Emergent Fast-Slow Dynamics in Multi-Agent Q-Learning for Networked Stochastic Games

Yuxin Geng[1], **Wolfram Barfuss**[3, 4, 5*], **Xingru Chen**[2*]
[1]School of Mathematical Sciences, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]School of Artificial Intelligence, Beihang University, Beijing 100191, China
[3]Transdisciplinary Research Area Sustainable Futures, University of Bonn, Germany
[4]Center for Development Research, University of Bonn, Germany
[5]Institute for Food and Resource Economics, University of Bonn, Germany
yuxin.evol@gmail.com, wbarfuss@uni-bonn.de, xingrucz@gmail.com

Abstract

Understanding the emergence of collective behaviors of multi-agent systems requires investigating the learning dynamics. However, the theoretical analysis of large-scale graph-structured multi-agent reinforcement learning (MARL) systems remains challenging due to agent heterogeneity and the intrinsic coupling between state transitions and individual Q-value updates. In this work, we develop a unified theoretical framework that captures the evolution of agent behaviors at both individual and population levels. By leveraging the pair approximation technique from statistical physics, we derive a closed set of evolution equations that accurately describe the temporal dynamics of the system. Our analysis also reveals a separation of time scales. For small learning rates, state transitions equilibrate rapidly, while Q-value updates evolve slowly with stationary state distributions. Through extensive agent-based simulations, we validate the robustness of our theoretical results and explain the mechanisms that lead to the emergence of cooperation in social dilemmas. Our framework offers new perspectives for bridging complex systems science and MARL, providing insights for the design of cooperative and resilient AI.

Introduction

Reinforcement learning (RL) is a paradigm for sequential decision-making in complex environments, with applications across diverse fields (Sutton and Barto 2018). In sociology and psychology, RL has been used to study how animals and humans acquire and adapt their behaviors (Lee et al. 2004; Dayan and Daw 2008). In computer science and artificial intelligence, RL is a foundational framework for modeling and improving autonomous decision-making (Mnih et al. 2015; Silver et al. 2017). Recently, applying RL to large language models has led to substantial improvements in their capabilities (Bai et al. 2022; Guo et al. 2025). Although single-agent RL has achieved remarkable progress over the years, multi-agent reinforcement learning (MARL) remains a thriving field with many open challenges, including non-convergence, system instability, and catastrophic failures that may cause a complete breakdown of coordination among agents (Hammond et al. 2025; Pan et al. 2025). In addition, individuals can sometimes spontaneously exhibit emergent behaviors, such as cooperation (Leibo et al. 2017) and social learning (Ndousse et al. 2021), that are not explicitly encoded by the underlying rules. Therefore, it is essential to understand MARL systems from the perspective of collective learning dynamics (Barfuss et al. 2025). Such insights not only guide the development of more effective learning algorithms but also inform the design of mechanisms that promote social fairness and cooperation, mitigate social dilemmas, and enhance collective welfare (Hughes et al. 2018; Dafoe et al. 2020; Tacchetti et al. 2025).

One of the central challenges in understanding the dynamics of large-scale MARL systems is managing the complex patterns of interaction among agents. A straightforward modeling approach is to describe the system at the individual level, which involves formulating a set of evolution equations for each agent. Early work investigated the relationship between the dynamics of RL algorithms and the replicator equation in two-player settings. Börgers and Sarin (1997) and Tuyls, Verbeeck, and Lenaerts (2003) demonstrated that, in two-player normal-form games, Cross learning and Q-learning correspond to replicator and replicator-mutator dynamics, respectively. The problem becomes more complex when the environment itself can change dynamically. From the perspective of a single agent, its surrounding environment is non-stationary, as neighboring agents are continuously adapting their behaviors. This results in a feedback coupling between the evolution of agents' policies and the transitions of environmental states. Hennes, Tuyls, and Rauterberg (2009); Hennes, Kaisers, and Tuyls (2010) extended the replicator dynamics framework to accommodate changing environments. More recently, Barfuss, Donges, and Kurths (2019) introduced a batch learning framework to analyze the dynamics of temporal-difference learning in stochastic games.

However, this individual-level approach rapidly becomes intractable as the size of the system increases, due to the well-known "curse of dimensionality." Rather than following each agent's trajectory separately, recent studies have instead adopted a population-level perspective, focusing on modeling the distribution of agent behaviors across the entire system. Hu, Leung, and Leung (2019) and Hu et al. (2022) derived a continuity equation for multi-agent Q-

*Corresponding authors
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

learning in normal-form games. Regarding different interaction structures, Leung, Hu, and Leung (2022) investigated the effects of local interactions and incomplete information; Chu et al. (2022) analyzed the dynamics on regular graphs; and Hussain et al. (2025) examined the convergence of myopic stateless Q-learning in random network systems. For dynamic environments, Chu et al. (2023) studied myopic stateless Q-learning dynamics in stochastic games.

The existing studies above either treat the environment as static (i.e., assuming that agents adopt stateless learning rules) or adopt a deterministic mean-field limit that overlooks the intrinsic stochasticity arising from the exploratory policies, the state transitions, and the heterogeneity of neighbor distributions. To achieve a comprehensive understanding of emergent behaviors in complex and dynamic multi-agent environments, an effective theoretical framework for large-scale MARL systems should capture the above stochasticities, and also link the individual-level and population-level phenomena.

In this work, we consider a graph-structured population of agents engaged in pairwise stochastic games (Shapley 1953), with agents adapting their policies using Q-learning (Watkins and Dayan 1992). We introduce a unified theoretical framework that can scale effectively to large, decentralized settings. To simultaneously represent individual strategies and their game states, we employ a heterogeneous graph to characterize the population. Starting from a microscopic description of the system, we derive a stochastic differential equation (SDE) that governs the evolution of a representative agent's Q-value. This SDE naturally corresponds to a Fokker-Planck equation (FPE), which describes the evolution of the distribution of Q-values on the population level. Meanwhile, the evolution of the distribution of environmental states can be captured by a master equation.

