An Agent-Centric Dynamical Systems Perspective on Multi-Agent Reinforcement Learning

James Rudd-Jones, María Pérez-Ortiz, Mirco Musolesi

Keywords: Multi-Agent Reinforcement Learning, Dynamical Systems, Agent-Centric Behaviour, Stability Analysis, Sensitivity Analysis

Summary

Analysing learning in Multi-Agent Reinforcement Learning (MARL) environments is challenging, in particular with respect to *individual* decision-making. Practitioners frequently struggle to compare training runs due to the inherent stochasticity in algorithms arising from random dithering exploration, environment transition noise, and stochastic gradient updates to name a few. Traditional analytical approaches, such as replicator dynamics, oft rely on mean-field approximations to remove stochastic effects, but this simplification, whilst able to provide general overall trends, can lead to dissonance between analytical predictions and actual agent realisations. We propose modelling MARL training as a *coupled stochastic dynamical system*, capturing both agent interactions and environmental characteristics. Leveraging tools from dynamical systems theory, we pragmatically analyse the stability and sensitivity of agent behaviour, which are key dimensions for their practical deployments, for example, in presence of strict safety requirements, while preserving the stochastic nature of practical MARL, and in doing so explore methods to enact better control. This framework allows us to rigorously study the inherent stochasticity of MARL, providing a deeper understanding of system behaviour for the design and control of multi-agent learning processes.

Contribution(s)

1. We formalise individual MARL learning updates as coupled discrete time dynamical systems in parameter space, covering both deterministic and stochastic settings.
   Context: Prior dynamical analyses of MARL typically focus on population level approximations such as replicator dynamics (Maynard Smith, 1982; Hofbauer & Sigmund, 1998; Tuyls et al., 2006), which average out stochasticity and algorithmic structure. In contrast, our formulation models the agent-level parameter updates of individual agents and their interactions with the environment as a coupled stochastic dynamical system, enabling analysis of practical MARL algorithms including stochasticity.
2. We introduce a practical methodology for diagnosing stability and sensitivity in deep MARL using tools from dynamical systems theory, even when closed form analysis is impossible, enabling a comprehensive characterisation of the system's dynamical regimes, transitions, and stability properties.
   Context: We use classical dynamical systems tools to provide insight into system stability and attractor structure (Farmer, 1982; Lyapunov, 1992; Eckmann et al., 1995; Grassberger & Procaccia, 1984), enabling practitioners to characterise fixed points, cycles, and chaotic regimes.
3. We empirically demonstrate the approach across multiple MARL games and learning algorithms, analysing regime transitions and hyperparameter sensitivity.
   Context: Our work provides a principled pragmatic empirical approach to identify attractor structures, oscillatory behaviour, and transitions between dynamical regimes under different learning rules and hyperparameters.

An Agent-Centric Dynamical Systems Perspective on Multi-Agent Reinforcement Learning

James Rudd-Jones[1], **María Pérez-Ortiz**[1], **Mirco Musolesi**[1,2]
{james.rudd-jones.22,maria.perez,m.musolesi}@ucl.ac.uk

[1]**Centre for Artificial Intelligence, Department of Computer Science, University College London, London, UK**
[2]**Department of Computer Science and Engineering, University of Bologna, Bologna, Italy**

Abstract

Analysing learning in Multi-Agent Reinforcement Learning (MARL) environments is challenging, in particular with respect to *individual* decision-making. Practitioners frequently struggle to compare training runs due to the inherent stochasticity in algorithms arising from random dithering exploration, environment transition noise, and stochastic gradient updates to name a few. Traditional analytical approaches, such as replicator dynamics, oft rely on mean-field approximations to remove stochastic effects, but this simplification, whilst able to provide general overall trends, can lead to dissonance between analytical predictions and actual agent realisations. We propose modelling MARL training as a *coupled stochastic dynamical system*, capturing both agent interactions and environmental characteristics. Leveraging tools from dynamical systems theory, we pragmatically analyse the stability and sensitivity of agent behaviour, which are key dimensions for their practical deployments, for example, in presence of strict safety requirements. This framework allows us to rigorously study the inherent stochasticity of MARL, providing a deeper understanding of system behaviour.

1 Introduction

Multi-Agent Reinforcement learning (MARL) routinely exhibits abrupt performance shifts, oscillations, and inconsistent game equilibria during training (Mazumdar et al., 2020; Goll et al., 2024). RL practitioners observe these phenomena across divergent training runs even under identical architectures and hyperparameters, making stability during training and asymptotic performance both difficult to reason about and costly to validate empirically (Henderson et al., 2018). Unstable training can lead to brittle policies that fail under small perturbations (Henderson et al., 2018; Agarwal et al., 2021). In safety critical applications these effects are particularly problematic (Garcıa & Fernández, 2015; Amodei et al., 2016). Assessing a learning system's stability and sensitivity to parameters provides crucial insight into robustness, generalisation, and long term adaptability.

There is a long tradition of framing learning as a dynamical system: neuroscience advocated for dynamical systems views of agents and environments (Favela, 2020); evolutionary game theory models population adaptation via dynamical systems (Maynard Smith, 1982); and several RL algorithms have dynamical systems based convergence proofs. In MARL, evolutionary game theory studies population level adaptation via replicator dynamics (Maynard Smith, 1982). However, these studies either (i) focus on *population level* and *deterministic dynamics*, or (ii) treat *environment as a dynamical system* while representing the learning rule only implicitly. By focusing on deterministic population averages, the stochastic path dependent dynamics of individual parameter updates are obscured, motivating an individual level perspective to understand stability, convergence, and emergent behaviour.

We study decision making in MARL using dynamical systems theory tools from an *individual* or *agent-centric* perspective, modelling each agent's learning update as a discrete time dynamical system and treat agents and environment as *coupled dynamical systems*. This formulation captures the actual update rules used in MARL, including stochasticity, and allows us to analyse stability and sensitivity directly at the level of individual decision making. This approach allows us to leverage the mature toolbox of dynamical systems theory, including methods for analysing: stationary (or invariant) distributions (Farmer, 1982; Scargle, 1989), limit cycles, quasi-cycles, random and strange attractors, and Lyapunov stability (Lyapunov, 1992). We provide a pragmatic use of these tools in the context of MARL[1] with an analysis focus on agent emergent interaction dynamics. Concretely, the contributions of this paper can be summarised as follows:

- We formalise individual MARL learning updates as coupled discrete time dynamical systems in parameter space, covering both deterministic and stochastic settings.
- We introduce a practical methodology for diagnosing stability and sensitivity in deep MARL using tools from dynamical systems theory, even when closed form analysis is impossible.
- We empirically demonstrate the approach across multiple MARL games and learning algorithms, analysing regime transitions and hyperparameter sensitivity.

