, \ldots, \theta_K\}$. Bayesian models inherently assume aligned temporal structures (e.g., $t = 1, \ldots, T$) and parametric likelihood forms, which limit their flexibility in handling temporal misalignment.

Finite Mixture Models. Finite mixture models, such as those proposed by Bardsley and Moffatt [11], assume that the data are generated from a mixture of a predefined number of K distributions, for instance reciprocators, strategists, altruists, and free-riders. The likelihood function is expressed as:

$$f(y_{it}) = \sum_{k=1}^{K} \pi_k f(y_{it} \mid \theta_k),$$

where π_k are the mixing proportions ($\sum_{k=1}^{K} \pi_k = 1$) and $f(y_{it} \mid \theta_k)$ is the parametric density associated with type k. The parameters $\{\pi_k, \theta_k\}$ are estimated via maximum likelihood or Bayesian methods. The number of types K can be determined using criteria such as BIC or AIC. Like Bayesian models, finite mixture models typically assume aligned temporal structures and parametric forms for the densities $f(y_{it} \mid \theta_k)$. Thus, they do not recover an underlying latent decision-making mechanism; instead, the structure driving behavior is pre-specified by the modeler.

Classifier-Lasso (C-Lasso). C-Lasso—proposed by Su et al. [98]—identifies latent types in panel data through a penalized regression approach. For a model $y_{it} = x_{it}' \beta_i + \epsilon_{it}$, where β_i varies across individuals but is homogeneous within groups, C-Lasso minimizes the following penalized objective:

$$Q(\beta, \alpha) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}' \beta_i\right)^2 + \lambda \sum_{i=1}^{N} \sum_{k=1}^{K} \|\beta_i - \alpha_k\|,$$

where α_k are group-level parameters, β_i are individual-level parameters, and λ is a tuning parameter. The penalty term clusters β_i into K groups, automatically determining group membership and the number of groups. C-Lasso assumes that the time series are aligned across individuals but does not require parametric likelihoods, making it more flexible than Bayesian or finite mixture models. Importantly, though C-Lasso partitions the data into latent groups based on the observed behavior, it does not recover an underlying latent reward function or decision-making mechanism; instead, it estimates fixed group-specific parameters, with the latent structure being imposed rather than dynamically inferred.

Hierarchical Clustering. [50] proposes to use Hierarchical Clustering with Ward's method to minimize within-cluster variance and using Manhattan distance to measure dissimilarity between contribution strategies.

B Details on the Data

Table S.1. **Data Overview.** Raw data were collected from various published studies on public good behavioral experiments. The inclusion criteria were: conducting a standard public good game, and providing participants with the average contribution of other group members after each round.

Study	Group Size	Periods	Participants
Diederich et al., 2016 [37]	10	7	410
Diederich et al., 2016 [37]	40	7	200
Diederich et al., 2016 [37]	100	7	500
Engel et al., 2014 [41]	4	30	228
Engel et al., 2021 [42]	4	10	236
Engel and Kurschilgen, 2013 [43]	4	30	20
Engel and Kurschilgen, 2019 [44]	4	30	288
Engel and Kurschilgen, 2020 [45]	4	30	432
Engel and Rockenbach, 2014 [46]	4	10	102
Engel and Rockenbach, 2024 [47]	3	20	30
Engel and Rockenbach, 2024 [47]	5	20	252
Kosfeld et al., 2009 [74]	4	20	216
Nikiforakis and Normann, 2008 [88]	4	10	24
Total			**2938**

