MODELING AND RECOVERING SOCIAL REWARD DIFFERENCES IN AUTISM WITH MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The clinical assessment of Autism Spectrum Disorder (ASD) relies heavily on observing dyadic interaction, yet existing computational models are overwhelmingly single-agent frameworks that fail to account for reciprocal social dynamics. Here, we operationalize the Social Motivation Theory of ASD in a multi-agent reinforcement learning (MARL) environment, providing a testbed to evaluate how alterations in social reward shape movement patterns. By varying a scalar social reward parameter, ρ, across eight conditions from baseline ($\rho = 1.0$) to profound deficit ($\rho = 0.01$), joint policy optimization with Proximal Policy Optimization (PPO) generates behavioral phenotypes consistent with ASD without prescribing top-down behavioral rules. Mean peer distance provides a sensitive metric that scales inversely with ρ (Spearman $r = -0.952$, $p < 0.001$). Control experiments confirm that this effect is specific to social reward attenuation rather than a general reduction in reward magnitude, consistent with the atypical interpersonal distance reported in clinical cohorts. To determine whether these kinematic signatures are informative, we train a bidirectional LSTM on trajectory sequences. The network recovers the latent parameter ρ at $5.06\times$ chance accuracy across nine classes. This recovery remains well above chance after removing explicit relational distance features and improves with extended observation windows. Additionally, mixed-ρ pairings show directionally consistent coordination degradation, offering a preliminary reinforcement-learning account of the Double Empathy Problem. These findings demonstrate that the social reward differences hypothesized in ASD can be modeled in simulation and recovered from dyadic kinematic data.

1 INTRODUCTION

Early and accurate identification of Autism Spectrum Disorder (ASD) remains one of the most consequential unsolved problems in developmental medicine. Current diagnostic practice relies almost exclusively on structured behavioral observation by trained clinicians—a process that is notoriously subjective, resource-intensive, and typically completed no earlier than age four. This timeline is problematic, as it falls well after critical windows for early intervention have begun to close. Despite decades of research, no objective biological or behavioral biomarker has achieved the sensitivity and specificity required for routine screening. The cost of this persistent gap is measured in delayed interventions, missed developmental opportunities, and inequitable access to diagnosis across socioeconomic groups.

Computational psychiatry has made substantial progress in addressing these challenges by modeling individual-level reward processing deficits in various psychiatric conditions using reinforcement learning (RL) frameworks Montague et al. (2012); Huys et al. (2016); Maia & Frank (2011). When applied to ASD, these approaches have successfully illuminated differences in probabilistic reward learning, prediction error signaling, and value-based decision-making Solomon et al. (2011); Kohls et al. (2013). However, existing computational models share a fundamental limitation: they are primarily single-agent frameworks that model an individual's internal reward computation in isolation. Social behavior, by definition, is not a solo computation; it is an emergent property of mutual adaptation between agents who must respond to each other in real time. Although the Social Motivation Theory of ASD Chevallier et al. (2012) proposes that the reduced intrinsic reward value of social

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stimuli constitutes the primary upstream deficit, this theory has never been formally implemented in a learning system capable of producing social behavioral phenotypes endogenously.

In this work, we introduce a multi-agent reinforcement learning (MARL) framework designed to address this gap directly. We train two agents simultaneously using Proximal Policy Optimization (PPO) on a shared policy within a common environment containing three landmarks. A single scalar parameter, ρ—representing the social reward weight, is systematically varied in eight conditions that range from $\rho = 1.0$ (neurotypical) to $\rho = 0.01$ (profound social reward deficit). This approach allows us to implement the core construct of the Social Motivation Theory as a formal reward parameterization. Crucially, no behavioral rules are scripted in our environment. Instead, the social approach and social indifference arise as optimal policies under their respective reward structures through joint policy optimization. To recover these parameters from observation, a bidirectional LSTM classifier is trained on 200-step kinematic trajectory sequences. This frames the clinical phenotyping problem as a task of latent reward parameter identification from synthetic kinematic trajectory sequences.

Our experiments reveal three principal findings. First, mean peer distance serves as a robust computational biomarker of social reward deficit. This metric scales inversely with ρ across the full spectrum with a large effect size (Spearman $r = -0.952$, $p < 0.001$; Kruskal-Wallis $H = 598.98$, $p = 4 \times 10^{-125}$, Cohen's $d = 1.116$) and remains stable across various reward functional forms in ablation experiments. Second, we find that the bidirectional LSTM can recover ρ from trajectory sequences with $56.27\% \pm 0.67\%$ accuracy. This is $5.06\times$ above the chance for a 9-class problem, with peer distance, velocity, and heading change identified as the primary carriers of the latent signal. Third, classification accuracy increases with observation length, rising steeply through the first 100 steps before plateauing. This indicates that social behavioral phenotypes accumulate gradually and become progressively more detectable with sustained observation, rather than appearing at a discrete temporal threshold. Furthermore, mixed-ρ dyads show directionally consistent coordination degradation, providing what we believe is the first reinforcement learning operationalization of the "Double Empathy Problem" Milton (2012).

Our contributions are as follows.

- **A Multi-Agent Foundation for Social Motivation Theory:** We introduce the first MARL framework grounding Social Motivation Theory as a formal reward parameterization. We demonstrate that ASD-consistent behavioral phenotypes arise from a single scalar reward parameter under joint policy optimization, entirely without the use of scripted behavioral rules.
- **Establishment of Robust Computational Biomarkers:** We identify mean peer distance as a robust computational biomarker of the social reward deficit. Our results show that this metric is specific to the social reward parameter and remains stable even when the reward functional forms are altered.
- **Latent Parameter Identification from Kinematics:** We demonstrate that a bidirectional LSTM can recover latent social reward structures from kinematic trajectory sequences at $5.06\times$ above chance accuracy. This frames the task of behavioral phenotyping as a formal problem of latent parameter identification from synthetic kinematic data.
- **Graded Temporal Accumulation of Detectability:** We show that phenotype detectability increases with observation length, accumulating steeply through the first 100 steps before plateauing, suggesting, if this generalizes beyond simulation, that observation duration is a relevant design variable for behavioral screening.

2 Related Work

Computational psychiatry has established reinforcement learning as a principled framework to characterize mental dysfunction as aberrant reward computation Montague et al. (2012); Huys et al. (2016). Applied to ASD, single-agent RL models have illuminated differences in probabilistic reward learning Solomon et al. (2011), prediction error signaling, and mesocorticolimbic responses to social versus monetary rewards Kohls et al. (2013). Maia & Frank (2011) demonstrate the broader utility of RL for mapping computational parameters to psychiatric and neurological conditions. However, these frameworks model one individual's internal reward processing in isolation, and

they cannot represent the dyadic interaction dynamics that are central to ASD's social behavioral presentation. Our work extends computational psychiatry into the multi-agent domain, where social behavioral phenotypes emerge from genuine agent co-adaptation rather than individual reward computation.