By integrating the master equation that describes state transitions with the FPE that characterizes the evolution of Q-values, and leveraging the pair approximation technique (Hauert and Szabó 2005; Ohtsuki et al. 2006; Chu et al. 2023) from statistical physics, we obtain a complete set of equations for the MARL system. Although Q-value updates and state transitions are coupled, we find that for small learning rates, these processes occur on two distinct time scales and can thus be analyzed separately. In particular, the environmental states quickly reach their stationary distribution, allowing the agents' learning dynamics to be effectively studied within this stable environment. This separation of time scales enables us to derive a reduced FPE that accurately captures the dynamics of the system. We validate the effectiveness of our theoretical framework through extensive agent-based simulations. In addition, our model successfully predicts the emergence of cooperation even in social dilemmas with adverse payoff structures.

In summary, we present a unified theoretical framework that characterizes the dynamics of multi-agent reinforcement learning in stochastic games within large-scale, graph-structured populations. We identify that state transitions and Q-value updates occur on two distinct time scales and accordingly derive a reduced FPE with time-scale separation that accurately captures the dynamics of the system at both microscopic and macroscopic levels. Our theoretical analysis bridges complex system science and multi-agent reinforcement learning, revealing new insights into the emergence of collective behaviors, especially the emergence of cooperation in large-scale graph-structured multi-agent systems.

Model

Although the theoretical framework developed in this paper is applicable to broader scenarios, including multiplayer games and other temporal-difference learning algorithms, we focus our analysis on a graph-structured population interacting in pairwise stochastic games, and Q-learning for clarity. The components of our model are formally defined in the following subsections, and the overall procedure is outlined in Algorithm 1.

Environment: Stochastic Game

In contrast to normal-form games, in which the environmental state remains static, a stochastic game $\mathcal{G} = \langle \mathcal{N}, \mathcal{S}, \mathcal{A}, T, R \rangle$ is characterized by multiple states with dynamic state transitions. Here, $\mathcal{N} = \{1, 2\}$ denotes the set of players, and $\mathcal{S} = \{s_1, \ldots, s_K\}$ denotes the set of states. At each time step, both players simultaneously select an action from the action set $\mathcal{A}(s)$ available in the current state s. For simplicity, we assume that the action set is identical across all states, i.e., $\mathcal{A}(s) \equiv \mathcal{A} = \{a_1, \ldots, a_M\}$ for all s.

When both players are in state s and select the joint action $\boldsymbol{a} = (a, \tilde{a})$, each receives a reward determined by the reward function $R : \mathcal{N} \times \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \mathbb{R}$, which specifies the reward of each player under the joint action $\boldsymbol{a}$ in state s. Since we assume a symmetric reward structure, the rewards can be represented by a payoff matrix $\boldsymbol{U}(s)$:

$$\begin{array}{c} \\ a_1 \\ a_2 \\ \vdots \\ a_M \end{array} \begin{array}{c} \begin{array}{cccc} a_1 & a_2 & \cdots & a_M \end{array} \\ \begin{pmatrix} U(s,a_1,a_1) & U(s,a_1,a_2) & \cdots & U(s,a_1,a_M) \\ U(s,a_2,a_1) & U(s,a_2,a_2) & \cdots & U(s,a_2,a_M) \\ \vdots & \vdots & \ddots & \vdots \\ U(s,a_M,a_1) & U(s,a_M,a_2) & \cdots & U(s,a_M,a_M) \end{pmatrix} \end{array}. \tag{1}$$

Here, $U(s, a_i, a_j)$ denotes the reward for the row player when selecting action a_i against the column player using a_j in state s.

Given the current state s and joint action $\boldsymbol{a}$, the subsequent environmental state $s' \in \mathcal{S}$ is determined by the Markovian transition kernel $T : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. Specifically, $T(s, \boldsymbol{a}, s')$ gives the probability of transitioning from state s to state s', given the joint action $\boldsymbol{a}$.

Population Structure: Heterogeneous Graph

Each agent i maintains a Q-function $Q_t(i, s, a)$ for each state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, representing the value of executing action a in state s at time t. To simultaneously capture the population structure and the environmental state, we represent the system as a heterogeneous graph $G = (V, E, Q_t, s_t)$ of average degree k. Here, $V = \{i | i \in \{1, 2, \ldots, N\}\}$ denotes the set of agents, and $E =$

$\{e_{ij}|i,j \in V, i$ interacts with $j\}$ denotes the pairwise interaction relationships between agents. Each edge $e_{ij} \in E$ represents the stochastic game interaction between agents i and j. Under this graph representation, the Q-value $Q_t(i,s,a)$ becomes the attribute of node i, and the environmental state of the interaction between agents i and j is characterized as an edge attribute $s_t(i,j)$ that evolves over time.

Policy Update Rule: Q-learning

The agent's policy $X_t(i,s,a)$ specifies the probability of selecting each action $a \in \mathcal{A}$ in a given state s, and is defined by the Boltzmann distribution (softmax):

$$X_t(i,s,a) = \frac{\exp\left[\beta Q_t(i,s,a)\right]}{\sum_{a' \in \mathcal{A}} \exp\left[\beta Q_t(i,s,a')\right]}, \tag{2}$$

where $\beta > 0$ is the inverse temperature, controlling the exploration-exploitation trade-off. As $\beta \to \infty$, agent i deterministically selects the action with the highest Q-value; i.e., $X_t(i,s,a) = 1$ if $a = \arg\max_{a'} Q_t(i,s,a')$, and 0 otherwise. Conversely, as $\beta \to 0$, agent i selects actions uniformly at random, i.e., $X_t(i,s,a) = 1/M$ for all $a \in \mathcal{A}$.

Agents improve their policies by updating Q-values using the temporal-difference (TD) learning rule:

$$Q_{t+1}(i,s,a) = Q_t(i,s,a) + \alpha\, \delta_t(i,s,a), \tag{3}$$

where the learning rate $\alpha \in (0,1)$ determines the weight assigned to new information in the update process, and $\delta_t(i,s,a)$ is the TD error, which relies on information obtained through interactions with their neighbors. Specifically, at time step t, i interacts with each of its neighbors j in the current state $s_t(i,j)$ by executing its pre-selected action $a_t(i, s_t(i,j))$, with neighbor j simultaneously executing its own action. As a result of this interaction, agent i receives a reward $r_t(i, s_t(i,j), j)$ and observes the next state $s_{t+1}(i,j)$. The TD error for this interaction and Q-learning is given by:

$$\begin{aligned} &r_t(i, s_t(i,j), j) + \gamma \max_{a'} Q_t(i, s_{t+1}(i,j), a') \\ &\quad - Q_t(i, s_t(i,j), a_t(i, s_t(i,j))), \end{aligned} \tag{4}$$

where γ is the discount factor, quantifying the degree to which future rewards are valued relative to immediate rewards. These TD errors are then aggregated for each state-action pair (s,a) and averaged over the relevant neighbors j to obtain the TD error $\delta_t(i,s,a)$.