C Additional Clustering Methods Comparisons

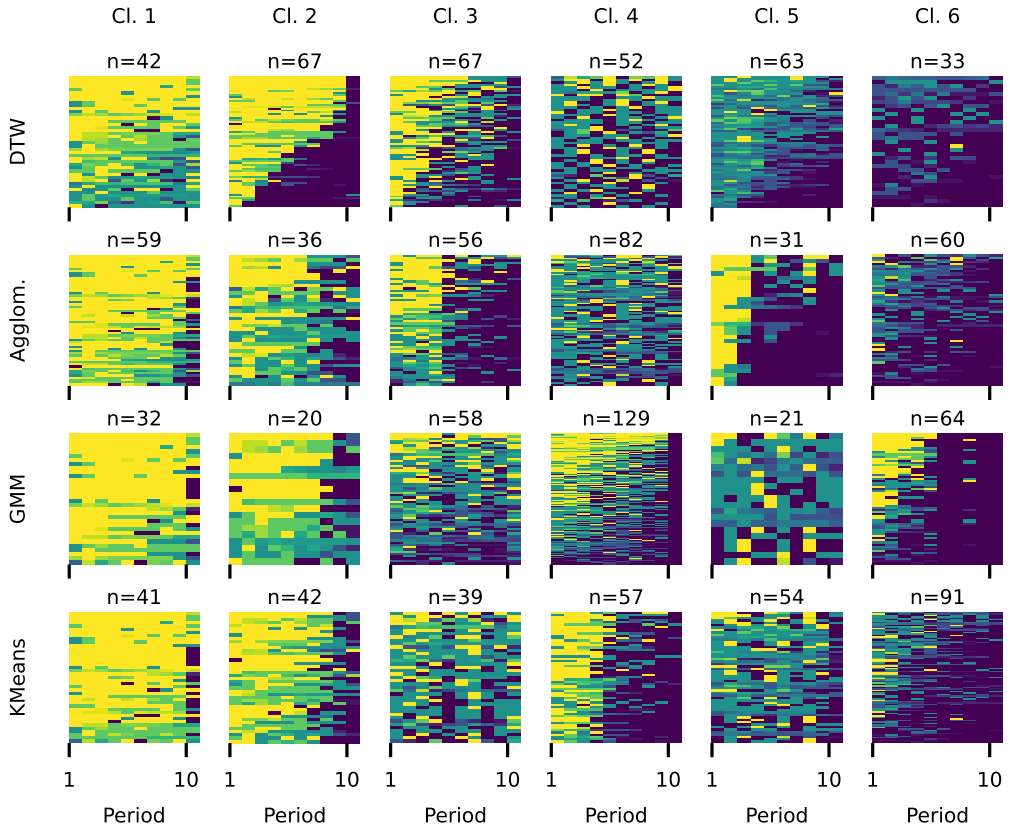

Fig. S.1. **Comparison of Distance Metrics.** DTW-based clustering (top row) identifies temporal patterns, exemplified by the pyramid shape in cluster 2. In contrast, Euclidean-based methods (bottom row) detect point-wise differences, shown by e.g. the binary splits in agglomerative cluster 3 and k-means cluster 4.

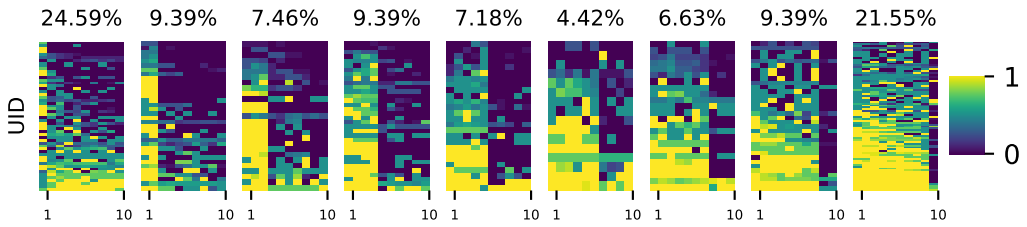

Fig. S.2. **Distribution of optimal threshold points across rounds.** Each subplot represents the percentage of players who showed their most dramatic change in contribution level at that round. There, we plot heatmaps of contributions, subsetted into subplots regarding the round in which the switching process happens. Interestingly, the visualization suggests that a partitioning based on this point of strategy switch is initially appealing, as is the interpretation suggested in the second row of Figure 2 A. However, these interpretations lose explanatory power when viewed in light of the distribution of switching points shown in Figure 2 B.