2 Background

Multi-agent reinforcement learning (MARL) has achieved strong empirical results across competitive multiplayer games (Vinyals et al., 2019; Berner et al., 2019), cooperative and competitive social dilemmas (Leibo et al., 2017; Anastassacos et al., 2020; 2021; Dasgupta & Musolesi, 2025), and even environmental policy derivation (Zhang et al., 2025; Rudd-Jones et al., 2025a;b). Despite these successes, MARL remains substantially more challenging to train and analyse than its single-agent counterpart. Non-stationarity induced by concurrent learners often produces oscillatory behaviours, instability, and brittle dependence on hyperparameters, complicating reproducibility (Bloembergen et al., 2015). MARL guarantees convergence only under restrictive assumptions such as of a zero-sum game (Hussain et al., 2023) or in the mean field limit (Yang et al., 2018). As a result, a general theoretical understanding of MARL convergence remains elusive.

Evolutionary Game Theory & Replicator Dynamics. Evolutionary Game Theory provides a population-level perspective on adaptation over time via Replicator Dynamics (Maynard Smith, 1982; Hofbauer & Sigmund, 1998), where the agent centric system is reformulated as an aggregate level dynamical system that denotes the evolving community share of agent strategies. MARL updates and replicator dynamics are strongly linked: the population share of each strategy can be related to the probability of taking a certain action, as if viewing through the lens of a single agent (Bloembergen et al., 2015). For example, Cross learning (Cross, 1973) admits a formal correspondence to replicator dynamics in normal-form games, converging in the continuous time limit (Börgers & Sarin, 1997), with extensions to certain stochastic and sequential games (Hennes et al., 2009; Galstyan, 2013; Hennes et al., 2020). However, there is a disconnect between replicator dynamics theory and practical MARL implementations: classical replicator dynamics assumes infinite populations and omits key algorithmic features such as exploration, bootstrapping, and function approximation that are inherent sources of stochasticity. Consequently, replicator dynamics provides powerful insights but fails to capture finite agent effects and stochastic learning.

Stochastic and Finite-Population Evolutionary Dynamics. To bridge idealised infinite-population models and the noisy reality of multi-agent systems, evolutionary game theory increasingly incorporates finite-population effects and stochasticity. Finite-population evolutionary dynamics capture demographic noise and drift (Nowak & Sigmund, 2004), since finite-agent constraints can disrupt deterministic attractors causing transitions between equilibria that classical replicator dynamics would classify as completely stable. Similarly, evolutionary dynamics with stochastic shocks (Fuden-

[1]For a more in-depth description of dynamical systems theory, we refer the reader to the existing excellent resources in this area, e.g., (Strogatz, 2024).

berg & Harris, 1992; Imhof, 2005) evaluate how environmental volatility perturbs payoffs and dictates long-term equilibrium selection. However, a fundamental limitation of both these approaches is their strict reliance on population-level frequency space, which tracks macro-level proportions of strategies across a generalised pool. This aggregate abstraction fails to map onto individual, trajectory-dependent parameter updates of learning agents, ignoring critical algorithmic mechanisms like bootstrapping, experience replay, or neural function approximation. Furthermore, analytically solving their underlying continuous-time stochastic differential equations becomes mathematically intractable as strategies or states scale, rendering these frameworks unusable for deep MARL.

Other Deterministic Methods. Alternatively, Barfuss et al. (2019) look at deriving the deterministic limit of three common RL algorithms - Q-learning, SARSA, and actor critic learning. By separating the adaption timescale (learning) from the interaction timescale (game dynamics) they map stochastic update rules found in traditional RL algorithms (e.g., ϵ-greedy exploration) to continuous deterministic flows (Barfuss et al., 2019). Thus enabling them to leverage deterministic dynamical systems theory to understand the attractors of the systems. However, adjusting algorithms so that they act in deterministic ways reduces much of their performance as well as the ability to scale to larger state spaces with function approximation. What is gained in theoretical predictive performance comes at the cost of scalable implementations.

Chaos in MARL. One of the major difficulties in MARL is the prevalence of chaotic dynamics, which makes proving convergence particularly challenging. For example, Sato et al. (2002) demonstrate how replicator dynamics can generate complex orbits as well as chaotic attractors and Galla & Farmer (2013) and Sanders et al. (2018) show that two player and many player games respectively can exhibit fixed points, limit cycles, or chaotic behaviour. These findings raise the fundamental question of whether fixed points are consistently attainable in such systems at all, or whether approximations and assumptions used to enforce convergence inadvertently strip away the very dynamics that could be essential for capturing the richness of agent interactions. In this sense, the "chaos" observed in MARL is a fundamental characteristic, posing significant challenges for both theory and practice.

3 Our Approach

In this work we focus on understanding individual decision making of MARL agents, an agent-centric rather than population level viewpoint. Our perspective is therefore complementary to population based approaches: rather than modelling population averages, we directly analyse the learning dynamics of individual agents.

3.1 Definitions

Dynamical Systems. A dynamical system is a mathematical framework for describing the evolution of a system over time, formally consisting of a state space X together with an evolution rule ϕ_t that describes how the state changes. The state space is typically in $\mathbb{R}^n$ or a manifold, and the evolution rule can be either continuous or discrete in time as well as either deterministic or stochastic. Continuous time deterministic dynamics are expressed by an Ordinary Differential Equation (ODE) $\dot{x} = f(x(t)),\ x(0) = x_0$, where $f : X \to \mathbb{R}^n$ is a vector field, and the solution is the flow $\phi_t(x_0)$ that represents a trajectory starting from the initial state. Discrete time deterministic dynamics are described by an iterated map $x_{t+1} = F(x_t),\ x_0 \in X$, where $F : X \to X$ is a transformation and the system evolves by iteration of F.

One can visualise the state space (aka phase space) as a landscape of possible states that the dynamical system evolves through. Trajectories (realisations) of the system emit structure in the phase space, known as a phase plot/portrait. Attractors are locations in phase space the system moves towards regardless of initial conditions, indicating asymptotic behaviour. There are three main types of attractors: fixed points - point locations in which the system converges, limit cycles - attractors the

system periodically loops around, strange attractors - bounded region of phase space where irregular system behaviour leads to a fractal structure.

Coupled dynamical systems are comprised of multiple dynamical systems that have cross/coupling terms within f or F, such as the bidirectionally coupled Logistic Map (Sugihara et al., 2012):

$$\begin{aligned} x_{t+1} &= x_t(r_x - r_x x_t - \beta_{xy} y_t) \\ y_{t+1} &= y_t(r_y - r_y y_t - \beta_{yx} x_t), \end{aligned} \tag{1}$$

where $r_x, r_y, \beta_{xy}, \beta_{yx}$ are parameters that adjust the chaotic behaviour and coupling strength.