Theoretical Derivation

Before proceeding with the derivation, we introduce some notation. Since we aim to describe the evolution of Q-values at both individual and population levels, we characterize agents by their Q-value vectors rather than by fixed identities. For example, $\boldsymbol{X}(\boldsymbol{Q})$ denotes the policy of an individual with Q-value vector $\boldsymbol{Q} = [Q(s_1,a_1), Q(s_1,a_2), \cdots, Q(s_K,a_M)]$, and $\boldsymbol{\delta}(\boldsymbol{Q})$ denotes the TD error for such an individual. Plain symbols refer to specific components of their corresponding vector quantities. For instance, $Q(\boldsymbol{Q},s,a)$ and $X(\boldsymbol{Q},s,a)$ denote the Q-value and policy probability of the agent with Q-value $\boldsymbol{Q}$ in state s and action a. In addition, we abbreviate $Q(\boldsymbol{Q},s,a)$ as $Q(s,a)$ for convenience.

Algorithm 1: Multi-Agent Q-Learning for Pairwise Stochastic Games on a Graph

Input: Graph-structured population of size N, pairwise stochastic game $\mathcal{G}$, and maximum time steps T
for $t = 0, \ldots, T-1$ **do**
  for each agent i **do**
    for each state $s \in \mathcal{S}$ **do**
      Sample action $a_t(i,s) \sim X_t(i,s,\cdot)$
    end for
  end for
  for each connected unordered agent pair (i,j) **do**
    Agents i and j interact in state $s_t(i,j)$, receive rewards $r_t(i, s_t(i,j), j)$ and $r_t(j, s_t(i,j), i)$, and the state transitions to $s_{t+1}(i,j)$
  end for
  for each agent i **do**
    for each state $s \in \mathcal{S}$ where agent i interacted with some neighbor j in state s **do**
      Compute average TD-error $\delta_t(i,s,a)$ and update Q-value
      $Q_{t+1}(i,s,a) \leftarrow Q_t(i,s,a) + \alpha\delta_t(i,s,a)$
    end for
  end for
end for

Individual-Level Dynamics: SDE

To investigate the population-level evolution of Q-value distributions, it is essential to first understand how Q-values evolve at the individual level. The dynamics of a representative agent's Q-values are influenced by multiple sources of randomness, including stochastic state transitions and exploratory policies. To capture both the typical learning trajectory and the random fluctuations that arise during the learning process, we adopt an SDE framework to model the evolution of a representative individual's Q-value vector $\boldsymbol{Q}$. This SDE describes how the Q-value vector $\boldsymbol{Q}$ changes infinitesimally over an infinitesimal time interval $\mathrm{d}t$:

$$\mathrm{d}\boldsymbol{Q} \;=\; \boldsymbol{\mu}(\boldsymbol{Q})\,\mathrm{d}t \;+\; \sqrt{\boldsymbol{\Sigma}(\boldsymbol{Q})}\,\mathrm{d}\boldsymbol{\xi}_t. \tag{5}$$

Here, we omit the subscript t for the vector quantities $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for simplicity. The term $\boldsymbol{\mu}(\boldsymbol{Q})$ is the drift vector, which represents the deterministic component of the Q-value dynamics, indicating the average direction and rate of change. Each entry $\mu_t(\boldsymbol{Q},s,a)$ corresponds to the expected change in the specific Q-value $Q(s,a)$ per unit time. Compared to the ordinary differential equation (ODE) framework, the SDE framework contains an additional term $\sqrt{\boldsymbol{\Sigma}(\boldsymbol{Q})}\,\mathrm{d}\boldsymbol{\xi}_t$, accounting for the stochastic fluctuations in Q-value updates. The matrix $\sqrt{\boldsymbol{\Sigma}(\boldsymbol{Q})} \in \mathbb{R}^{(KM)\times(KM)}$ characterizes the magnitude and correlation structure of these fluctuations across different state-action pairs, where $\sqrt{\cdot}$ denotes the matrix square root. In $\boldsymbol{\Sigma}(\boldsymbol{Q})$, the diagonal entries represent the variances of the one-step change in each $Q(s,a)$, while the off-diagonal entries represent the covariances between changes in different Q-values. The notation $\boldsymbol{\xi}_t$ denotes a standard $(K \times M)$-dimensional Wiener process. In the following, we sketch the derivation of $\boldsymbol{\mu}(\boldsymbol{Q})$,

leaving the full details of $\boldsymbol{\mu}(\boldsymbol{Q})$ and $\boldsymbol{\Sigma}(\boldsymbol{Q})$ to the Supplementary Material.

When a focal agent with Q-value vector $\boldsymbol{Q}$ selects action a, its opponent chooses action $\tilde{a}$ in state s, and the environment transitions to s', the TD error $\delta_t(\boldsymbol{Q}, s, a, \tilde{a}, s')$ for this interaction is given by:

$$U(s, a, \tilde{a}) + \gamma \max_{a'} Q(s', a') - Q(s, a). \tag{6}$$

The average TD error $\bar{\delta}_t(\boldsymbol{Q}, s, a)$ when the focal agent selects action a in state s is computed by averaging over the distribution of the opponent's Q-value vector $\tilde{\boldsymbol{Q}}$, the opponent's selected action $\tilde{a}$, and the next state s'. Formally, this average is given by:

$$\begin{aligned}\bar{\delta}_t(\boldsymbol{Q}, s, a) := \int \sum_{\tilde{a}, s'} X(\tilde{\boldsymbol{Q}}, s, \tilde{a}) \delta(\boldsymbol{Q}, s, a, \tilde{a}, s') \cdot \\ T(s, a, \tilde{a}, s') p_t(\tilde{\boldsymbol{Q}} | \boldsymbol{Q}, s) \, \mathrm{d}\tilde{\boldsymbol{Q}},\end{aligned} \tag{7}$$

where we use the shorthand $\mathrm{d}\tilde{\boldsymbol{Q}}$ for the product of differentials $\mathrm{d}\tilde{Q}(s_1, a_1), \mathrm{d}\tilde{Q}(s_1, a_2), \cdots, \mathrm{d}\tilde{Q}(s_K, a_M)$. The conditional probability distribution of the opponent's Q-value vector $\tilde{\boldsymbol{Q}}$ is defined as:

$$p_t(\tilde{\boldsymbol{Q}} | \boldsymbol{Q}, s) = \frac{p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})}{\int p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}) \, \mathrm{d}\tilde{\boldsymbol{Q}}}, \tag{8}$$

where $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$ is the probability density of the edge that is state s and connected to agents with Q-values $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$