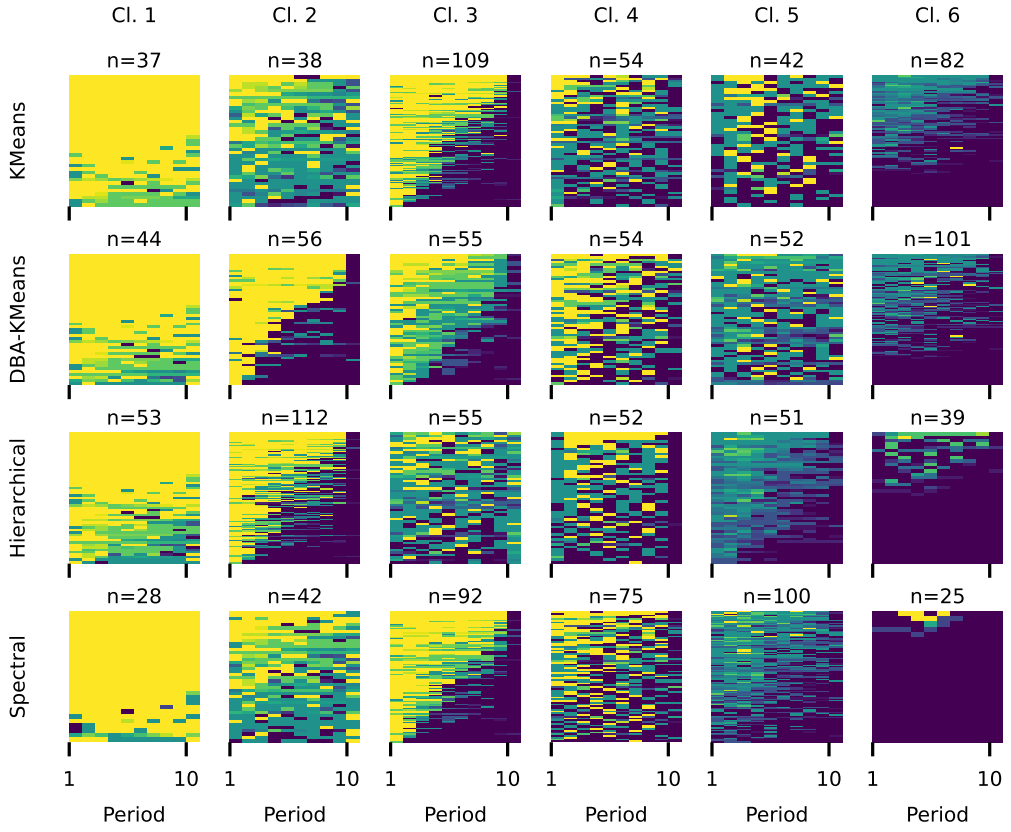

Fig. S.3. **Comparison of clustering algorithms in combination with DTW.** Specifically, we examine combinations of DTW with K-means, hierarchical clustering, and spectral clustering. For this analysis, we have fixed the number of clusters to six, focusing on the data subset with 10 rounds. Our results indicate that spectral clustering most distinctly separates the data.

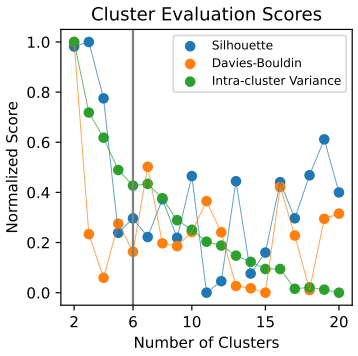

Fig. S.4. **Finding the optimal number of clusters k.** Normalized cluster evaluation scores (Silhouette, Davies-Bouldin, and Intra-cluster Variance) for different numbers of clusters. The vertical grey line marks the selected solution of 6 clusters, which balances evaluation metrics and ensures meaningful cluster sizes (100 members).