Multi-agent reinforcement learning has been used to study emergent social behavior, cooperation, and communication, but not to model neurodevelopmental conditions. Jaques et al. (2019) introduce causal influence as an intrinsic social reward in MARL, demonstrating that social reward terms shape agent coordination, but do not vary reward weights across agents to produce phenotypic variation. Leibo et al. (2017) study prosocial and antisocial emergence in sequential social dilemmas, varying the environmental structure rather than the agent's reward parameters. Most closely related to our approach, Namasse et al. (2026) apply multi-agent simulation to ADHD, and Björne & Balkenius (2005) demonstrate that a single RL parameter sweep produces ASD-like attentional profiles in a single-agent system. Neither grounds their parameterization in Social Motivation Theory, models social reward specifically, nor recovers the latent parameter from behavioral trajectories. Our framework is the first to vary a social reward weight across agents in a MARL environment, grounded in an established neuroscientific theory of ASD, with emergent peer-distance dynamics as the primary behavioral signature.

A growing body of work identifies movement kinematics as a promising source of behavioral biomarkers of ASD. Anzulewicz et al. (2016) achieve high accuracy classifying ASD from tablet gesture trajectories using handcrafted features, and Luongo et al. (2024) extend this using LSTM on raw touch movement sequences for binary ASD versus neurotypical classification. In parallel, Rabinowitz et al. (2018) demonstrate that neural networks can infer latent agent parameters (reward preferences and decision-making type) from gridworld trajectories, establishing trajectory-based latent parameter recovery as a viable methodology. Empirically, interpersonal distance has been identified as a core behavioral correlate of ASD in multiple laboratory paradigms Gessaroli et al. (2013); Asada et al. (2016); Farkas et al. (2024). Our work differs from Luongo et al. (2024) in that we recover a continuous latent reward parameter across nine conditions from synthetic multi-agent trajectories rather than performing binary clinical classification on real patient data. We extend Rabinowitz et al. (2018) to the multi-agent social domain and connect latent parameter recovery to a clinically grounded theory. We introduce a dynamic peer distance, arising from policy optimization rather than measured in a structured laboratory task, as a computational biomarker across the full phenotypic spectrum.

The Double Empathy Problem Milton (2012) reframes autistic-nonautistic social difficulties as a bidirectional disjuncture in reciprocity rather than a unilateral deficit. Although empirical support has grown through behavioral studies Crompton et al. (2020); Sasson et al. (2017), formal computational accounts remain scarce. Recent efforts include a mathematical model of the Double Empathy Problem using dynamical systems (Calderoli et al., 2026) and a Bayesian account of interpersonal misattunement (Bolis et al., 2018). However, neither instantiates the problem as an emergent consequence of mismatched reward structures in agents that learn through environmental interaction. We provide the first reinforcement learning operationalization of this hypothesis, quantifying the degree of socio-cognitive mismatch as $|\rho_0 - \rho_1|$. This frames coordination degradation as a distributed property of the dyad, offering a benchmark for future work in more complex, high-dimensional social environments.

3 METHODS

3.1 ENVIRONMENT

We simulate social interaction using the `simple_spread_v3` environment from the Multi-Particle Environment suite Mordatch & Abbeel (2018), accessed via the `mpe2` package. The environment places two agents and three landmarks in a continuous two-dimensional space. Both agents act in a continuous action space. We set `local_ratio`=0.0, which exposes the other agent's relative position in each agent's observation; the environment's built-in reward is replaced by the reward in Equation 2. Each episode runs for a maximum of 100 steps during training. Evaluation episodes are extended to 200 steps to allow richer behavioral characterization; agents operating beyond their training horizon may exhibit distributional shift, which we acknowledge as a limitation of the temporal analysis (Section 4.3). Both agents train simultaneously in the shared environment, ensuring

that social behavioral signatures arise from genuine agent co-adaptation rather than interaction with a fixed or scripted partner.

Each agent receives a 12-dimensional observation vector at each timestep: self-velocity ($[0:2]$), self-position ($[2:4]$), three landmark relative positions ($[4:10]$), and the other agent's relative position ($[10:12]$). Inter-agent distance is the norm of this relative position, equivalently:

$$d_{\text{peer}} = \|\mathbf{p}_0 - \mathbf{p}_1\|_2 \tag{1}$$

where $\mathbf{p}_0$ and $\mathbf{p}_1$ denote the positions of agent 0 and agent 1.

3.2 Reward Function

Each agent receives an independent reward signal at each timestep, computed from its own observation and its own ρ value:

$$R_i = \frac{\alpha}{d_{\text{toy},i} + 1} + \frac{\rho_i}{d_{\text{peer}} + 1} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{2}$$

where $d_{\text{toy},i}$ is agent i's Euclidean distance to its nearest landmark, d_{peer} is the inter-agent distance defined in Equation 1, $\alpha = 1.0$ is a fixed non-social curiosity weight held constant across all conditions, ρ_i is agent i's social reward weight, and ε_i is an independent Gaussian noise term with $\sigma = 0.05$ modeling sensory uncertainty. Rewards are not shared or summed across agents: each agent optimizes strictly under its own ρ_i value, ensuring that the behavioral phenotype of each agent is a direct consequence of its own reward parameterization and preserving the interpretability of the $\rho \rightarrow$ phenotype mapping under both homogeneous and heterogeneous reward conditions.

The parameter ρ_i directly operationalizes the Social Motivation Theory of ASD Chevallier et al. (2012): when $\rho_i = 1.0$, social proximity earns equal reward to object proximity, corresponding to neurotypical social motivation; as ρ_i decreases toward zero, the social reward term diminishes, and the agent's policy becomes increasingly indifferent to peer position. In the primary phenotype sweep, both agents share the same ρ value across eight conditions: $\{1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.05, 0.01\}$, spanning neurotypical to profound social reward deficit. In the Double Empathy experiment (Section 4.4), agents are assigned different ρ values within the same episode, operationalizing the degree of social motivation mismatch as $|\rho_0 - \rho_1|$.

The non-social weight α is kept constant across all conditions to isolate the effect of ρ. To verify that peer distance is driven by ρ specifically rather than overall reward magnitude, we additionally train a baseline condition with $\alpha = 0.1$ and $\rho = 1.0$, reducing non-social motivation while preserving full social motivation. A second ablation replaces the inverse-distance functional form with a linear reward form to test the robustness of findings across reward parameterizations.