Recall that the Q-learning update is given by Equation (3), with the TD error defined in Equation (4), and that only the Q-value corresponding to the action chosen and executed by the focal agent is updated. Therefore, the drift term corresponds exactly to the expected change in $Q(s, a)$ per unit time:

$$\mu_t(\boldsymbol{Q}, s, a) = \alpha[1 - (1 - p_t(s|\boldsymbol{Q}))^k] X(\boldsymbol{Q}, s, a) \bar{\delta}_t(\boldsymbol{Q}, s, a). \tag{9}$$

where $p_t(s|\boldsymbol{Q})$ is the proportion of stochastic games involving an individual with Q-values $\boldsymbol{Q}$ that are in state s.

Population-Level Dynamics: FPE and Master Equation

The evolution of agents' Q-values is fundamentally intertwined with the evolution of states: updates to Q-values depend on the rewards received in particular states and the subsequent state transitions, while the states themselves are influenced by the agents' policies, which are in turn determined by their Q-values. Although the SDE in Equation (5) captures the microscopic, stochastic evolution of Q-values for an individual agent, it does not account for the dynamics of state transitions. Moreover, our primary interest often lies in understanding the collective behaviors and emergent phenomena that arise within large populations.

To address this, we focus on two key macroscopic quantities: the distribution of Q-values across agents in the population, denoted by $p_t(\boldsymbol{Q})$, and the conditional state distribution $p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$, which describes the distribution of interaction states between agents with Q-values $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$. Equivalently, these two quantities can be jointly represented by the distribution $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$. In deriving the evolutionary dynamics of this joint distribution, we decompose it into three components: the distribution of the focal agent's Q-value, the distribution of the neighbor's Q-value, and the distribution of the interaction state between them.

From SDE to FPE When shifting our perspective from a single agent to the entire population, the SDE that describes an individual agent's trajectory becomes the generator for an evolving ensemble of agents in the high-dimensional Q-value space. Each agent's path is still driven by the SDE with drift $\boldsymbol{\mu}(\boldsymbol{Q})$ and diffusion $\boldsymbol{\Sigma}(\boldsymbol{Q})$, but we are now interested in how the ensemble of agents is distributed in the Q-value space, i.e., how the population-level density $p_t(\boldsymbol{Q})$ evolves over time. The population-level dynamics corresponding to the SDE are exactly the following FPE:

$$\begin{aligned}\frac{\partial p_t(\boldsymbol{Q})}{\partial t} =& - \nabla_{\boldsymbol{Q}} \cdot \left[\boldsymbol{\mu}(\boldsymbol{Q}) \, p_t(\boldsymbol{Q})\right] \\ &+ \frac{1}{2} \nabla_{\boldsymbol{Q}} \cdot \Big\{ \nabla_{\boldsymbol{Q}} \cdot \left[\boldsymbol{\Sigma}(\boldsymbol{Q}) \, p_t(\boldsymbol{Q})\right] \Big\}.\end{aligned} \tag{10}$$

Here, $\nabla_{\boldsymbol{Q}} \cdot (*) = \sum_{s,a} \partial_{Q(s,a)}(*)$ is the divergence operator, and $\nabla_{\boldsymbol{Q}} \cdot \{\nabla_{\boldsymbol{Q}} \cdot (*)\} = \sum_{s,a,s',a'} \partial^2_{Q(s,a)Q(s',a')}(*)$ corresponds to the sum of all elements of the Hessian matrix of the given function. Analogous to its microscopic counterpart, the FPE contains two transport mechanisms. The convective term $-\nabla_{\boldsymbol{Q}} \cdot \left[\boldsymbol{\mu}(\boldsymbol{Q}) \, p_t(\boldsymbol{Q})\right]$ transfers probability mass along the deterministic learning flow dictated by the drift $\boldsymbol{\mu}$. The density increases in regions where the drift points inward and decreases where it points outward, much like how a current concentrates or disperses suspended particles. Complementing this advection, the diffusive contribution $\frac{1}{2} \nabla_{\boldsymbol{Q}} \cdot \{\nabla_{\boldsymbol{Q}} \cdot [\boldsymbol{\Sigma}(\boldsymbol{Q}) \, p_t(\boldsymbol{Q})]\}$ captures the cumulative impact of random fluctuations encoded in the covariance matrix $\boldsymbol{\Sigma}$, which gradually fan out high-density peaks and smooth sharp gradients, broadening the distribution in direct proportion to the local diffusion strength.

At the individual level, the SDE provides a particle-based mechanics of learning for a single agent, with $\boldsymbol{\mu}$ prescribing the expected increment of $\boldsymbol{Q}$ and $\boldsymbol{\Sigma}$ setting its fluctuations. At the population level, the FPE provides a continuum, fluid-like portrait, where the same $\boldsymbol{\mu}$ steers the centroid of the density ensemble and the same $\boldsymbol{\Sigma}$ governs the rate at which it disperses. The SDE and the FPE form a dual description, as an ensemble of a large number of SDE trajectories, and the solutions of the FPE that are initialized with the same density $p_0(\boldsymbol{Q})$ will evolve identically in distribution.