Markov Decision Process/Markov Game. RL utilises Markov Decision Processes (MDP), defined by the tuple $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$. Where $\mathcal{S}$ is the set of states, $\mathcal{A}$ the set of actions, $P(s' \mid s, a)$ the transition probability of moving to state s' after taking action a in state s, $R(s, a)$ the reward function, and $\gamma \in [0, 1)$ a discount factor. At each timestep t, the agent observes $s_t \in \mathcal{S}$, chooses $a_t \in \mathcal{A}$, receives reward $r_t = R(s_t, a_t)$, and transitions to $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Generalising for multiple agents, the MDP becomes a stochastic game or Markov Game (MG), a multi-agent MDP with N agents defined by the tuple $\left\langle \mathcal{S}, \{\mathcal{A}^i\}_{i=1}^N, P, \{R^i\}_{i=1}^N, \gamma \right\rangle$. Where $\mathcal{S}$ is the shared state space, $\mathcal{A}^i$ is the action space of agent i, $P(s' \mid s, a^1, \ldots, a^N)$ is the transition probability dependent on the joint action $\mathbf{a} = (a^1, \ldots, a^N)$, and $R^i(s, \mathbf{a})$ is the reward function for agent i. Unlike the single-agent case, where the environment is stationary under a fixed policy, multi-agent systems are inherently non-stationary since the dynamics depend on the evolving policies of all agents.

Single-Agent RL as a Dynamical System. Consider a discrete time dynamical system for a singular agent. We initially assume a fully deterministic framework, and subsequently identify and discuss the sources of stochasticity:

$$s_{t+1} = f(s_t, a_t), \tag{2}$$

where a_t is the agent action, s_t the state at time step t, and $f(\cdot, \cdot)$ our deterministic transition function that is affected by agent actions. Similarly to Beer (1995) we model a coupled dynamical system of the learning RL based agent and the environmental dynamical system:

$$\theta_{h+1} = g(\theta_h, \mathbf{s}_h, \mathbf{a}_h), \tag{3}$$

where h is a learning update step, relating to a certain number of time steps t. $\boldsymbol{\theta}_h$ (a scalar or vector) is the agent state at update h, perhaps the model parameters, and an update depends on the trajectory of states and actions gained in that update window. Actions are generated from a policy π using the environment state s_t and the agent state $a_t = \pi(x_t; \theta_h)$ *We can understand the stability of the function g by analysing what kind of attractor it settles upon.* This is the simplest definition form, but can be extended to incorporate practical implementations such as replay buffers, target networks, or off policy learning, where the samples for updates may originate from different behavioural policies.

MARL as a Dynamical System. For clarity we assume just two agents in this system but the definition can be expanded to general agents. Upon introducing multiple agents our discretised transition dynamics become:

$$s_{t+1} = f(s_t, a_t^1, a_t^2), \tag{4}$$

where the superscript identifies the agent. Creating a coupled dynamical system between the updates of individual agents as their trajectories are dependent on the updates of other agents, defined as:

$$\begin{aligned} \theta_{h+1}^1 &= g^1(\theta_h^1, \mathbf{s}_h, \mathbf{a}_h^1) \\ \theta_{h+1}^2 &= g^2(\theta_h^2, \mathbf{s}_h, \mathbf{a}_h^2), \end{aligned} \tag{5}$$

where an agent updates its internal representation given a set of historical states and its own actions. Different MARL algorithms change g^i (e.g., independent learners or opponent modelling), altering the coupling strength/structure. Additional variables can be added creating further coupling, such as opponent actions or other feature representations used for agent updates. In the following, we

focus on the simplest case where an agent makes decisions about the system only from global state information and its own action, defined in agent parameter space:

$$\theta^1_{h+1} = g^1\left(\theta^1_h, f\left(s_t, \pi_{\theta^1}(s_t), \pi_{\theta^2}(s_t)\right), \pi_{\theta^1}(s_t)\right)$$
$$\theta^2_{h+1} = g^2\left(\theta^2_h, f\left(s_t, \pi_{\theta^1}(s_t), \pi_{\theta^2}(s_t)\right), \pi_{\theta^2}(s_t)\right). \tag{6}$$

Our goal is to *measure and compare* stability and sensitivity of g across MARL algorithms, games, and hyperparameters, beyond only fixed point existence.

The Impacts of Stochasticity. Equation 6 presents the simplest coupled dynamical system for a fully deterministic setting. In practical settings, multiple sources of stochasticity arise within the system. Traditionally, these stochastic elements have been approximated deterministically to enable analytical tractability. In contrast, we explicitly model these sources of randomness to more faithfully capture the behaviour of MARL agents as a coupled stochastic dynamical system:

$$\theta^1_{h+1} = g^1\left(\theta^1_h, f\left(s_t, \pi_{\theta^1}(s_t) + \xi_t, \pi_{\theta^2}(s_t) + \xi_t\right) + \eta_t, \pi_{\theta^1}(s_t) + \xi_t\right) + \zeta_h$$
$$\theta^2_{h+1} = g^2\left(\theta^2_h, f\left(s_t, \pi_{\theta^1}(s_t) + \xi_t, \pi_{\theta^2}(s_t) + \xi_t\right) + \eta_t, \pi_{\theta^2}(s_t) + \xi_t\right) + \zeta_h, \tag{7}$$

where ξ_t represents stochasticity in exploration strategies or policy sampling, η_t the environment transition noise, and ζ_h the stochastic gradient updates. Below, we present a version that subsumes all sources of stochasticity into a single combined stochastic effect term ν_h:

$$\theta^1_{h+1} = g^1\left(\theta^1_h, f\left(s_t, \pi_{\theta^1}(s_t), \pi_{\theta^2}(s_t)\right), \pi_{\theta^1}(s_t)\right) + \nu_h$$
$$\theta^2_{h+1} = g^2\left(\theta^2_h, f\left(s_t, \pi_{\theta^1}(s_t), \pi_{\theta^2}(s_t)\right), \pi_{\theta^2}(s_t)\right) + \nu_h. \tag{8}$$

Comparison to Replicator Dynamics. We compare our agent centric perspective to the population level replicator dynamics formulated as a continuous time dynamical system:

$$\dot{u}_i = u_i\left[w_i(\mathbf{u}) - \bar{w}(\mathbf{u})\right], \tag{9}$$

where $\mathbf{u} = (u_1, u_2, ..., u_n)$ represents the population state vector quantifying the percentage of the population that belong to each of the n strategies, $w_i(\cdot)$ is the fitness function of a specific strategy, and $\bar{w}(\cdot)$ is the average fitness of the total population. As the $\bar{w}(\cdot)$ term is an expectation over the whole population, this averages out many of the stochastic effects.