3.3 Training

All agents are trained using Proximal Policy Optimization Schulman et al. (2017), which provides stable policy gradient updates for cooperative multi-agent environments with continuous action spaces. We adopt the parameter-sharing paradigm Yu et al. (2022): both agents share a single MLP policy network, but each agent receives only its own local observation and computes its own independent reward signal. Agents do not share observations, hidden states, or any form of explicit communication channel. This design, sometimes termed *parameter sharing with individual rewards*, is well-established in cooperative MARL Yu et al. (2022) and ensures that coordination emerges purely from shared policy optimization under individual reward signals, precluding any implicit information sharing that would confound the peer-distance results.

Training hyperparameters are: learning rate 3×10^{-4}, rollout buffer size 2048, batch size 64, 10 optimization epochs per rollout, discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.95$, clip range 0.2 and entropy coefficient 0.01. Each of the eight ρ conditions is trained across five independent random seeds, producing 40 training runs in total. Each run proceeds for a maximum of 500,000 timesteps with early stopping applied when the rolling mean episode reward changes by less than

1% across 50 consecutive rollouts. All experiments are implemented using Stable Baselines 3 v2.8.0 Raffin et al. (2021) and PettingZoo v1.25.0 Terry et al. (2021).

All experiments were conducted on an NVIDIA A100 GPU, with each RL training run requiring approximately 10–12 minutes of wall-clock time (total ≈8 hours for the primary sweep) and LSTM training requiring approximately 90 minutes per fold.

3.4 Behavioral Feature Extraction and Statistical Analysis

Following training, each model is evaluated across 100 episodes of 200 steps each. Seven behavioral features are extracted per episode, averaged across both agents: (1) mean peer distance, computed as the episode mean of d_{peer}; (2) Shannon entropy of the two-dimensional spatial occupancy distribution discretized onto a 20×20 grid; (3) mean velocity, the average per-step displacement magnitude; (4) velocity variance; (5) path tortuosity, defined as total path length divided by straight-line displacement between episode start and end positions; (6) spatial dispersion, the mean standard deviation of agent positions; and (7) social convergence slope, the linear regression coefficient of d_{peer} over episode timesteps, with negative values indicating active approach and values near zero indicating social indifference.

Differences in behavioral features across ρ conditions are assessed using the Kruskal-Wallis H-test, a non-parametric alternative to one-way ANOVA that does not assume normality. Significant omnibus tests are followed by Dunn's post-hoc comparisons with Bonferroni correction for multiple comparisons. The effect sizes are reported as Cohen's d relative to the neurotypical baseline ($\rho = 1.0$). These tests treat evaluation episodes as observations; because episodes are nested within five trained policies per condition, p-values should be interpreted as descriptive, and seed-level variability is reflected in the reported standard deviations. Classification accuracies are reported as mean $\pm$ standard deviation across the five cross-validation folds.

3.5 LSTM Classifier

To assess whether the latent social reward structure can be recovered from behavioral trajectories, we train a bidirectional LSTM classifier on 200-step movement sequences, since the signatures of interest (approach toward or drift away from the peer) are defined by how movement unfolds over time, which episode-level summary statistics discard Luongo et al. (2024). Each timestep is represented by a 6-dimensional feature vector: $(x, y, d_{\text{peer}}, d_{\text{landmark}}, v, \Delta\theta)$, where all channels are averaged across the two agents, so that x and y denote the dyad centroid, d_{peer} is inter-agent distance, d_{landmark} is distance to the nearest landmark, v is the instantaneous velocity, and $\Delta\theta$ is heading change between consecutive steps. Feature channels are normalized to zero mean and unit variance using statistics computed on the training fold to prevent data leakage across cross-validation splits. Although the input includes pre-computed relational features (d_{peer}, d_{landmark}) alongside raw positional coordinates, the classifier must learn to integrate these features across the full temporal sequence to recover ρ — a non-trivial inference problem given the high within-class variance documented in Table 1.

The classifier architecture consists of a 2-layer bidirectional LSTM with hidden dimension 64 and dropout 0.3, followed by a multi-layer perceptron (MLP) head with layer normalization and dropout, mapping the latent temporal representations to the nine-class ρ space via $\text{Linear}(128 \to 32) \to \text{ReLU} \to \text{Dropout}(0.3) \to \text{Linear}(32 \to 9)$. The model is trained using Adam optimizer Kingma & Ba (2015) with learning rate 10^{-3} and weight decay 10^{-4}, with a ReduceLROnPlateau scheduler and early stopping with patience of 15 epochs. Classification performance is evaluated using 5-fold stratified cross-validation across 9,600 trajectories that span nine ρ conditions. Each condition contributes 1,000 trajectories (5 seeds $\times$ 200 evaluation episodes) except $\rho = 1.0$, which contributes 1,600 trajectories due to the inclusion of evaluation episodes from the $\alpha = 0.1$, $\rho = 1.0$ specificity baseline condition, which shares the same ρ class label.

To characterize how detectability depends on observation duration, we train a separate classifier for each window length $L \in \{25, 50, 100, 150, 200\}$ on the first L steps of each trajectory (no padding), each evaluated by 5-fold stratified cross-validation with the same architecture and training procedure.

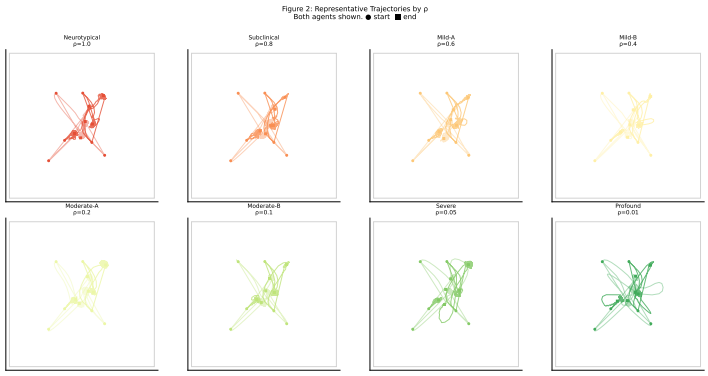

Figure 1: **Representative trajectories by ρ condition (3 episodes per panel).** Both agents shown; circle = start, square = end. High-ρ agents converge toward peers; low-ρ agents exhibit landmark-anchored disorganized exploration. Color: green (neurotypical) to red (profound).