Pair Approximation for Neighbor Distribution Tracking the evolution of the distribution $p_t(\boldsymbol{Q})$ of Q-values requires the conditional distribution $p_t(\tilde{\boldsymbol{Q}}|\boldsymbol{Q}, s)$ of the first-order neighbor's Q-values. In turn, tracking $p_t(\tilde{\boldsymbol{Q}}|\boldsymbol{Q}, s)$ requires the conditional distribution $p_t(\tilde{\tilde{\boldsymbol{Q}}}|\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}, \tilde{s})$ of the second-order neighbor's Q-values. If this cascade of dependencies were fully resolved, we would need to track the Q-values of all agents and all the states governing their interactions, which is exactly the individual-level description and is

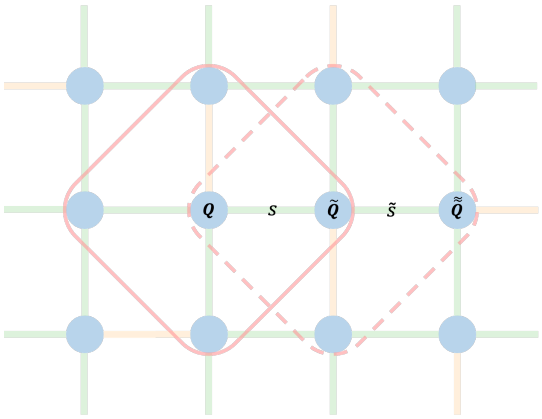

Figure 1: Pair approximation for neighbor distribution. This figure illustrates the pair approximation technique under a lattice population structure. Each blue node represents an agent with a specific Q-value vector, and each edge represents a stochastic game between the connected agents, with the edge color indicating the game state. With node $\boldsymbol{Q}$ as the focal agent, we approximate the distribution $p_t(\tilde{\tilde{\boldsymbol{Q}}}|\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}, \tilde{s})$ of the second-order neighbor $\tilde{\tilde{\boldsymbol{Q}}}$ (dashed box) using the distribution $p_t(\tilde{\tilde{\boldsymbol{Q}}}|\tilde{\boldsymbol{Q}}, s)$ of the first-order neighbor (solid box), as if we treat the neighbor $\tilde{\boldsymbol{Q}}$ as the focal agent.

computationally intractable for a large population size N. To manage this complexity, we employ the pair approximation technique. As illustrated in Figure 1, this approach truncates the infinite series of higher-order correlations by approximating the distribution for a neighbor's neighbor using the distribution for a direct neighbor:

$$p_t(\tilde{\tilde{\boldsymbol{Q}}}|\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}, \tilde{s}) \approx p_t(\tilde{\tilde{\boldsymbol{Q}}}|\tilde{\boldsymbol{Q}}, s). \tag{11}$$

Consequently, the neighbor distribution $p_t(\tilde{\boldsymbol{Q}}|\boldsymbol{Q}, s)$ evolves according to the same dynamic equation (Equation (10)) that governs the marginal distribution $p_t(\boldsymbol{Q})$. As a result, the system's dynamics can be fully described by the joint distribution of the triplet $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$.

Master Equation for State Transitions To obtain the evolutionary dynamics of $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$, it remains to account for the effect of state transitions on the joint distribution of the triplet. When only state transition is considered, the transition rate $\lambda_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ from s to s' between two agents with Q-values $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ is given by the transition kernel averaged over the policies of both players:

$$\lambda_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) = \sum_{a,\tilde{a}} X_t(\boldsymbol{Q}, s, a) X_t(\tilde{\boldsymbol{Q}}, s, \tilde{a}) T(s, a, \tilde{a}, s'). \tag{12}$$

To describe state transitions at the population level, we multiply the transition rate by the density at state s to obtain the probability flux from s to s' per unit time:

$$J_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) = p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}) \lambda_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}). \tag{13}$$

The temporal evolution of the joint probability density $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$ is determined by the net difference between the probability flux into state s from all other states and the flux out of state s to all other states. This conservation law leads to the master equation that describes the dynamics of $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$:

$$\frac{\partial p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})}{\partial t}\Big|_{\alpha=0} = \sum_{s' \neq s} \Big[J_{s's}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) - J_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \Big]. \tag{14}$$

Full Characterization of the System By coupling the FPEs arising from the SDEs for $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$, and the master equation for the state transitions, we finally arrive at the dynamic equation for $p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})$ that accounts for both Q-value updates and state transitions. The resulting equation is as follows:

$$\begin{aligned} &\frac{\partial p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})}{\partial t} \\ &= \underbrace{\sum_{s' \neq s} \Big[J_{s's}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) - J_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \Big]}_{\text{state-transition inflow/outflow}} \\ &\quad - \underbrace{\sum_{\boldsymbol{F} \in \{\boldsymbol{Q}, \tilde{\boldsymbol{Q}}\}} \nabla_{\boldsymbol{F}} \cdot \Big[\boldsymbol{\mu}(\boldsymbol{F})\, p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}) \Big]}_{\text{Q-values drift (deterministic change)}} \\ &\quad + \underbrace{\frac{1}{2} \sum_{\boldsymbol{F} \in \{\boldsymbol{Q}, \tilde{\boldsymbol{Q}}\}} \nabla_{\boldsymbol{F}} \cdot \Big\{ \nabla_{\boldsymbol{F}} \cdot \big[\boldsymbol{\Sigma}(\boldsymbol{F})\, p_t(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}}) \big] \Big\}}_{\text{Q-values diffusion (stochastic fluctuations)}}. \end{aligned} \tag{15}$$

Separation of Time-Scales

In the FPE component of Equation (15), the master equation is of order $\mathcal{O}(1)$ in α, the drift term is of order $\mathcal{O}(\alpha)$, and the diffusion term is of order $\mathcal{O}(\alpha^2)$. This allows us to express the evolution equations for $p_t(\boldsymbol{Q})$ and $p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ as follows:

$$\frac{\partial p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})}{\partial t} = - \sum_{\boldsymbol{F} \in \{\boldsymbol{Q}, \tilde{\boldsymbol{Q}}\}} \nabla_{\boldsymbol{F}} \cdot \Big[\boldsymbol{\mu}(\boldsymbol{F}) p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \Big] + \mathcal{O}(\alpha^2). \tag{16}$$

$$\begin{aligned} \frac{\partial p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})}{\partial t} &= \sum_{s'} \lambda_{s's}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \cdot p_t(s'|\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \\ &\quad - \sum_{s'} \lambda_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \cdot p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) + \mathcal{O}(\alpha). \end{aligned} \tag{17}$$

When the learning rate α is sufficiently small (i.e., $\alpha \ll 1$), state transitions happen significantly faster than Q-value updates. On this fast time scale, the conditional state distribution $p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ rapidly approaches the slow manifold defined by $\frac{\partial p_t(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})}{\partial t} = 0$, which ensures zero net probability flux for any state s. As a result, the slow variables $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ evolve along the slow manifold, as if the system is embedded in a stationary environment. Throughout this work,

