Journal of High School Science

Instability as insight: reinterpreting learning dynamics in repeated games through bounded rationality

Mahadevaiah P

Submitted: November 8, 2025, Revised: version 1, May 16, 2026, version 2, May 16, 2026, version 3, May 17, 2026, version 4, May 19, 2026, version 5, May 19, 2026
Accepted: May 20, 2026

Abstract
The literature on repeated games has traditionally treated adaptive strategic behavior through frameworks such as reinforcement learning and evolutionary dynamics, both of which largely evaluate learning in terms of convergence to equilibrium or stochastic fluctuations around it. While analytically convenient, this perspective overlooks a fundamental feature of real multi-agent systems: agents operate under bounded cognition, finite computational resources, imperfect information, and continually changing strategic environments. As a result, their behavior is rarely perfectly stationary. This paper revisits Q-learning and replicator dynamics through the framework of bounded rationality, arguing that instability, oscillation, and partial convergence should not be viewed merely as failures of learning or artifacts of imperfect models. Rather, they are often natural consequences of adaptation under realistic constraints. Behaviors commonly dismissed as noise—including persistent cycling, metastability, and non-convergent trajectories—may instead contain information about strategic experimentation, exploration, and ongoing adaptation. Building on this perspective, we propose an initial operational framework for studying bounded-rational learning dynamics using tools from information theory and dynamical systems analysis, including entropy measures, Lyapunov exponents, information flow metrics, and continuous measures of rationality. These tools allow instability to be analyzed quantitatively rather than treated solely as a deviation from equilibrium. By shifting attention from equilibrium attainment to the structure of adaptation itself, this framework broadens the study of learning in repeated games and suggests new empirical, computational, and analytical approaches for investigating how agents learn within complex strategic environments. More generally, treating instability as a potentially informative feature rather than as residual noise may contribute to a richer mathematical understanding of adaptive flexibility in both artificial and natural systems.

Keywords
Adaptive learning dynamics, Transfer learning, Sequential games, Cross-game generalization, Optimization history dependence, Non-equilibrium learning, Reinforcement learning, Path dependence, Information retention, Emergent behavioral structure

Pratyush Mahadevaiah, Dougherty Valley High School, 733 Clifton Ct, San Ramon, CA, 94582, USA.
pratyush.mahadevaiah@gmail.com

https://doi.org/10.64336/001c.163090

1. Introduction

In the context of repeated games, convergence towards equilibrium, stability, or mutual consistency has historically been the theoretical and evaluative goal of learning. Traditional accounts posit, that given enough experience, rational agents will eventually coordinate with one another on equilibrium strategies that maximize their expected utilities. This model underpins much of our economic reasoning, our thinking about learning, and about artificial intelligence. The promise held by the convergence paradigm is precisely predictability: a world where characteristically learning is done, and rational order takes over. However, as we will argue in this work, the emphasis on predictability and convergence may obscure important aspects of how learning unfolds in adaptive systems. The following sections outline the conceptual foundations and historical appeal of convergence as the dominant paradigm for understanding learning in games (1).

2. Methods

2.1 Literature review

This paper conducts a systematic review of the literature on game theory, learning dynamics, bounded rationality, and information theory. The review was performed using major academic databases, including JSTOR, ScienceDirect, SpringerLink, IEEE Xplore, and Google Scholar. Searches employed combinations of keywords such as *repeated games*, *learning in games*, *bounded rationality*, *multi-agent reinforcement learning*, *replicator dynamics*, *information-theoretic entropy*, and *Lyapunov stability*. The resulting literature was examined to identify major theoretical developments, empirical findings, and emerging connections among adaptive learning, bounded rationality, and dynamical systems perspectives. Priority was given to peer-reviewed journal articles, conference papers published in leading venues such as NeurIPS and AAAI, and scholarly monographs published by academic presses including MIT Press, Cambridge University Press, and Princeton University Press. Preference was also given to papers with a high citation index or those which presented important or fundamental algorithms or theories.

An initial bibliography of more than 120 publications was assembled and screened for relevance to the central focus of this study: learning dynamics in repeated games under cognitive, informational, and computational constraints. Sources were subsequently categorized into several thematic areas, including classical convergence theory, canonical learning models, empirical evidence of non-convergence, bounded rationality frameworks, and information-theoretic interpretations of adaptive behavior. Within each category, publications were evaluated according to their underlying assumptions, methodological approaches, and conclusions regarding learning dynamics and strategic adaptation.

The final corpus consisted of approximately 60 peer-reviewed sources selected to provide both historical continuity and coverage of contemporary developments. Care was taken to maintain representation across multiple disciplinary perspectives, including economics,

artificial intelligence, behavioral decision-making, and dynamical systems theory, while emphasizing influential and foundational contributions to the study of learning in repeated games.