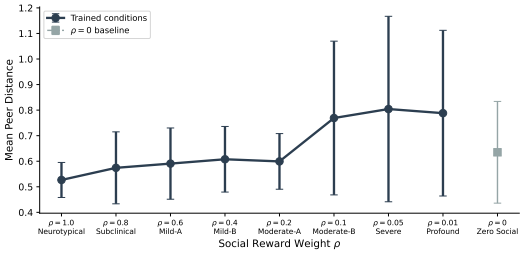

Figure 2: **Mean peer distance scales inversely with social reward weight ρ.** Error bars = std across 5 seeds $\times$ 100 episodes. Grey dashed point = ρ=0.0 baseline (0.635 ± 0.199). Spearman $r = -0.952$, $p < 0.001$; KW $H = 598.98$, $p = 4 \times 10^{-125}$, Cohen's $d = 1.116$.

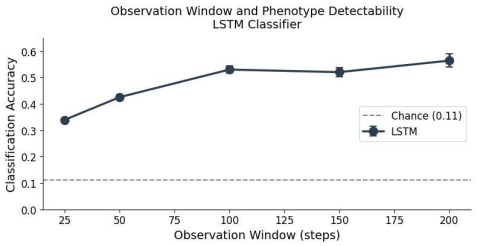

Figure 3: **Classification accuracy versus observation window length.** Accuracy rises steeply through the first 100 steps and then plateaus, indicating that social behavioral phenotypes accumulate gradually rather than at a discrete threshold. Error bars denote standard deviation across five folds; the dashed line marks chance (0.111, nine classes).

4 Results

4.1 Behavioral Phenotypes Under Varying Social Reward

We hypothesized that varying ρ across the social motivation spectrum would produce qualitatively distinct behavioral phenotypes as consequences of policy optimization, without any scripted behavioral rules. All 40 training runs successfully converged across the eight ρ conditions and five random seeds. Final episode rewards decreased monotonically with ρ: from 158.0 ± 2.5 at $\rho = 1.0$ to 92.5 ± 6.1 at $\rho = 0.01$, confirming that the reward function produced the intended motivational gradient throughout the spectrum (Figure S1).

The primary behavioral finding concerns mean peer distance. Agents trained under higher social reward weights maintained significantly closer proximity to their peers throughout evaluation episodes, with mean peer distance scaling inversely with social reward weight throughout the full ρ spectrum (Spearman $r = -0.952$, $p < 0.001$; Kruskal-Wallis $H = 598.98$, $p < 10^{-124}$, Cohen's $d = 1.116$ for $\rho = 1.0$ versus $\rho = 0.05$). Neurotypical agents ($\rho = 1.0$) maintained a mean peer distance of 0.527 ± 0.069, increasing progressively to 0.804 ± 0.363 at $\rho = 0.05$ and 0.788 ± 0.324 at $\rho = 0.01$ (Table 1; Figure 2). The response is not uniform across ρ: Bonferroni-corrected Dunn's tests show that the gradient is driven by a small transition between $\rho = 1.0$ and 0.8 ($p = 1.2 \times 10^{-3}$) and a dominant transition between $\rho = 0.2$ and 0.1 ($p = 1.9 \times 10^{-12}$), while conditions within the 0.8–0.2 band and within the 0.1–0.01 band are not mutually distinguishable. This step-like structure

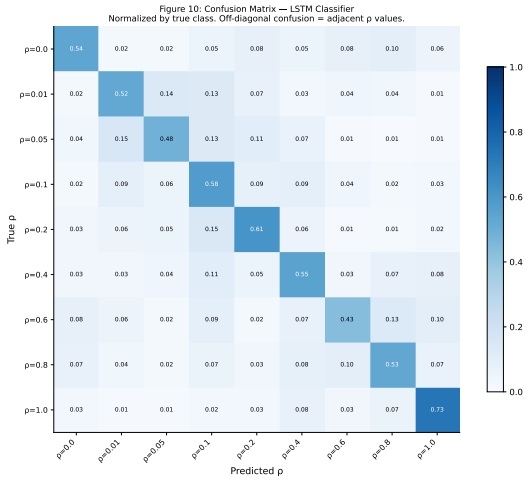

Figure 4: **Normalized confusion matrix (5-fold aggregate).** Rows = true ρ; columns = predicted ρ. Errors concentrate between adjacent values, indicating an ordered behavioral manifold.

arises from the learning dynamics rather than from the reward, which is linear in ρ. Notably, the $\rho = 0$ baseline (0.635 ± 0.199) yields lower peer distance than conditions with $\rho \leq 0.1$, suggesting that a small positive social reward may alter exploration in ways that increase separation relative to no social reward; we leave a mechanistic account of this non-monotonicity to future work. Representative trajectory plots confirm the qualitative distinction: neurotypical agents converge toward their peers within episodes, whereas low-ρ agents exhibit disorganized exploration anchored to landmarks rather than peers (Figure 1; Figure S2).

The social convergence slope, the linear regression coefficient of d_{peer} over episode timesteps, provided a complementary signature: neurotypical agents exhibited the most negative slope (-0.0005 ± 0.0008), indicating active approach behavior, while profound deficit agents showed slopes near zero (-0.0001 ± 0.0008), indicating indifference to peer position. All seven behavioral features differed significantly across ρ conditions (Kruskal-Wallis $p < 0.05$; Tables 1 and S1).

Two ablation experiments confirmed the robustness and specificity of the peer distance finding. First, replacing the inverse-distance reward form with a linear form preserved the peer distance gradient across ρ conditions (from 0.475 at $\rho = 1.0$ to 0.600 at $\rho = 0.01$), confirming that the finding is not an artifact of the specific reward parameterization (Figure S5). Second, a baseline condition with reduced non-social curiosity ($\alpha = 0.1, \rho = 1.0$) produced a mean peer distance of 0.507 ± 0.058, indistinguishable from the standard neurotypical condition (0.527 ± 0.069), confirming that peer distance is driven specifically by ρ rather than by overall reward magnitude (Figure S6).

4.2 Latent Parameter Recovery

We hypothesized that the behavioral signatures produced by different values of ρ would leave recoverable traces in the kinematic trajectory sequences, allowing latent reward parameter identification from kinematic data. A bidirectional LSTM classifier trained on 6-dimensional 200-step trajectory sequences achieved a 5-fold cross-validation accuracy of $56.27\% \pm 0.67\%$ on a 9-class classification problem, representing $5.06\times$ above the chance level of 11.11% (Figure 4). Per-fold accuracy ranged from 55.6% to 57.6%, with low variance across the folds confirming the stability of the result rather than reflecting a single fortunate data split.