3.2 Dimensions of our Analysis

Equation 8 provides a general definition of a coupled dynamical system for two agents, which during learning generates a phase portrait in (θ^1, θ^2) for any combination of environment and MARL agents. Leveraging stochastic dynamical systems theory, we analyse system behaviour along three key dimensions:

- *Stability analysis:* Understanding the asymptotic performance of a MARL system allows us to rigorously validate the resulting game equilibria.
- *Sensitivity analysis:* Once the phase-space structure can be analysed, we can assess how parameter changes affect stability, providing insight into the system's sensitivity.
- *Control:* By quantifying stability and sensitivity in the coupled MARL dynamical system, we can inform strategies for improved control.

Our framework is agent-centric and data driven: rather than assuming deterministic structure, we analyse realised stochastic trajectories under arbitrary algorithm environment pairings. In this setting, classical notions (e.g. fixed points, attractors, stability) must be interpreted in a stochastic sense, as process noise perturbs trajectories and precludes strict invariance. Formally, each agent's update follows a Markov process $\theta_{h+1} = g(\theta_h, s_t, \boldsymbol{\pi}) + \nu_h$, and behaviour is characterised through a stationary distribution ρ^θ that satisfies $\theta_h \sim \rho^\theta \Rightarrow \theta_{h+1} \sim \rho^\theta$.

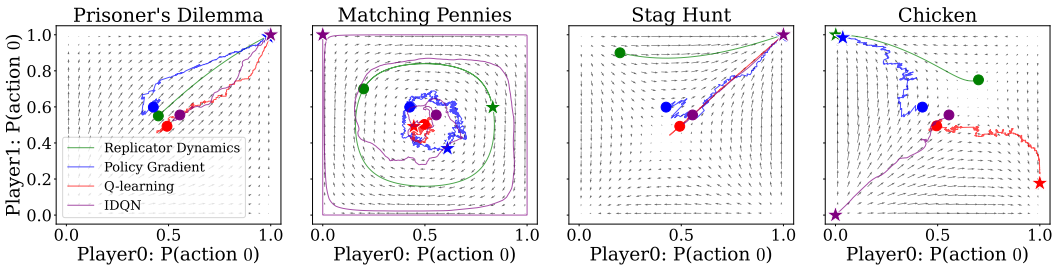

Figure 1: Comparison between replicator dynamics and Policy Gradient, Tabular Q-learning, and IDQN with Boltzmann exploration. Replicator dynamics are represented by the analytical vector field and an arbitrary initial condition realisation. Realisations start from circles and end at stars.

Although closed-form solutions for ρ^θ are generally intractable, ergodic approximations can be obtained via long-run simulations or multiple initialisations. The geometry of ρ^θ reveals qualitative system behaviour. Concentrated mass indicates a fixed point, while sustained oscillations appear as a "smeared" ring, characteristic of quasi-cycles induced by stochasticity. Quasi-cycles arise when deterministic dynamics would converge to a fixed point or limit cycle, but stochastic perturbations sustain oscillatory behaviour. To make this concrete, we introduce four diagnostics:

- *Stationary (invariant) distributions (Farmer, 1982; Scargle, 1989):* Probability distribution over θ. If the Frobenius norm $\|\Sigma\|_F$ of the covariance of ρ^θ is low it indicates fixed point convergence.
- *Lyapunov exponents (Lyapunov, 1992):* They quantify how quickly two trajectories diverge or converge, a hallmark of stability or chaos. For trajectories θ_h and θ'_h starting ϵ apart, $\lambda = \lim_{h\to\infty} \frac{1}{h} \log \frac{||\theta_h - \theta'_h||}{||\theta_0 - \theta'_0||}$. Negative exponents suggest convergence to a noisy fixed point, near-zero values indicate cycles or neutral stability, and positive exponents are evidence of chaos.
- *Recurrence plots (Eckmann et al., 1995):* Dynamical systems over time can visit states recurrently, visualising the pattern of these revisits indicates ordered, periodic, or chaotic behaviour. Defined as a binary matrix R_{ij} where $R_{ij} = 1 \quad \text{if} \quad ||\theta_i - \theta_j|| < \epsilon, \quad \text{else} \quad 0$.
- *Fractal dimensions of attractors (Theiler, 1990):* Chaotic attractors often have fractal geometry occupying a fractional dimension between integer values. Using states θ_i, calculating: $C(r) = \frac{1}{N^2} \sum_{i,j} \mathbf{1}\{||\theta_i - \theta_j|| < r\}$ denotes the probability two points are within distance r. For small r, $C(r) \sim r^{D_2}$ where D_2 is the correlation dimension. $D_2 \approx 0$ is a fixed point, $D_2 \approx 1$ a limit cycle, and D_2 non-integer > 1 a fractal attractor (signal of chaos).

Please refer to Appendix B for details on how these diagnostics are computed.

4 Experimental Settings

We evaluate all methods under a shared experimental protocol across four canonical stateless repeated games: Prisoner's Dilemma, Matching Pennies, Stag Hunt, and Chicken, and a larger multi-state cooperative environment, Simple Spread from Multi Particle Environments (MPE) (Mordatch & Abbeel, 2017). The matrix games capture standard social dilemma, zero-sum, coordination, and competitive structures, respectively, while Simple Spread provides a high-dimensional setting requiring sustained coordination between three agents. We use two or three independent learner algorithms that closely align with replicator dynamics. Tabular Q-learning approximates replicator dynamics in stateless games in the continuous time, infinitesimal step-size limit if it uses Boltzmann (Softmax) exploration Bloembergen et al. (2015). Policy gradient (REINFORCE with baseline) reduces to replicator dynamics in the mean-field limit, since subtracting the baseline corresponds to subtracting the average payoff (Bernasconi et al., 2025). To scale to Simple Spread, we use Independent Deep Q-Networks (IDQN) (Tampuu et al., 2017), introducing function approximation and additional stochasticity. Appendix A presents detailed descriptions of these environments.

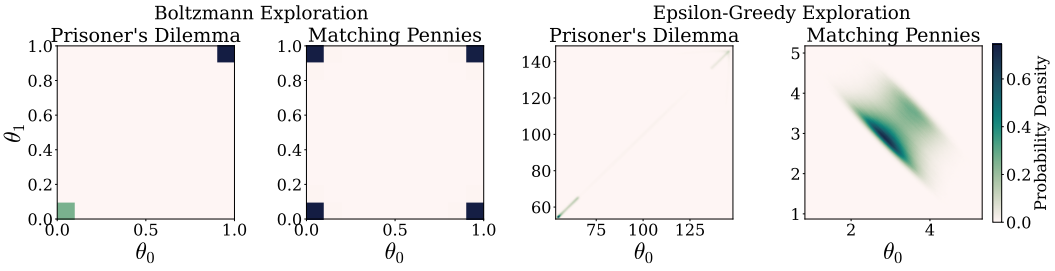

Figure 2: Stationary distributions calculated from realisations of training two IDQN agents. The two left figures are agents using Boltzmann exploration; the parameters can be interpreted as probabilities of taking action 0. On the Boltzmann plots bin counts are intentionally very low so it is clear where the stationary distribution has density. The two right figures are agents using ϵ-greedy exploration. Parameters cannot be interpreted as action probabilities which is why their scale is larger.