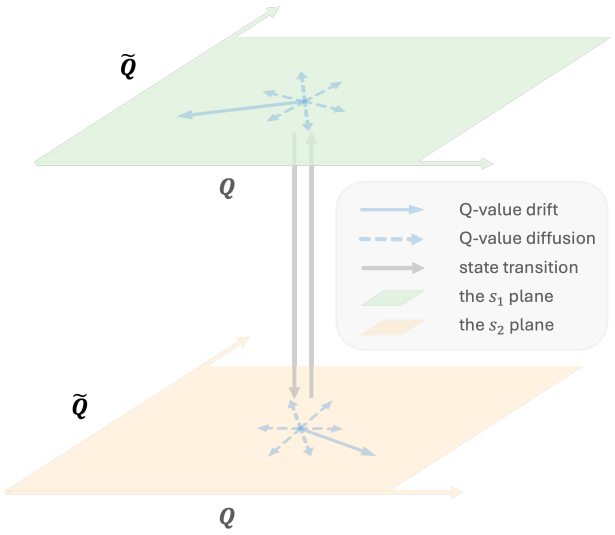

Figure 2: Visualization of Eq. (15) for two-state stochastic games. Each point represents a pair of agents with Q-values $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ in a specific state. For visualization purposes, $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$ are simplified to one-dimensional values. The probability mass $p(\boldsymbol{Q}, s, \tilde{\boldsymbol{Q}})\mathrm{d}\boldsymbol{Q}\mathrm{d}\tilde{\boldsymbol{Q}}$ is transferred between corresponding positions on the two planes through the state-transition flows $J_{ss'}$ and $J_{s's}$ (gray arrows). Meanwhile, the mass evolves on each plane through drift arising from average deterministic learning dynamics (solid blue arrows) and diffusion driven by stochastic fluctuations (dashed blue arrows).

such quantities under stationary state distributions are denoted with an asterisk (e.g., $p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$, $\mu^*(\boldsymbol{Q})$), indicating that all fast state inflows and outflows have reached equilibrium in Eq. (17).

Under this time-scale separation, the stationary state distribution $p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ becomes the function of Q-value pairs $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$. Specifically, the stochastic game between these two agents induces a Markov chain over the state space $\mathcal{S}$, where the transition probability from s to s' is exactly the transition rate $\lambda_{ss'}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ in Equation (12). The stationary distribution $p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ is the unique solution to the following set of equations for all $s \in \mathcal{S}$:

$$p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) = \sum_{s'} p^*(s'|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})\lambda_{s's}(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}). \tag{18}$$

Consequently, the dynamics can be fully described by Q-value pairs with the state distribution reaching its stationary form:

$$\frac{\partial p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})}{\partial t} = -\sum_{\boldsymbol{F} \in \{\boldsymbol{Q}, \tilde{\boldsymbol{Q}}\}} \nabla_{\boldsymbol{F}} \cdot \Big[\boldsymbol{\mu}^*(\boldsymbol{F}) p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}})\Big], \tag{19}$$

where in the calculation of $\boldsymbol{\mu}^*(\boldsymbol{Q})$, we have

$$p^*(\tilde{\boldsymbol{Q}}|\boldsymbol{Q}, s) = \frac{p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \cdot p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})}{\int p_t(\boldsymbol{Q}, \tilde{\boldsymbol{Q}}) \cdot p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})\, \mathrm{d}\tilde{\boldsymbol{Q}}}. \tag{20}$$

Here, $p^*(s|\boldsymbol{Q}, \tilde{\boldsymbol{Q}})$ denotes the stationary distribution of state s between two interacting agents with Q-values $\boldsymbol{Q}$ and $\tilde{\boldsymbol{Q}}$.

The resulting Equation (19) is an FPE without diffusion. In other words, under time-scale separation, although agents randomly select their actions to interact, the learning dynamics of each agent follow a deterministic trajectory.

Experiments

In this section, we conduct agent-based simulations to validate the theoretical predictions. We first consider a two-state Prisoner's Dilemma game (Hilbe et al. 2018) on a regular lattice. In each state, agents can choose between two actions: cooperation (action a_1) or defection (action a_2). Cooperation incurs a cost and provides a benefit to the other agent, so that defection always yields a higher payoff. However, mutual cooperation yields higher social welfare, creating a social dilemma scenario. The payoff matrices for the two states are given by:

$$\boldsymbol{U}(s_1) = \begin{pmatrix} b_1 - c_1 & -c_1 \\ b_1 & 0 \end{pmatrix}, \boldsymbol{U}(s_2) = \begin{pmatrix} b_2 - c_2 & -c_2 \\ b_2 & 0 \end{pmatrix}, \tag{21}$$

where $c_1, c_2 > 0$ represent the costs of cooperation in states s_1 and s_2, respectively, and $b_1, b_2 > 0$ are the corresponding benefits received when the opponent cooperates. The transition structure captures the intuition that mutual cooperation tends to maintain favorable conditions, while defection leads to environmental degradation. The system transitions to state s_1 with probability p_1 and to s_2 with probability $1 - p_1$ if both agents cooperate (choose a_1), and with probability p_2 and $1 - p_2$ if at least one defects (chooses a_2).

Figure 3 presents the theoretical predictions and agent-based simulation trajectories averaged over 10 runs for the two-state Prisoner's Dilemma game. We observe a clear time-scale separation in the system dynamics. As shown in subfigure (a), the state distribution converges to the theoretical stationary distribution (marks) within just one time step, demonstrating the rapid equilibration of the state dynamics relative to the Q-value evolution. Subfigures (b) and (c) display the mean policies of agents in states s_1 and s_2, respectively. Our FPE theoretical model accurately captures the policy evolution of the system, with theoretical predictions closely matching the simulation results. Despite both states presenting social dilemma scenarios, agents converge to full cooperation in both states.