Inspection of the confusion matrix revealed that the classification errors were concentrated between adjacent ρ values rather than randomly distributed across classes (Figure 4). Conditions at the extremes of the spectrum, $\rho = 1.0$ and $\rho = 0.01$ were classified with the highest per-class precision, while intermediate conditions produced the greatest confusion with their immediate neighbors. This structured error pattern indicates that the classifier captures an ordered behavioral manifold instead of arbitrary clusters, consistent with the continuous social motivation spectrum proposed by Chevallier et al. (2012). The permutation importance analysis identified peer distance, velocity, and heading

change as the primary carriers of the latent social signal, with peer distance contributing the largest decrease in classification accuracy upon permutation, followed by the velocity and heading change (Figure S4). To verify that latent recovery does not depend on the peer-distance feature that appears directly in the reward function, we retrained the classifier on reduced feature sets under 5-fold cross-validation. Using position and kinematics alone (x, y, velocity, heading change), with both relational features (d_{peer}, d_{landmark}) removed, accuracy is $48.0\% \pm 2.1\%$ ($4.32\times$ chance), retaining 85% of the full-feature performance. Position alone yields $21.7\% \pm 1.2\%$; adding velocity and heading change, neither of which appears in the reward function, accounts for a 26-point gain. Recovery is therefore driven by movement dynamics arising under policy optimization rather than by direct access to the optimized distance term. The main six-feature classifier reported above is unaffected; this analysis complements it as a robustness check.

4.3 Observation Length and Phenotype Detectability

To characterize how phenotype detectability depends on observation duration, we trained a separate classifier for each observation window $L \in \{25, 50, 100, 150, 200\}$ steps, each evaluated by 5-fold stratified cross-validation on genuine (unpadded) trajectories of the corresponding length. This replaces our earlier protocol of applying a single 200-step model to zero-padded inputs, which artificially suppressed accuracy at short windows. Accuracy rises steeply with observation length through the first 100 steps, from $33.9\% \pm 0.4\%$ at $L = 25$ ($3.05\times$ chance) to $53.0\% \pm 1.4\%$ at $L = 100$ ($4.77\times$ chance), and then plateaus ($52.0\% \pm 1.7\%$ at $L = 150$; $56.4\% \pm 2.5\%$ at $L = 200$; Figure 3). This graded accumulation, rather than a discrete threshold, indicates that social behavioral phenotypes become progressively more detectable with sustained observation, with most of the recoverable signal accumulating within the first 100 steps.

If this generalizes beyond simulation, it would suggest that longer structured observation improves reliable detection, while brief behavioral samples carry attenuated but non-trivial signal. The task of mapping simulation steps to real observation time is left to future empirical work. Agents were trained on 100-step episodes; the observation window at which detectability saturates is not resolved within our evaluation horizon, and characterizing the full accuracy-versus-duration curve motivates future work with matched longer-horizon training and evaluation.

4.4 Double Empathy: Coordination Under Mismatched Reward Weights

To provide a preliminary computational test of the Double Empathy Problem Milton (2012), we evaluated coordination metrics across six dyad conditions that span matched and mismatched ρ pairings. The convergence ratio, defined as the normalized reduction in inter-agent distance over the course of an episode, was highest for matched neurotypical dyads ($\rho_0 = \rho_1 = 1.0$: 0.370 ± 0.305) and lowest for a mismatched dyad ($\rho_0 = 1.0, \rho_1 = 0.1$: 0.347 ± 0.285), with matched ASD dyads ($\rho_0 = \rho_1 = 0.1$: 0.369 ± 0.287) coordinating nearly as effectively as matched neurotypical dyads. Matched dyads coordinate better than mismatched dyads, consistent with the Double Empathy prediction; however, the ordering among mismatched pairs is not monotonic in $|\rho_0 - \rho_1|$ and falls within one standard deviation, so we treat it as directional only.

However, the effect sizes are small: the difference between the best-coordinating (matched NT) and worst-coordinating (mismatched) dyad conditions is 0.023 in convergence ratio, well within one standard deviation of either condition. We therefore report these results as directionally consistent with the Double Empathy hypothesis rather than as a confirmation of it, and present this experiment as the first reinforcement learning operationalization of the hypothesis rather than a definitive empirical test. The limited expressiveness of the two-agent, three-landmark environment likely constrains the magnitude of the mismatch effect; richer environments with explicit communication channels and longer interaction horizons are expected to produce larger effects.

5 Discussion

The identification of mean peer distance as a robust computational biomarker of social reward deficit, scaling inversely with ρ (Spearman $r = -0.952$, $p < 0.001$), connects our framework directly to an established empirical literature: laboratory studies report atypical interpersonal distance regulation in autistic individuals Gessaroli et al. (2013); Asada et al. (2016); Farkas et al. (2024). Our work

Table 1: Primary behavioral features by ρ condition (mean $\pm$ std across 5 seeds $\times$ 100 episodes). KW: Kruskal-Wallis H-statistic. d: Cohen's d for $\rho = 1.0$ versus $\rho = 0.05$.

ρ	Label	Peer Distance	Entropy (bits)	Conv. Slope
1.00	Neurotypical	0.527 ± 0.069	2.494 ± 0.836	-0.0005 ± 0.0008
0.80	Subclinical	0.574 ± 0.141	2.149 ± 0.754	-0.0004 ± 0.0007
0.60	Mild-A	0.591 ± 0.139	2.380 ± 1.103	-0.0004 ± 0.0008
0.40	Mild-B	0.608 ± 0.128	2.263 ± 0.931	-0.0004 ± 0.0008
0.20	Moderate-A	0.599 ± 0.109	2.370 ± 1.149	-0.0003 ± 0.0007
0.10	Moderate-B	0.769 ± 0.301	2.006 ± 1.026	-0.0002 ± 0.0008
0.05	Severe	0.804 ± 0.363	1.874 ± 0.822	-0.0002 ± 0.0007
0.01	Profound	0.788 ± 0.324	2.098 ± 1.189	-0.0001 ± 0.0008
KW H		598.98	180.11	99.70
p-value		4×10^{-125}	2×10^{-35}	1×10^{-18}
Cohen's d		1.116	0.747	0.444

demonstrates that this signature arises under policy optimization from reduced social reward weight in a multi-agent framework, without scripted behavioral rules. Its robustness across reward functional forms and specificity to ρ rather than overall reward magnitude provide computational support for the central claim of Social Motivation Theory Chevallier et al. (2012): within this model, social reward deficit is sufficient to produce social behavioral differences.

The temporal finding has potential clinical implications. Classification accuracy increases with observation length, rising steeply through the first 100 steps before plateauing, indicating that social behavioral phenotypes accumulate gradually rather than at a discrete threshold. If this generalizes beyond simulation, longer structured observation should improve reliable detection, while brief behavioral samples may capture attenuated signal.

The Double Empathy experiment produced directionally consistent but small effects (Section 4.4). We attribute this to the minimal environment, which lacks explicit communication, theory-of-mind mechanisms, and extended interaction horizons, and present it as the first reinforcement learning operationalization of the hypothesis Milton (2012) and a benchmark for richer environments.