5 Understanding Individual Decision-Making in MARL

5.1 Stability Analysis

Stability analysis in MARL concerns asymptotic behaviour and its relation to game equilibria. Evolutionary game theory characterises equilibria as fixed points of deterministic dynamics (Hofbauer & Sigmund, 1998). When replicator dynamics apply, the induced flow admits a closed-form vector field. As our approach is data-driven, we visualise empirical trajectories in parameter space (policy traces). Figure 1 compares phase portraits with policy traces from Q-learning, Policy Gradient, and IDQN with Boltzmann exploration, where parameters correspond to action probabilities. With sufficiently small learning rates, Q-learning and Policy Gradient approximate replicator flow. In contrast, IDQN with Boltzmann exploration introduces additional stochasticity from function approximation, gradient updates, and target networks, yielding behaviours such as expanding cycles in Matching Pennies and drift toward defection in Chicken.

Figure 2 shows stationary distributions for Prisoner's Dilemma and Matching Pennies under IDQN with Boltzmann and ϵ-greedy exploration. Under Boltzmann exploration, the stationary distribution largely concentrates near the sink points predicted by replicator dynamics in Figure 1. In Prisoner's Dilemma this appears as strong concentration near mutual defection, while Matching Pennies exhibits spread across parameter space consistent with cycling behaviour, seen in Figure 1. Some mass in the bottom left indicates that for some initial conditions convergence does not match replicator dynamics. Under ϵ-greedy exploration, dynamics become more varied: Prisoner's Dilemma converges to one of two fixed points, whereas Matching Pennies displays an approximate limit cycle, consistent with quasi-cyclic dynamics.

Further quantitative evidence (Table 1) supports these observations. In Prisoner's Dilemma, $\lambda_{\max} \approx 0$ and fractal dimension $D_2 \approx 0$ indicate convergence to a stable fixed point. In contrast, Matching Pennies exhibits positive $\lambda_{\max}$ and larger D_2, consistent with cyclical behaviour. Figure 3 presents recurrence plots for both games (with the identity line masked). In Prisoner's Dilemma, long diagonal bands parallel to the identity line dominate, indicating deterministic and predictable dynamics (a stochastic process exhibits almost none) (Marwan et al., 2007). The pattern here suggests the system follows a regular, repeating path through its state space. Matching Pennies instead shows regularly spaced diagonals, characteristic of periodic motion, along with arc like structures associated with quasi-periodicity (Facchini & Kantz, 2007).

Thus far, we have considered settings amenable to direct visualisation. As the number of agents or parameters grows, phase portraits become impossible. Although empirical diagnostics remain well-defined, intuitive inspection is hindered. High-dimensional systems typically require projection (e.g., via Principal Component Analysis (Hotelling, 1933)), incurring information loss. Recurrence plots offer an alternative, providing a two-dimensional representation of the dynamics (Eckmann et al., 1995), capturing structural properties such as determinism and convergence (Figure 4).

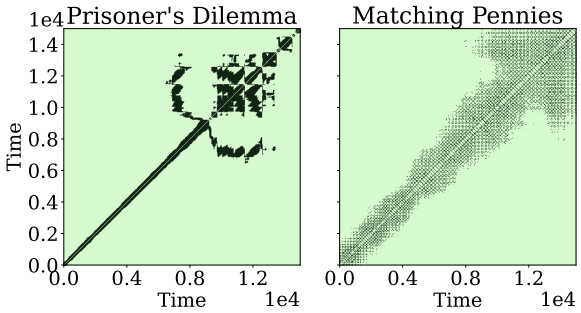

Figure 3: Recurrence plots from a realisation of training two IDQN agents, indicating when the coupled dynamical system of all agents visits the same area in phase space at the time on the x-axis and y-axis. Intuitively, this means locations are marked when $\theta_i \approx \theta_j$ when $i = x, j = y$, with identity band masked out as it is always recurrent when $i = j$ so adds no extra information.

Table 1: Diagnostics for stochastic fixed points, limit/quasi-cycles, and chaotic attractors with 95% confidence intervals over 8 random seeds.

Environment	$\|\Sigma\|_F$	$\lambda_{\max}$	D_2
Prisoners' Dilemma	0.102 ± 0.041	≈ 0	0.438 ± 0.042
Matching Pennies	2.351 ± 0.103	0.039 ± 0.002	1.154 ± 0.075
Stag Hunt	0.553 ± 0.068	≈ 0	0.628 ± 0.050
Chicken	0.118 ± 0.022	≈ 0	0.760 ± 0.034
Simple Spread	0.956 ± 0.015	≈ 0	0.441 ± 0.064

Pairing the recurrence plots with Table 1 for the higher-dimensional cooperative setting of Simple Spread, the diagnostics reveal dynamics between deterministic convergence and cyclical instability. The relatively large Frobenius norm ($||\Sigma||_F = 0.956$) indicates moderate parameter variance among the three agents. Furthermore, the near-zero maximum Lyapunov exponent and the correlation dimension ($D_2 = 0.441$) suggest that the system settles into a state of stability. Rather than collapsing to a strict fixed point ($D_2 \approx 0$), the agents' joint policy forms a quasi-cycle driven by the inherent stochasticity of IDQN's function approximation. Figure 4 corroborates these quantitative findings: exhibiting a broad, dense diagonal band with diffuse edges stemming from exploration noise and multi-agent non-stationarity in contrast to sharp diagonal lines seen in deterministic settings.

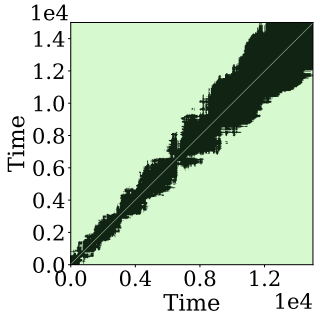

Figure 4: Recurrence plot from a realisation of training three IDQN agents in Simple Spread.

To summarise, analysing the MARL system as an agent-centric coupled dynamical system enables a practitioner to clearly understand the stability and asymptotic performance of an agent's *true* behaviour. Further, our quantitative and qualitative insights are able to scale to environments and agent numbers that can't be easily visualised, opening the door for analysis in any environment. When working with classical state based games, the insights from replicator dynamics generally align with our findings, indicating the usefulness of a paired approach for understanding MARL stability.