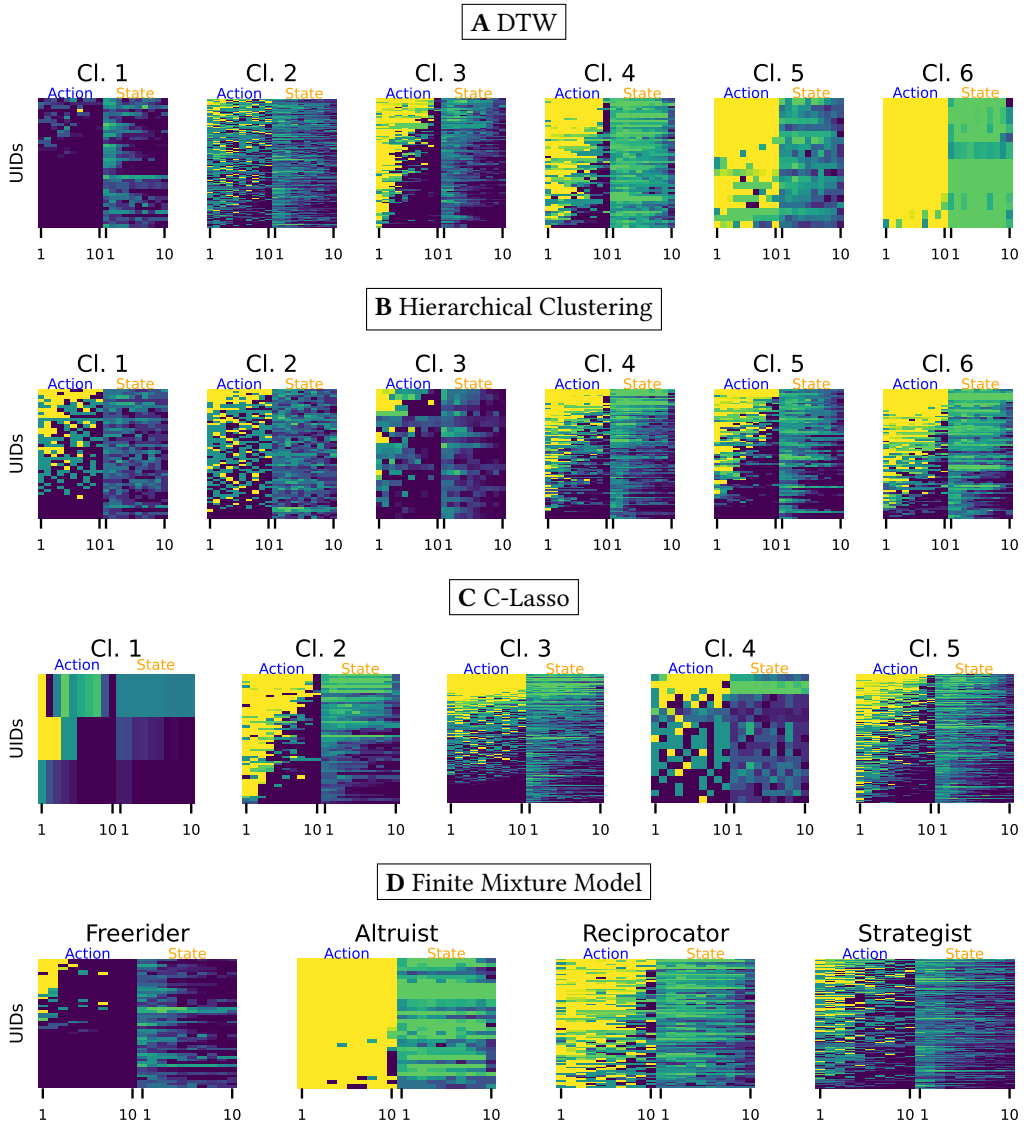

Fig. S.5. **Comparison of partitioning methods.** (A) Dynamic Time Warping (DTW) and Spectral Clustering, (B) Hierarchical Clustering [50], (C) Classifier-Lasso [98], and (D) Finite Mixture Models [11]. We argue that DTW provides the sharpest separation of temporal patterns in our panel data.