This work has four principal limitations: the simulation environment is intentionally minimal and does not model joint attention, communication, or temporal reciprocity; ρ is a single scalar that does not capture the full complexity of social motivation; all findings are derived entirely from simulation and require empirical validation against real clinical behavioral data; and the homogeneous parameter-sharing architecture does not reflect the architectural heterogeneity of real social dyads. Consistent with a neurodiversity-informed perspective, the behavioral phenotypes produced by low-ρ agents should be understood as rational adaptations to altered reward parameterization rather than as deficits or malfunctions, since agents are optimal under their own reward structure. These limitations position the present contribution as a computational foundation, grounded in an established clinical theory, that establishes a framework for subsequent empirical validation against real behavioral data.

6 Conclusion

We introduced a multi-agent reinforcement learning framework grounded in the Social Motivation Theory of ASD, demonstrating that a single scalar social reward weight ρ is sufficient to produce a full spectrum of social behavioral phenotypes without scripted behavioral rules. Mean peer distance serves as a robust computational biomarker scaling inversely with social reward weight across eight conditions (Spearman $r = -0.952$, $p < 0.001$; Cohen's $d = 1.116$), and a bidirectional LSTM recovered the latent reward parameter from the kinematic trajectory sequences at $5.06\times$ above chance accuracy on a nine-class problem. Classification accuracy increased with observation length, accumulating steeply through the first 100 steps before plateauing, suggesting that, within this framework, reliable phenotype detection benefits from sustained observation.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing of the manuscript and to help debug code. We have not used generative AI tools for research ideation, experimental design, or the generation of data or results. All experiments were run by the authors, and every reported number was checked against the underlying result files. We have reviewed all AI-assisted work: suggested text edits were reviewed and revised by the authors, and all code changes arising from debugging assistance were tested and verified for correctness. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The environment, observation space, and reward function are specified in Section 3, including the reward ablation and specificity baseline. PPO hyperparameters, the number of random seeds per condition, the early-stopping criterion, software versions, and compute resources are given in Section 3.3. Behavioral features and statistical tests are defined in Section 3.4. The LSTM architecture, input features, normalization procedure, and 5-fold stratified cross-validation protocol, including the per-window and reduced-feature analyses, are described in Section 3.5 and Section 4.2. All classification results are reported as mean $\pm$ standard deviation across five folds, and supplementary figures and tables are provided in the Appendix.

REFERENCES

Anna Anzulewicz, Krzysztof Sobota, and Jonathan T Delafield-Butt. Toward the autism motor signature: Gesture patterns during smart tablet gameplay identify children with autism. *Scientific reports*, 6(1):31107, 2016.

Kosuke Asada, Yoshikuni Tojo, Hiroo Osanai, Atsuko Saito, Toshikazu Hasegawa, and Shinichiro Kumagaya. Reduced personal space in individuals with autism spectrum disorder. *PloS one*, 11 (1):e0146306, 2016.

Petra Björne and Christian Balkenius. A model of attentional impairments in autism: first steps toward a computational theory. *Cognitive Systems Research*, 6(3):193–204, 2005. ISSN 1389-0417. doi: https://doi.org/10.1016/j.cogsys.2004.11.003. URL https://www.sciencedirect.com/science/article/pii/S1389041704000749. Epigenetic Robotics.

Dimitris Bolis, Joshua Balsters, Nicole Wenderoth, Cristina Becchio, and Leonhard Schilbach. Beyond autism: Introducing the dialectical misattunement hypothesis and a bayesian account of intersubjectivity. *Psychopathology*, 50(6):355–372, 2018.

Enrique Calderoli, Maria Cristina Varriale, and Flávio Kapczinski. A distinct communication strategies model of the double empathy problem. *arXiv preprint arXiv:2602.02562*, 2026.

Coralie Chevallier, Gregor Kohls, Vanessa Troiani, Edward S Brodkin, and Robert T Schultz. The social motivation theory of autism. *Trends in cognitive sciences*, 16(4):231–239, 2012.

Catherine J Crompton, Danielle Ropar, Claire VM Evans-Williams, Emma G Flynn, and Sue Fletcher-Watson. Autistic peer-to-peer information transfer is highly effective. *Autism*, 24(7): 1704–1712, 2020.

Kinga Farkas, Orsolya Pesthy, Karolina Janacsek, and Dezső Németh. Interpersonal distance theory of autism and its implication for cognitive assessment, therapy, and daily life. *Perspectives on Psychological Science*, 19(1):126–136, 2024.

Erica Gessaroli, Erica Santelli, Giuseppe di Pellegrino, and Francesca Frassinetti. Personal space regulation in childhood autism spectrum disorders. *PloS one*, 8(9):e74959, 2013.

Quentin JM Huys, Tiago V Maia, and Michael J Frank. Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature neuroscience*, 19(3):404–413, 2016.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pp. 3040–3049. PMLR, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Gregor Kohls, Martin Schulte-Rüther, Barbara Nehrkorn, Kristin Müller, Gereon R Fink, Inge Kamp-Becker, Beate Herpertz-Dahlmann, Robert T Schultz, and Kerstin Konrad. Reward system dysfunction in autism spectrum disorders. *Social cognitive and affective neuroscience*, 8(5): 565–572, 2013.

Joel Z. Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, pp. 464–473, Richland, SC, 2017. International Foundation for Autonomous Agents and Multiagent Systems.

Maria Luongo, Roberta Simeoli, Nicola Milano, and Michela Ponticorvo. Enhancing early autism diagnosis through machine learning: Exploring raw motion data for classification. *PLoS ONE*, 19, 01 2024.

Tiago V Maia and Michael J Frank. From reinforcement learning models to psychiatric and neurological disorders. *Nature neuroscience*, 14(2):154–162, 2011.

Damian E.M. Milton. On the ontological status of autism: the 'double empathy problem'. *Disability & Society*, 27(6):883–887, 2012. doi: 10.1080/09687599.2012.710008. URL https://doi.org/10.1080/09687599.2012.710008.

P. Read Montague, Raymond J. Dolan, Karl J. Friston, and Peter Dayan. Computational psychiatry. *Trends in Cognitive Sciences*, 16(1):72–80, 2012. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2011.11.018. URL https://www.sciencedirect.com/science/article/pii/S1364661311002518. Special Issue: Cognition in Neuropsychiatric Disorders.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Zineb Namasse, Zineb Hidila, Mohamed Tabaa, Mounia Elhaddadi, and Samar Mouchawrab. Multi-agent reinforcement learning model simulation for attention-deficit hyperactivity disorder children. *Applied Sciences*, 16(4), 2026. ISSN 2076-3417. doi: 10.3390/app16042158. URL https://www.mdpi.com/2076-3417/16/4/2158.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4218–4227. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/rabinowitz18a.html.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. URL http://jmlr.org/papers/v22/20-1364.html.