In Figure 4, we explore the mechanism that drives the emergence of cooperation. In subfigures (a) and (b), the payoff matrices are identical for both states ($b_1 = b_2 = 1.2$). In such a case, cooperation does not emerge since there is no environmental incentive to promote it. However, as shown in subfigures (c) and (d), the dynamics change when mutual cooperation leads the system to the prosperous state s_1 ($b_1 = 5$) and defection leads to the degraded state s_2 ($b_2 = 1.2$). Although defection can yield a higher immediate reward, the degradation into s_2 results in lower future payoffs. Agents that sufficiently value future rewards (e.g., with $\gamma = 0.8$) can recognize this trade-off and become more cooperative. This conclusion for large-scale MARL populations is consistent with previous findings in other multi-

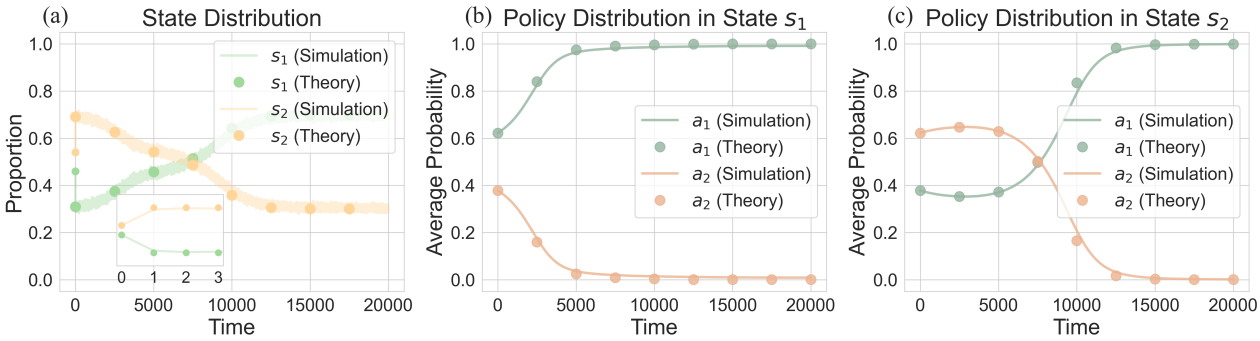

Figure 3: Simulation results and theoretical predictions for the two-state Prisoner's Dilemma game on a regular lattice. We omit variance indicators as all trajectories are nearly identical across simulations. In addition, cooperation behavior emerges as the average probability of cooperation (action a_1) in both states converges to 1, even though both states are social dilemmas. Parameters: $N = 100$, $\alpha = 0.001$, $\beta = 1$, $\gamma = 0.8$, $b_1 = 5$, $b_2 = 1.2$, $c_1 = c_2 = 0.5$, $p_1 = 0.8$ and $p_2 = 0.3$.

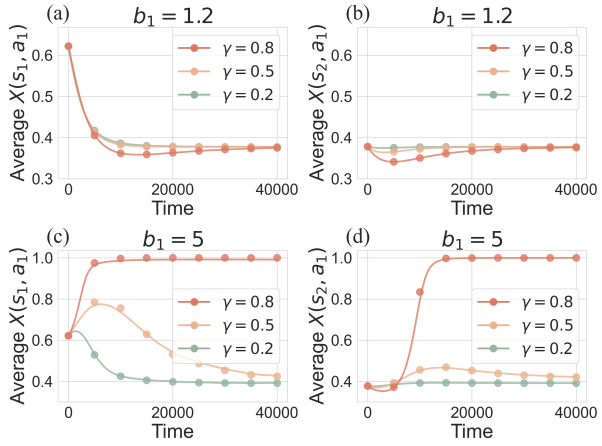

Figure 4: Mechanism of the emergence of cooperation. Except for the parameters shown in the figure, all other parameters are the same as in Figure 3.

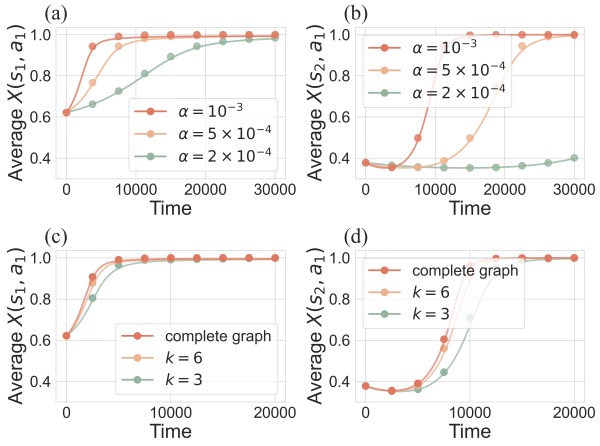

Figure 5: Parameter sensitivity analysis. Our theoretical framework accurately captures the dynamics across different parameter values. Except for the parameters shown in the figure, all other parameters are the same as in Figure 3.

agent systems (Nowak 2006; Hilbe et al. 2018; Su et al. 2019; Barfuss et al. 2020).

In Figure 5, we examine the robustness of our model in the learning rate α and the degree k of the graph. The timescale separation holds effectively with $\alpha < 10^{-3}$, as in subfigures (a) and (b), the theoretical results derived from the stationary distribution closely match the simulation trajectories. In subfigures (c) and (d), we vary the network topology by employing random regular graphs with different degrees. As the degree k increases, individuals converge faster and the learning trajectory gradually approaches that of a complete graph (where $k = N - 1$). When an individual has k neighbors, it effectively performs Q-learning with a batch size of k. Therefore, increasing the network degree is equivalent to increasing the batch size, which accelerates the individual learning process. To further validate the generalizability of our theoretical framework, we include additional cross-validation between theoretical predictions and simulation results in the Supplementary Material.

Conclusion

Beyond accurately predicting agent behaviors, our framework provides the following additional contributions. For networked MARL systems in dynamic environments, our use of a heterogeneous graph representation enables the application of the pair approximation technique to derive closed-form equations describing the joint evolution of Q-values and states. These equations reveal an intrinsic timescale separation between rapid state mixing and slower Q-value updates, allowing us to apply fast-slow systems theory for further analysis. Intuitively, when the learning rate is small, although the environmental states continuously transition, the distribution of states remains stable before the agents' policies undergo a significant change. These theoretical approaches demonstrate that statistical physics and complex systems science can significantly enhance our understanding of MARL systems.

In experiments, we extend the findings on collective behavior to large-scale MARL systems, showing that the emergence of cooperation requires agents sufficiently value future rewards. From the perspectives of dynamical systems and control theory, our framework enables the characterization of equilibrium solutions and their stability, providing an efficient approach to enhancing the controllability of such systems. For example, it informs the design of mechanisms that promote cooperation (McAvoy et al. 2025). In summary, our work expands the methodological toolbox for studying large-scale, networked MARL systems in dynamic environments and offers insights for the development of cooperative and resilient AI.