D Full Sample Plots

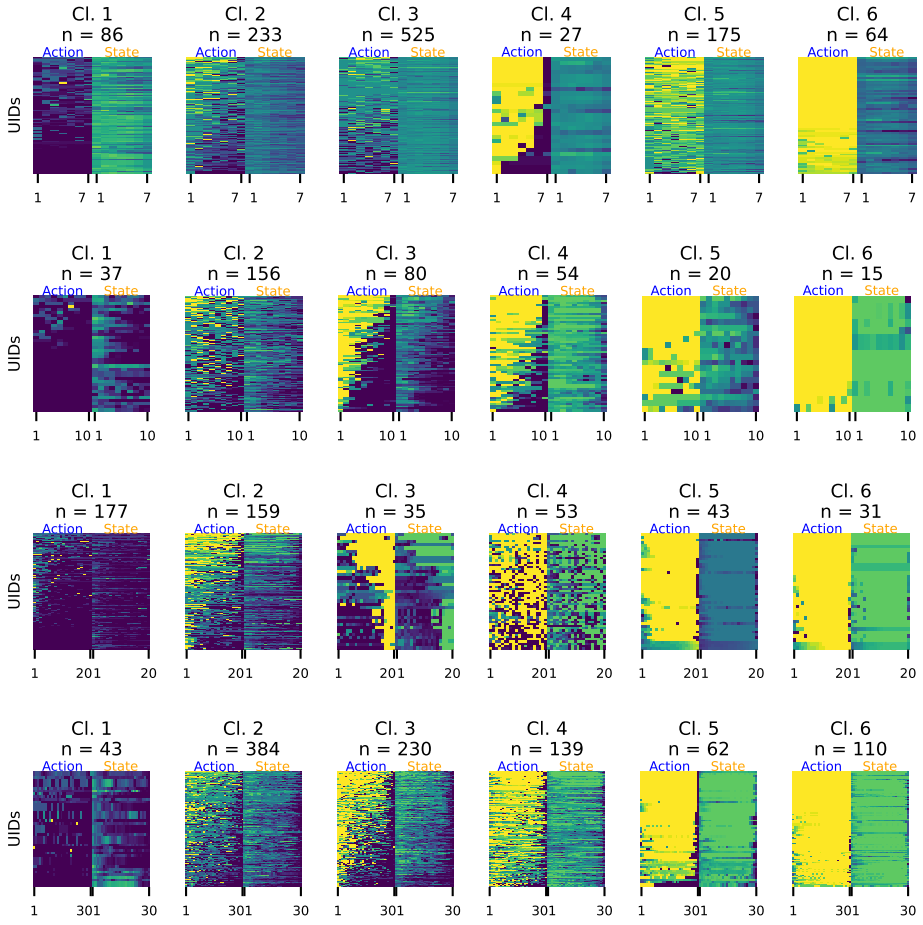

Fig. S.6. **Raw data of actions and states per cluster.** The y-axis represents different UIDs (displayed in the subplot title). The left side of each subplot represents the participants' contributions (actions), while the right side shows the average contribution of their group members (state).