Noah J. Sasson, Daniel J. Faso, Jack Nugent, Sarah Lovell, Daniel P. Kennedy, and Ruth B. Grossman. Neurotypical peers are less willing to interact with those with autism based on thin slice judgments. *Scientific Reports*, 7, 2017. URL https://api.semanticscholar.org/CorpusID:11659119.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Marjorie Solomon, Anne C. Smith, Michael J. Frank, Stanford Ly, and Cameron S. Carter. Probabilistic reinforcement learning in adults with autism spectrum disorders. *Autism Research*, 4 (2):109–120, 2011. doi: https://doi.org/10.1002/aur.177. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/aur.177`.

Jordan Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35:24611–24624, 2022.

A Appendix

This appendix contains supplementary figures and tables referenced in the main paper. Figure S1 shows training curves that confirm convergence for all 40 training runs. Figure S2 shows spatial occupancy heatmaps for all eight ρ conditions. Figure S3 shows full behavioral feature distributions. Figure S5 shows reward function ablation results for Shannon entropy and convergence slope across the reward functional forms; robustness of peer distance gradient is reported numerically in Section 4.1. Figure S6 shows the baseline comparison confirming peer distance is specific to ρ rather than the overall reward magnitude. Table S1 reports secondary behavioral features omitted from the main document for space.

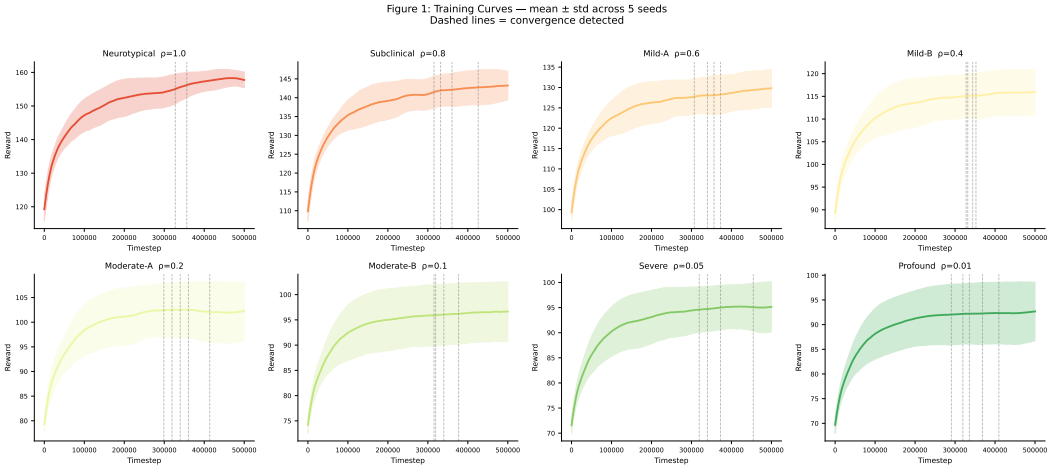

Figure S1: **Training curves for all eight ρ conditions (mean $\pm$ std, 5 seeds).** Episode reward decreases monotonically with ρ, confirming the intended motivational gradient. Dashed vertical lines indicate early stopping at detected convergence. All 40 runs converged successfully within 500,000 timesteps.

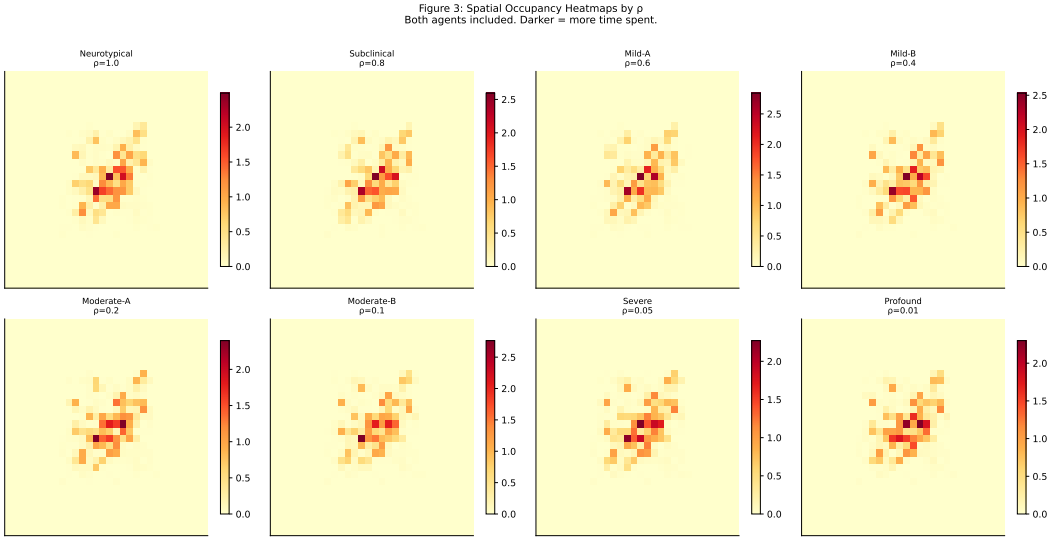

Figure S2: **Spatial occupancy heatmaps by ρ condition (20 evaluation episodes, both agents).** Darker regions indicate greater time spent. Neurotypical agents distribute across the full space during social coordination; low-ρ agents cluster near landmarks, reflecting the shift from peer-anchored to object-anchored spatial behavior.