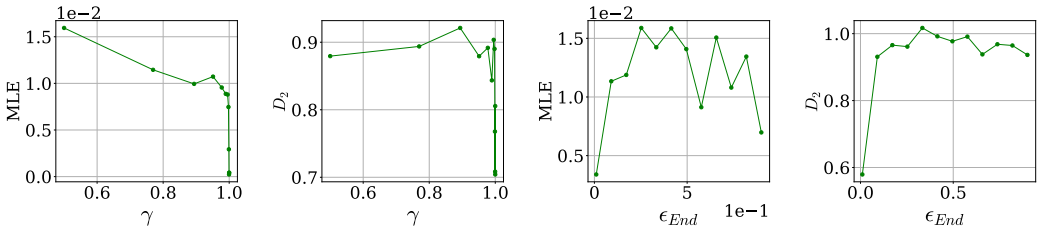

Figure 5: Varying γ, the discounting parameter in IDQN, and ϵ_{End} the end value for ϵ-greedy exploration in IDQN, to understand their impact on the coupled dynamical system attractor via the Maximum Lyapunov Exponent (MLE) and fractal dimension D_2 over 8 random seeds.

5.2 Sensitivity Analysis

We analyse how MARL dynamics change under variations in hyperparameters and learning rules, focusing on topological shifts in phase space induced by different choices of g. We sweep the discount factor γ and terminal exploration rate ϵ_{End} in IDQN (with $\epsilon = 0.9$ and exponential decay), estimating the maximal Lyapunov exponent and correlation dimension D_2 from post-burn-in policy traces across seeds (Figure 5). Increasing γ suppresses cycling in Matching Pennies as $\gamma \to 1$. Setting $\epsilon_{\text{End}} = 0$ removes exploration noise, yielding convergence to a fixed point ($\lambda_{\max}, D_2 \to 0$). Interestingly, small positive ϵ_{End} induces cyclical behaviour, while larger values reduce structure, suggesting convergence toward smaller radius limit cycles or quasi-cycles compared to the case with $\epsilon_{\text{End}} \approx 0.4$. Recurrence plots (Figure 6) corroborate this transition: the far left figure shows a clear limit cycle with determinism, the centre left a quasi-cycle as the spread of diagonal lines indicates higher stochasticity, the centre right has hallmarks of the beginnings of chaos, and finally the far right is almost purely stochastic with very faint structure. These sensitivity experiments reveal not only the phase space *shape*, but also *why* it assumes that shape under different hyperparameter settings, providing a principled connection between tuning choices and long-term dynamical behaviour.

5.3 Control

In the previous sections, having characterised stability and sensitivity through the lens of dynamical systems theory, we propose closing the loop by using these metrics to actively shape MARL behaviour. Rather than serving purely as post-hoc diagnostics, quantities such as the maximal Lyapunov exponent and correlation dimension can define control objectives (e.g., suppressing cycling by driving $\lambda_{\max} < 0$ and $D_2 \to 0$). One possible approach is stability aware MARL, where dynamical metrics act as pseudo-rewards or guide meta-learning of hyperparameters toward desired phase space properties. Such schemes complement reward maximisation by steering learning toward regimes that are not only high performing but also stable and predictable.

6 Implications

Our central aim is to analyse MARL at the level where decisions occur: individual learning updates. Population based approaches (e.g., replicator dynamics) offer aggregate descriptions but smooth out heterogeneity, stochasticity, and algorithmic structure that are central to practical MARL. An agent-centric dynamical systems perspective instead models the coupled update dynamics that drive learning and interaction. This perspective enables principled analysis of stability and sensitivity under stochasticity and function approximation. Tools such as invariant distributions, Lyapunov exponents, and recurrence analysis characterise whether learning converges to fixed points, cycles, or chaotic regimes. Practically, stable equilibria correspond to consistent joint policies, while oscillatory or chaotic behaviour reflects persistent non-stationarity. By examining how hyperparameters or environment perturbations shift these regimes, we obtain measures of robustness and reproducibility. Although conducted in parameter space, results map to reward and behavioural outcomes, enabling assessment of whether convergence is desirable. Applicable to both tabular and deep MARL, this

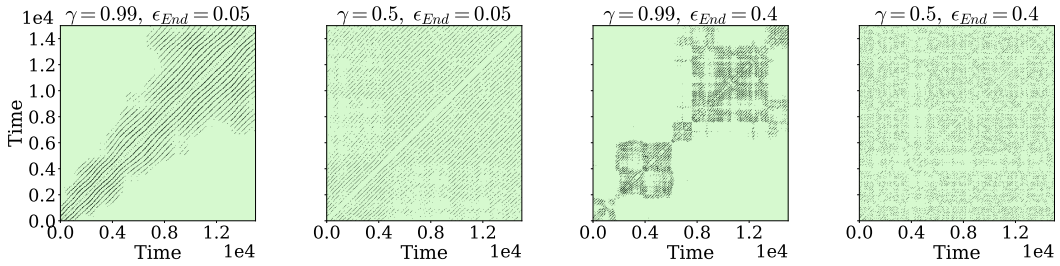

Figure 6: Recurrence plot calculated from a realisation of training two IDQN agents in Matching Pennies. All four combinations of $\gamma = \{0.5, 0.99\}$ and $\epsilon_{\text{End}} = \{0.4, 0.0\}$ are presented.

framework provides diagnostics for instability in high-dimensional settings where mean-field abstractions are intractable, clarifying how design choices shape long-term dynamics.

7 Conclusion

In this work, we proposed an agent-centric dynamical systems framework for analysing MARL, modelling individual learning updates as coupled stochastic dynamical systems, enabling stability and sensitivity analysis beyond deterministic population level abstractions. Using invariant distributions, Lyapunov exponents, recurrence structure, and fractal dimensions, we demonstrated how regime transitions, oscillations, and instability arise in both tabular and deep MARL, and how hyperparameters shape long-run behaviour. Our approach is primarily empirical, relying on finite sample trajectory estimates that may be sensitive to burn-in, dimensionality, and projection effects; scalability to very high dimensional multi-agent systems poses computational challenges; and we do not yet provide formal predictive guarantees linking algorithm design to specific attractor structures.

Future work proceeds in three directions. First, additional tools from dynamical systems theory could broaden stability analysis. Second, stochastic invariant distributions could be characterised analytically via the Fokker-Planck equation (Risken, 1989) or estimated using partial differential equation based methods (Li et al., 2021), enabling direct calculations of stability and sensitivity, as demonstrated by Leung et al. (2023). Third, diminishing sources of stochasticity (e.g., decaying exploration) raise open questions about their impact on attractor structure. Viewing learning as a coupled dynamical process provides a principled foundation for analysing and controlling emergent behaviour in multi-agent systems.