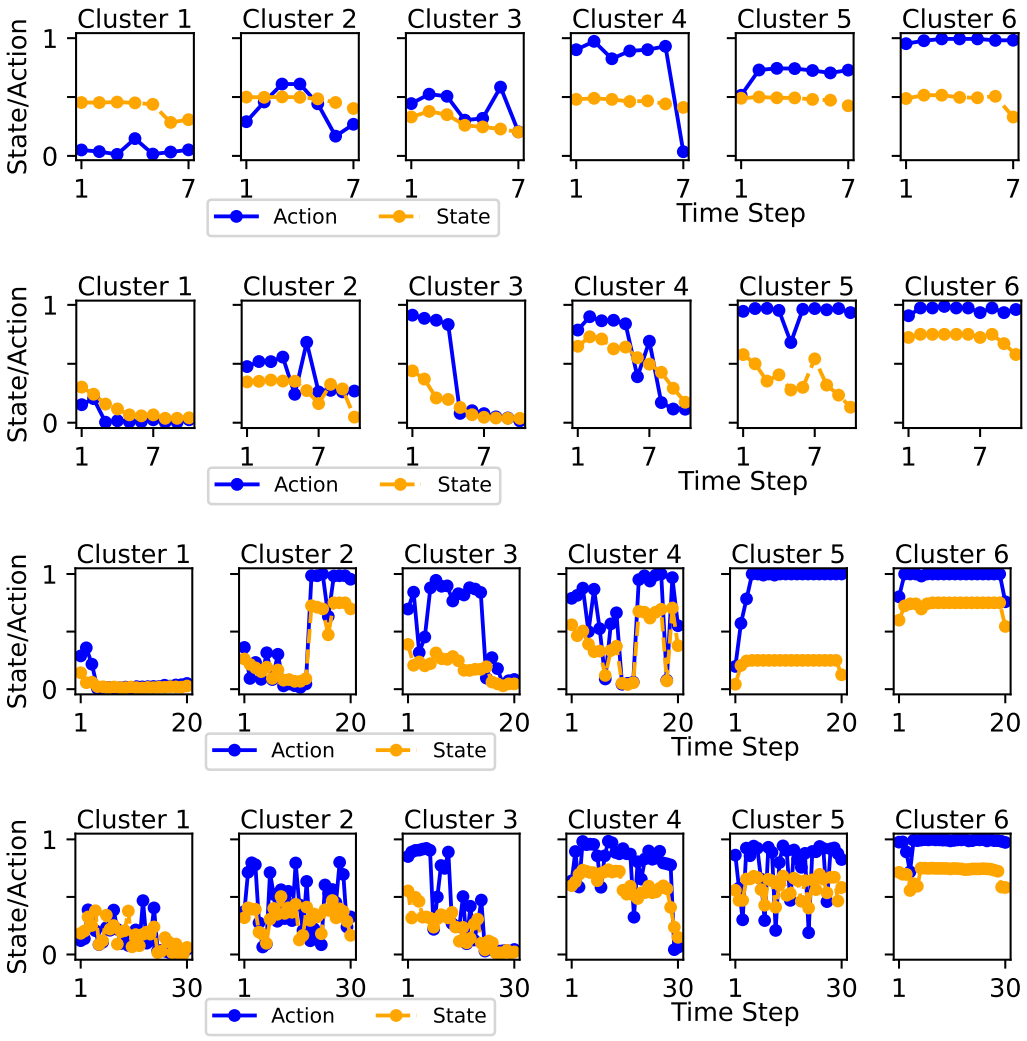

Fig. S.7. Centroids of actions and states per cluster and data subset.