Acknowledgments

XC acknowledges the support of the Beijing Natural Science Foundation (grant no. 1244045). WB acknowledges the support of the Cooperative AI Foundation.

References

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Barfuss, W.; Donges, J. F.; and Kurths, J. 2019. Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 99(4): 043305.

Barfuss, W.; Donges, J. F.; Vasconcelos, V. V.; Kurths, J.; and Levin, S. A. 2020. Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proceedings of the National Academy of Sciences*, 117(23): 12915–12922.

Barfuss, W.; Flack, J.; Gokhale, C. S.; Hammond, L.; Hilbe, C.; Hughes, E.; Leibo, J. Z.; Lenaerts, T.; Leonard, N.; Levin, S.; Sehwag, U. M.; McAvoy, A.; Meylahn, J. M.; and Santos, F. P. 2025. Collective cooperative intelligence. *Proceedings of the National Academy of Sciences*, 122(25): e2319948121.

Börgers, T.; and Sarin, R. 1997. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1): 1–14.

Chu, C.; Li, Y.; Liu, J.; Hu, S.; Li, X.; and Wang, Z. 2022. A Formal Model for Multiagent Q-Learning Dynamics on Regular Graphs. In *IJCAI*, 194–200.

Chu, C.; Yuan, Z.; Hu, S.; Mu, C.; and Wang, Z. 2023. A pair-approximation method for modelling the dynamics of multi-agent stochastic games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 5565–5572.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*.

Dayan, P.; and Daw, N. D. 2008. Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience*, 8(4): 429–453.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081): 633–638.

Hammond, L.; Chan, A.; Clifton, J.; Hoelscher-Obermaier, J.; Khan, A.; McLean, E.; Smith, C.; Barfuss, W.; Foerster, J.; Gavenčiak, T.; et al. 2025. Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

Hauert, C.; and Szabó, G. 2005. Game theory and physics. *American Journal of Physics*, 73(5): 405–414.

Hennes, D.; Kaisers, M.; and Tuyls, K. 2010. RESQ-learning in stochastic games. In *Adaptive and Learning Agents Workshop at AAMAS*, 8. Citeseer.

Hennes, D.; Tuyls, K.; and Rauterberg, M. 2009. State-coupled replicator dynamics. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, 789–796.

Hilbe, C.; Šimsa, Š.; Chatterjee, K.; and Nowak, M. A. 2018. Evolution of cooperation in stochastic games. *Nature*, 559(7713): 246–249.

Hu, S.; Leung, C.-w.; and Leung, H.-f. 2019. Modelling the dynamics of multiagent q-learning in repeated symmetric games: a mean field theoretic approach. *Advances in Neural Information Processing Systems*, 32.

Hu, S.; Leung, C.-W.; Leung, H.-f.; and Soh, H. 2022. The dynamics of q-learning in population games: A physics-inspired continuity equation model. *arXiv preprint arXiv:2203.01500*.

Hughes, E.; Leibo, J. Z.; Phillips, M.; Tuyls, K.; Dueñez-Guzman, E.; García Castañeda, A.; Dunning, I.; Zhu, T.; McKee, K.; Koster, R.; et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *Advances in Neural Information Processing Systems*, 31.

Hussain, A.; Leonte, D.; Belardinelli, F.; Huser, R.; and Paccagnan, D. 2025. Multi-Agent Q-Learning Dynamics in Random Networks: Convergence due to Exploration and Sparsity. *arXiv preprint arXiv:2503.10186*.

Lee, D.; Conroy, M. L.; McGreevy, B. P.; and Barraclough, D. J. 2004. Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research*, 22(1): 45–58.

Leibo, J. Z.; Zambaldi, V.; Lanctot, M.; Marecki, J.; and Graepel, T. 2017. Multi-agent reinforcement learning in sequential social dilemmas. *arXiv preprint arXiv:1702.03037*.

Leung, C.-w.; Hu, S.; and Leung, H.-f. 2022. Modelling the Dynamics of Multi-Agent Q-learning: The Stochastic Effects of Local Interaction and Incomplete Information. In *IJCAI*, 384–390.

McAvoy, A.; Sehwag, U. M.; Hilbe, C.; Chatterjee, K.; Barfuss, W.; Su, Q.; Leonard, N. E.; and Plotkin, J. B. 2025. Unilateral incentive alignment in two-agent stochastic games. *Proceedings of the National Academy of Sciences*, 122(25): e2319927121.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control

through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Ndousse, K. K.; Eck, D.; Levine, S.; and Jaques, N. 2021. Emergent social learning via multi-agent reinforcement learning. In *International Conference on Machine Learning*, 7991–8004. PMLR.

Nowak, M. A. 2006. Five rules for the evolution of cooperation. *science*, 314(5805): 1560–1563.

Ohtsuki, H.; Hauert, C.; Lieberman, E.; and Nowak, M. A. 2006. A simple rule for the evolution of cooperation on graphs and social networks. *Nature*, 441(7092): 502–505.

Pan, M. Z.; Cemri, M.; Agrawal, L. A.; Yang, S.; Chopra, B.; Tiwari, R.; Keutzer, K.; Parameswaran, A.; Ramchandran, K.; Klein, D.; et al. 2025. Why do multiagent systems fail? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10): 1095–1100.

Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. 2017. Mastering the game of go without human knowledge. *Nature*, 550(7676): 354–359.

Su, Q.; McAvoy, A.; Wang, L.; and Nowak, M. A. 2019. Evolutionary dynamics with game transitions. *Proceedings of the National Academy of Sciences*, 116(51): 25398–25404.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. The MIT Press, second edition.

Tacchetti, A.; Koster, R.; Balaguer, J.; Leqi, L.; Pislar, M.; Botvinick, M. M.; Tuyls, K.; Parkes, D. C.; and Summerfield, C. 2025. Deep mechanism design: Learning social and economic policies for human benefit. *Proceedings of the National Academy of Sciences*, 122(25): e2319949121.

Tuyls, K.; Verbeeck, K.; and Lenaerts, T. 2003. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the second International Conference on Autonomous Agents and Multiagent Systems*, 693–700.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8: 279–292.