Broader Impact Statement

The methods and findings presented in this work have the potential to support a wide range of socially beneficial applications. For example, improving the stability and robustness of multi-agent learning systems could enhance coordination in domains such as distributed robotics, traffic and resource management, and environmental monitoring, especially in areas where safety is paramount. Such applications can advance sustainability efforts, climate research, biodiversity protection, and responsible resource management. We advocate for responsible innovation and encourage the use of this work in ways that prioritise environmental sustainability, social equity, and the public good.

A Environments

We first consider a set of four stateless matrix games, which serve as canonical benchmarks for analysing multi-agent interaction dynamics. These environments consist of single-step interactions in which agents select actions simultaneously and receive rewards according to fixed payoff matrices. As there is no temporal component or state transition, they provide a controlled setting for examining coordination, competition, and equilibrium behaviour under simplified yet illustrative conditions. The environments are:

- *Prisoner's Dilemma*: a standard social dilemma with a unique Nash equilibrium (defection), though cooperation can arise under repeated interaction Axelrod & Hamilton (1981); Nowak (2006).
- *Matching Pennies*: a zero-sum game with a mixed-strategy equilibrium.
- *Stag Hunt*: a coordination game with multiple equilibria (cooperative "stag" or risk-dominant "hare").
- *Chicken*: a cooperative/competitive game where mutual aggression is costly, with two pure-strategy equilibria (one player swerves while the other does not) and a mixed-strategy equilibrium.

The pay-off matrices of these games are reported in the tables below.

Prisoner's Dilemma

	C	D
C	1,1	5,0
D	0,5	3,3

Matching Pennies

	H	T
H	1,-1	-1,1
T	-1,1	1,-1

Stag Hunt

	S	H
S	4,4	0,0
H	0,0	3,3

Chicken

	A	B
A	-1,-1	4,0
B	0,4	2,2

We also use the `Simple Spread MPE` environment from the `JaxMARL` benchmark suite (Rutherford et al., 2024). In this cooperative multi-agent task, multiple agents must coordinate to cover a set of target landmarks in a shared continuous space. Each agent aims to occupy a different landmark while avoiding collisions with other agents, requiring effective spatial coordination and decentralised decision making. Agents must distribute themselves across landmarks such that each landmark is covered by a different agent while minimising collisions. The environment's partially observable state and the need for coordinated positioning make it a useful testbed for studying cooperation and learning dynamics in MARL.

B Dynamical Systems Analysis Techniques: Additional Details

This appendix details the numerical methods used to analyse emergent agent dynamics, implemented via vectorised operations in `JAX` (Google Research, 2023) and `NumPy` (Harris et al., 2020).

B.1 Stationary Distribution Estimation

We estimate the invariant measure of the coupled stochastic learning process empirically from long-run policy trajectories. Simulating the system for n_{steps} iterations across n_{runs} random seeds and discarding a burn-in phase of n_{burn} steps, the stationary samples for agent parameters $\boldsymbol{\theta}_h = (\theta_h^1, \theta_h^2)$ at update step h are collected as:

$$\mathcal{S} = \bigcup_{i=1}^{n_{\text{runs}}} \{\boldsymbol{\theta}_h^{(i)} : h > n_{\text{burn}}\}. \tag{10}$$

For scalar two-agent parameters, these form a normalised 2D empirical density $\hat{p}(\theta^1, \theta^2)$; for higher-dimensional spaces or more agents, dimensionality reduction is applied prior to visualisation.

B.2 Covariance Analysis and Frobenius Norm

For a multivariate trajectory $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^n)$ over H total update steps, linear parameter dependencies are captured by the empirical covariance matrix $\Sigma = \frac{1}{H-1}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})$. To summarise stationary variability and coupling strength as a single scalar, we compute its Frobenius norm (Böttcher & Wenzel, 2008):

$$\|\Sigma\|_F = \sqrt{\sum_{i,j} \Sigma_{ij}^2}. \tag{11}$$

Unlike trace-based metrics, the Frobenius norm incorporates cross-covariance terms, providing a compact descriptor of joint fluctuation complexity.

B.3 (Maximum) Lyapunov Exponent Estimation

To quantify local sensitivity to initial conditions, we estimate the maximal Lyapunov exponent $\lambda_{\max}$ from multivariate trajectories $\boldsymbol{\theta} = [\theta_h]_{h=1}^{H} \in \mathbb{R}^{H\times\Theta}$ via nearest-neighbour divergence. For each state θ_i, its nearest neighbour $\theta_{j(i)}$ is found outside a Theiler window of $\pm w$ steps (Theiler, 1990). The Euclidean distance between forward trajectories at lag z, $d_i(z) = \|\theta_{i+z} - \theta_{j(i)+z}\|_2$, is tracked over $z \in [z_{\text{min}}, z_{\text{max}}]$, yielding:

$$\lambda_{\max} \approx \frac{d}{dt}\langle \log d(z)\rangle. \tag{12}$$

A positive $\lambda_{\max}$ denotes exponential divergence and chaotic learning dynamics.

B.4 Recurrence and Correlation Plots

Recurrence plots visualise phase-space revisitations by thresholding pairwise state distances:

$$R_{ij} = \mathbf{1}\{\|\theta_i - \theta_j\|_2 \le \varepsilon\}, \tag{13}$$

where ε is calibrated to achieve a target recurrence rate (e.g., 8%). The binary matrix R encodes temporal proximity: diagonal structures indicate epochs of predictability, while scattered points reflect stochastic or chaotic transitions (Eckmann et al., 1995; Marwan et al., 2007; Facchini & Kantz, 2007).

B.5 Correlation Dimension (Fractal Dimension)

We compute the correlation (fractal) dimension D_2 of the underlying attractor using the Grassberger–Procaccia algorithm (Grassberger & Procaccia, 1984). Given a delay embedding $\Theta_h = [\theta_h, \theta_{h+\tau}, \ldots, \theta_{h+(m-1)\tau}] \in \mathbb{R}^m$ over N reconstructed vectors, we evaluate the correlation sum across logarithmic radii r:

$$C(r) = \frac{2}{N(N-1)} \sum_{i<j} \mathbf{1}\{\|\Theta_i - \Theta_j\|_2 < r\}. \tag{14}$$

In the linear scaling regime where $C(r) \propto r^{D_2}$, regression of $\log C(r)$ against $\log r$ yields D_2.

B.6 Summary of Computed Metrics

In summary, for each analysed trajectory the following quantities are reported:

- Stationary distribution $\hat{p}(\boldsymbol{\theta})$: empirical invariant measure.
- $\|\Sigma\|_F$: Frobenius norm of the covariance, capturing overall variability.
- λ_{max}: largest Lyapunov exponent, measuring chaotic divergence.
- Recurrence plots: graphical tools used to visualize the times at which a dynamical system returns to states similar to previous ones, as defined by a chosen error tolerance.
- D_2: correlation (fractal) dimension of the reconstructed attractor.