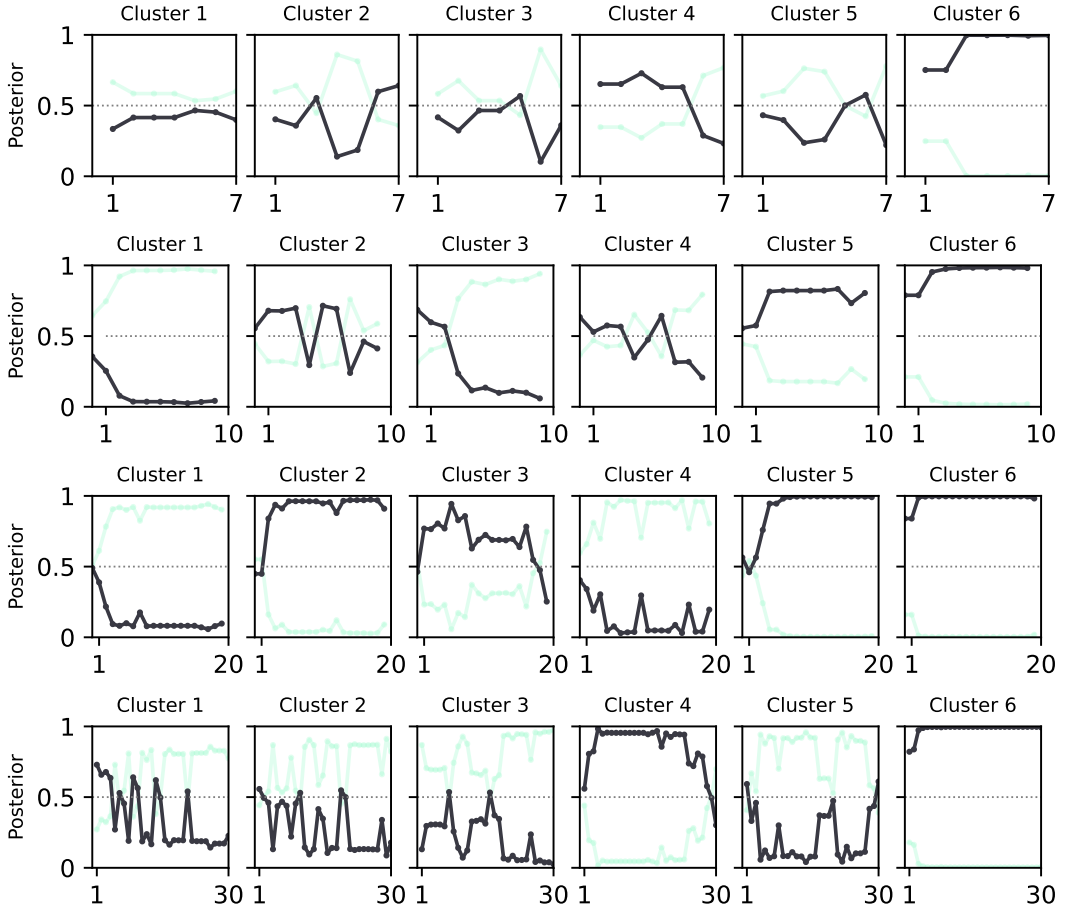

Fig. S.8. **Barycenters of posterior probabilities (intentions) per cluster**. Subplots show participant estimated posterior probability for intention 1 (in black) and 2 (in green), with intention 2 computed as 1 − intention 1).