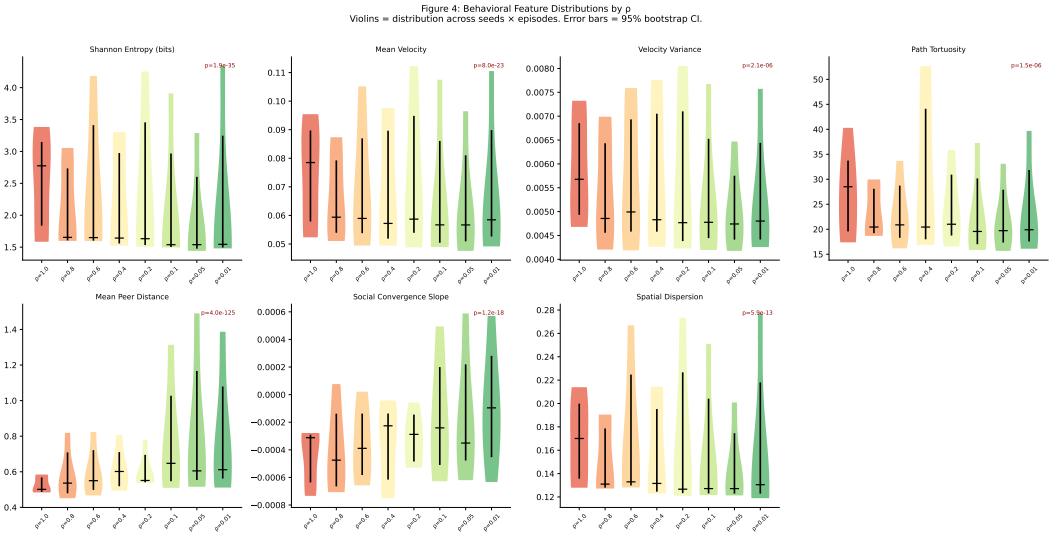

Figure S3: **Behavioral feature distributions by ρ condition.** Violin plots show the full distribution across 5 seeds $\times$ 100 evaluation episodes per condition. Error bars indicate 95% bootstrap confidence intervals. Kruskal-Wallis p-values annotated per panel. All seven features differ significantly across ρ conditions ($p < 0.05$). Color gradient from green (neurotypical) to red (profound deficit).

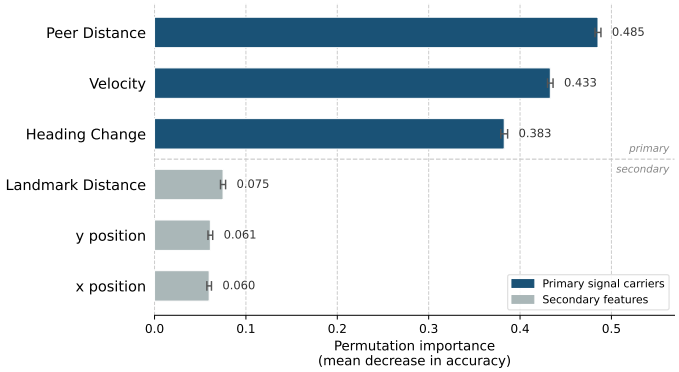

Figure S4: **Permutation importance of LSTM input features.** Mean decrease in accuracy upon permutation; error bars = ± 1 std over 100 repetitions. Peer distance, velocity, and heading change are the primary signal carriers.

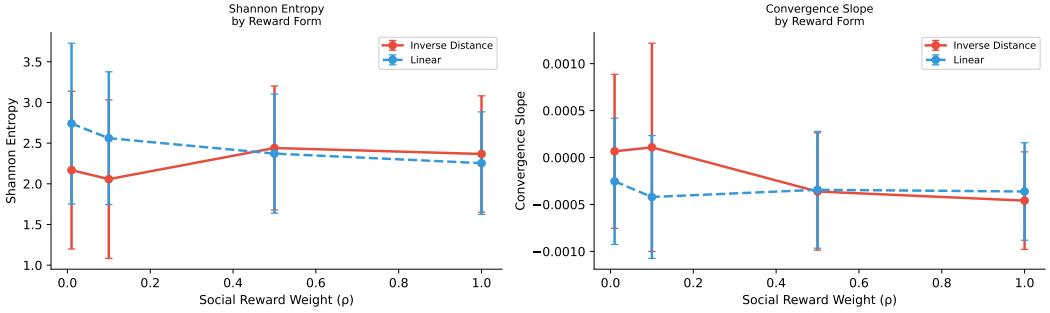

Figure S5: **Reward function ablation across entropy and convergence slope.** Shannon entropy and convergence slope are compared across inverse-distance and linear reward forms. The peer distance gradient (primary biomarker) was confirmed robust across forms via numerical comparison reported in Section 4.1.

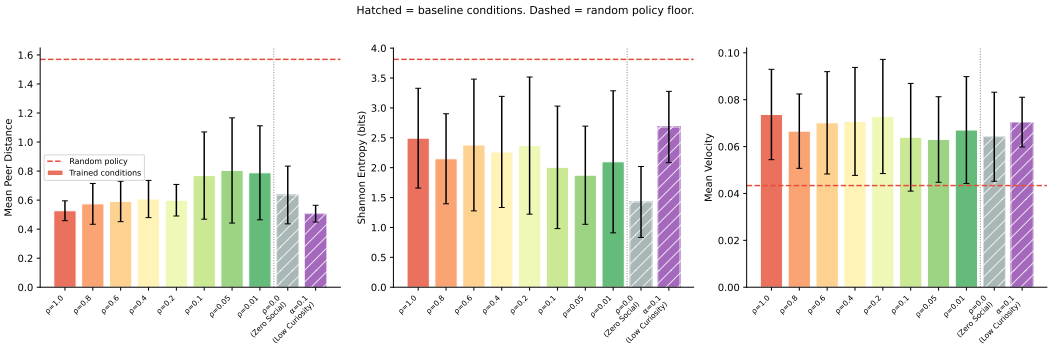

Figure S6: **Baseline comparison.** The α=0.1, ρ=1.0 condition produces peer distance indistinguishable from the neurotypical baseline, confirming peer distance is specific to ρ rather than overall reward magnitude.

Table S1: Secondary behavioral features by ρ condition (mean $\pm$ std across 5 seeds $\times$ 100 episodes).

ρ	Label	Mean Vel	Vel Var	Tortuosity	Dispersion
1.00	Neurotypical	0.074 ± 0.019	0.006 ± 0.003	26.71 ± 26.39	0.168 ± 0.059
0.80	Subclinical	0.067 ± 0.016	0.006 ± 0.003	23.55 ± 23.88	0.153 ± 0.059
0.60	Mild-A	0.070 ± 0.022	0.006 ± 0.003	23.03 ± 18.51	0.170 ± 0.075
0.40	Mild-B	0.071 ± 0.023	0.006 ± 0.003	30.35 ± 100.62	0.160 ± 0.064
0.20	Moderate-A	0.073 ± 0.024	0.006 ± 0.003	23.94 ± 17.68	0.167 ± 0.076
0.10	Moderate-B	0.064 ± 0.023	0.005 ± 0.003	22.19 ± 21.59	0.153 ± 0.074
0.05	Severe	0.063 ± 0.018	0.005 ± 0.003	21.77 ± 16.34	0.143 ± 0.065
0.01	Profound	0.067 ± 0.023	0.005 ± 0.003	23.19 ± 31.58	0.157 ± 0.080
KW H		119.91	38.84	39.62	71.98
p-value		8×10^{-23}	2×10^{-6}	1×10^{-6}	6×10^{-13}
Cohen's d		0.568	0.309	0.225	0.396