Together, these diagnostics provide a comprehensive characterisation of the dynamical complexity, stationarity, and stability properties of the interacting reinforcement learning agents.

Acknowledgments

James Rudd-Jones was supported by the Engineering and Physical Sciences Research Council through a Doctoral Training Partnership (DTP) PhD studentship (grant number EP/W524335/1 - Award 28684830).

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems (NeurIPS'21)*, 2021.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems using Reinforcement Learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI'20)*, 2020.

Nicolas Anastassacos, Julian Garcia, Stephen Hailes, and Mirco Musolesi. Cooperation and Reputation Dynamics with Reinforcement Learning . In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'21)*, 2021.

Robert Axelrod and William D Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

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

Randall D Beer. A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, 72(1-2):173–215, 1995.

Martino Bernasconi, Federico Cacciamani, Simone Fioravanti, Nicola Gatti, and Francesco Trovò. The evolutionary dynamics of soft-max policy gradient in multi-agent settings. *Theoretical Computer Science*, 1027:115011, 2025.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemysław Dębiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53, 2015.

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

Albrecht Böttcher and David Wenzel. The frobenius norm and the commutator. *Linear algebra and its Applications*, 429(8-9):1864–1885, 2008.

John G Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87(2):239–266, 1973.

Nayana Dasgupta and Mirco Musolesi. Investigating the impact of direct punishment on the emergence of cooperation in multi-agent reinforcement learning systems. *Autonomous Agents and Multi-Agent Systems*, 39(1):1–37, 2025.

Jean-Pierre Eckmann Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. In *Turbulence, Strange Attractors and Chaos*, pp. 441–445. World Scientific, 1995.

Andrea Facchini and Holger Kantz. Curved structures in recurrence plots: The role of the sampling time. *Physical Review E*, 75(3):036215, 2007.

J Doyne Farmer. Information dimension and the probabilistic structure of chaos. *Zeitschrift für Naturforschung A*, 37(11):1304–1326, 1982.

Luis H Favela. Dynamical systems theory in cognitive science and neuroscience. *Philosophy Compass*, 15(8):e12695, 2020.

Drew Fudenberg and Christopher Harris. Evolutionary dynamics with aggregate shocks. *Journal of Economic Theory*, 57(2):420–441, 1992.

Tobias Galla and J Doyne Farmer. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

Aram Galstyan. Continuous strategy replicator dynamics for multi-agent q-learning. *Autonomous Agents and Multi-agent Systems*, 26(1):37–53, 2013.

Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

David Goll, Jobst Heitzig, and Wolfram Barfuss. Deterministic Model of Incremental Multi-Agent Boltzmann Q-Learning: Transient Cooperation, Metastability, and Oscillations. *arXiv preprint arXiv:2501.00160*, 2024.

Brain Team Google Research. JAX: Autograd and XLA for high-performance machine learning research, 2023. URL https://github.com/google/jax.

Peter Grassberger and Itamar Procaccia. Dimensions and entropies of strange attractors from a fluctuating dynamics approach. *Physica D: Nonlinear Phenomena*, 13(1-2):34–54, 1984.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI'18)*, 2018.

Daniel Hennes, Karl Tuyls, and Matthias Rauterberg. State-Coupled Replicator Dynamics. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI'09)*, 2009.

Daniel Hennes, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duéñez-Guzmán, et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'20)*, 2020.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417, 1933.

Aamal Hussain, Francesco Belardinelli, and Georgios Piliouras. Beyond strict competition: Approximate convergence of multi agent q-learning dynamics. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI'23)*, 2023.

Lorenz A. Imhof. The long-run behavior of the stochastic replicator dynamics. *The Annals of Applied Probability*, 15(1B):1019–1045, 2005.

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 International Conference on Autonomous Agents and Multiagent Systems (AAMAS'17)*, 2017.

Chin-wing Leung, Shuyue Hu, and Ho-fung Leung. The stochastic evolutionary dynamics of softmax policy gradient in games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS'24)*, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*, 2021.

Aleksandr Mikhailovich Lyapunov. The general problem of the stability of motion. *International Journal of Control*, 55(3):531–534, 1992.

Norbert Marwan, M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, 2007.

John Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

Eric Mazumdar, Lillian J. Ratliff, Michael I. Jordan, and S. Shankar Sastry. Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'20)*, 2020.

Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

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

Martin A Nowak and Karl Sigmund. Evolutionary dynamics of biological games. *science*, 303 (5659):793–799, 2004.

Hannes Risken. Fokker-planck equation. In *The Fokker-Planck Equation: Methods of Solution and Applications*, pp. 63–95. Springer, 1989.

James Rudd-Jones, Mirco Musolesi, and María Pérez-Ortiz. Multi-Agent Reinforcement Learning Simulation for Environmental Policy Synthesis. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'25)*, 2025a.

James Rudd-Jones, Fiona Thendean, and María Pérez-Ortiz. Crafting desirable climate trajectories with reinforcement learning explored socio-environmental simulations. *Environmental Data Science*, 4:e41, 2025b.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Garðar Ingvarsson Juto, Timon Willi, Ravi Hammond, Akbir Khan, Christian Schroeder de Witt, et al. JaxMARL: Multi-Agent Reinforcement Learning Environments and Algorithms in JAX. *Advances in Neural Information Processing Systems (NeurIPS'24)*, 2024.

James BT Sanders, J Doyne Farmer, and Tobias Galla. The prevalence of chaotic dynamics in games with many players. *Scientific Reports*, 8(1):4902, 2018.

Yuzuru Sato, Eizo Akiyama, and J Doyne Farmer. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

Jeffrey D Scargle. An introduction to chaotic and random time series analysis. *International Journal of Imaging Systems and Technology*, 1(2):243–253, 1989.

Steven H Strogatz. *Nonlinear Dynamics and Chaos: with Applications to Physics, Biology, Chemistry, and Engineering*. Chapman and Hall/CRC, 2024.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. Multiagent cooperation and competition with deep reinforcement learning. *PLOS One*, 12(4):e0172395, 2017.

James Theiler. Estimating fractal dimension. *Journal of the Optical Society of America A*, 7(6): 1055–1073, 1990.

Karl Tuyls, Pieter Jan'T Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12(1): 115–153, 2006.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, volume 80, pp. 5571–5580, 2018.

Tianyu Zhang, Andrew Williams, Phillip Wozny, Kai-Hendrik Cohrs, Koen Ponse, Marco Jiralerspong, Soham Phade, Sunil Srinivasa, Lu Li, Yang Zhang, Prateek Gupta, Erman Acar, Irina Rish, Yoshua Bengio, and Stephan Zheng. AI for Global Climate Cooperation: Modeling Global Climate Negotiations, Agreements, and Long-Term Cooperation in RICE-N. In *Proceedings of the 42nd International Conference on Machine Learning (ICML'25)*, 2025.