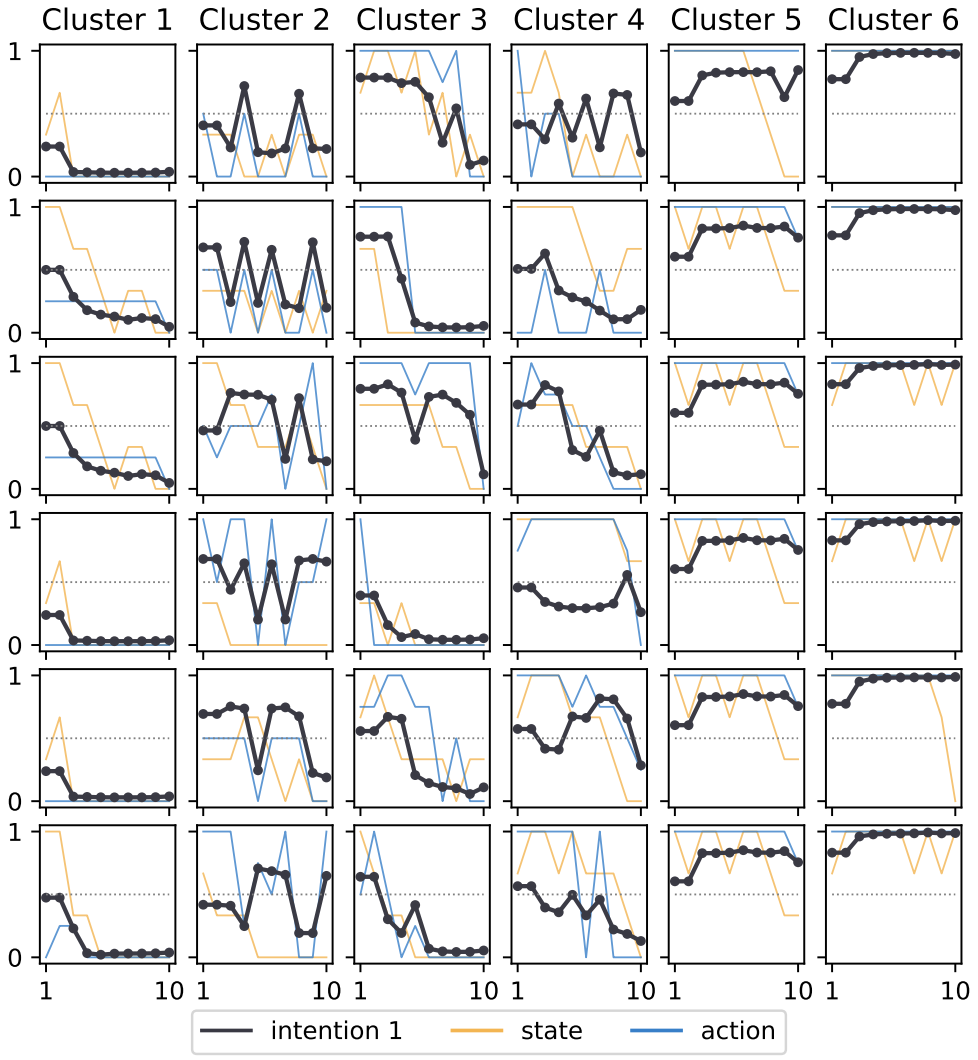

Fig. S.9. **Posterior intentions and contributions for sampled UIDs across clusters.** Subfigures show results for one stratified-sampled UID per cluster, based on average contribution. X-axis represents rounds; Y-axis displays normalized values. Subplots include participant states, actions, and estimated posterior probability for intention 1 (with intention 2 computed as 1 − intention 1).

E Transition Probabilities and Log Likelihoods for All Subsets

Table S.2. Transition probabilities for all subsets.

Max Period	Transition	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
7	Intention 1⟶1	0.718	0.614	0.623	0.738	0.696	0.722
	Intention 1⟶2	0.282	0.386	0.377	0.262	0.304	0.278
10	Intention 1⟶1	0.687	0.525	0.665	0.720	0.724	0.714
	Intention 1⟶2	0.313	0.475	0.335	0.280	0.276	0.286
20	Intention 1⟶1	0.719	0.641	0.742	0.579	0.837	0.740
	Intention 1⟶2	0.281	0.359	0.258	0.421	0.163	0.260
30	Intention 1⟶1	0.667	0.607	0.641	0.713	0.710	0.675
	Intention 1⟶2	0.333	0.393	0.359	0.287	0.290	0.325

Table S.3. Best Log Likelihood across 5×10 Folds by Subset and Cluster

Max Period	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
7	-0.073	-0.641	-0.677	-0.328	-0.506	-0.016
10	-0.100	-0.557	-0.447	-0.637	-0.091	-0.001
20	-0.201	-0.220	-0.603	-0.262	-0.101	-0.025
30	-0.317	-0.754	-0.635	-0.432	-0.461	-0.036