2.2.1 *Learning algorithms*

Q-learning agents maintained action-value estimates Q(*a*) updated via the standard rule $Q(a) \leftarrow Q(a) + \alpha\ (r + \gamma\ \max Q - Q(a))$, with learning rate $\alpha = 0.1$, discount factor $\gamma = 0.95$, initial Boltzmann temperature $\tau = 1.0$, and exponential temperature decay rate 0.9998. Policies were computed via the Boltzmann softmax function. Each simulation ran for T=10,000 timesteps in the primary analysis and T = 2,000-3,000 timesteps in the extended analyses, sufficient to observe both transient and asymptotic behavior given the temperature decay schedule. Replicator dynamics used timestep $dt = 0.01$ with Gaussian noise of scale 0.005.

2.2.2 *Diagnostic suite*

For each game, a ten-item diagnostic suite was computed: [1] strategy trajectory plots for both agents under both learning rules; [2] entropy time-series of the Boltzmann policy with sliding-window smoothing; [3] maximal Lyapunov exponent estimation via nearest-neighbor divergence; [4] transfer entropy between agents in both directions, measuring directional information flow; [5] KL divergence from uniform over time, tracking adaptive rationality; [6] PCA of joint strategy evolution, measuring attractor dimensionality; [7] frequency spectra via FFT, identifying quasi periodic structure; [8] eigenvalue analysis of the Jacobian of the replicator dynamics at quasi stationary points; [9] Poincare return maps; and [10] surrogate data significance testing via temporal shuffling, distinguishing structured dynamics from noise.

2.2.3 *Multi-seed robustness analysis*

To assess robustness to initial conditions, we conducted 20 independent simulation runs per game using different random seeds. For each run, we computed the Lyapunov proxy (mean log-norm of successive strategy differences), strategy entropy, and entropy rate in the final quarter of the trajectory. We reported mean plus or minus standard deviation, 95% confidence intervals, and the fraction of runs classified as cycling versus convergent (with a threshold of $H > 0.01$ bits for cycling). Statistical separation between game classes was assessed via box plot distributions of the Lyapunov proxy.

2.2.4 *Parameter sensitivity analysis*

To identify phase transitions and assess robustness of structured instability across parameter regimes, we performed sweeps across four dimensions for four representative games: learning rate alpha (0.01 to 0.5), temperature decay rate (0.999 to 1.0), replicator noise scale (0.0 to 0.1), and payoff perturbation magnitude (0.0 to 1.0), applied as additive Gaussian noise to the payoff matrix). For each parameter setting, we recorded the Lyapunov proxy and second-half entropy, and classified the dynamics as convergent, edge-of-chaos, or structured chaos.

2.2.5 *Temporal windowed analysis*

To distinguish stationary bounded cycling from adaptive dynamical reorganization, we divided

each trajectory into 10 equal temporal windows and computed per-window diagnostics: entropy, Lyapunov proxy, PCA variance concentration (PC1 percentage), strategy spread (mean standard deviation of strategy components), transfer entropy, and entropy rate. Linear trend lines with slope annotations were fitted to each metric across windows. The key diagnostic was the slope: flat slopes indicated stationary dynamics, while systematically declining entropy or increasing PCA concentration indicated progressive attractor reorganization.

2.2.6 *Cross-game transfer experiments*

To test whether game topology leaves persistent latent structure in the learned state, we designed two sets of transfer experiments. In each experiment, two Q-learning agents trained on a source game for 2,000 timesteps. Their complete internal state (Q-values and current temperature) was then preserved, and they continued learning on a target game for 3,000 additional timesteps. This transfer condition was compared against a control in which fresh agents (zero initialized Q-values, same temperature) learned the target game from scratch. The first set tested cross-topology transfer with four cases: (A) RPS to Prisoner's Dilemma, (B) Prisoner's Dilemma to RPS, (C) RPS to Stag Hunt, and (D) Stag Hunt to RPS. The second set tested adversarial-to-adversarial transfer with four additional cases: (E) RPS to Shapley, (F) Shapley to RPS, (G) RPS to Matching Pennies, and (H) Matching Pennies to RPS. These same-topology cases were additionally compared against convergent-to-adversarial baselines to isolate the specific contribution of adversarial pre-training. Each experiment was replicated across 15 independent random seeds. The primary comparison metrics were early-adaptation entropy (mean over the first 500 timesteps of phase 2), time to cycling (first timestep at which smoothed entropy exceeded 0.5 bits), early policy support (number of actions with mean probability exceeding 0.1), and early strategy spread (mean standard deviation of strategy components). Significance was assessed via Mann-Whitney U tests.

3. Theoretical background and conceptual framework

3.1 The classic goal of repeated games: Nash equilibrium

Since its emergence in the mid-20th century, game theory, with the Nash equilibrium (12) as its primary organizing principle, has undergone a transformation. The Nash equilibrium represents a strategy profile such that no agent can deviate from the strategy profile to their payoff advantage, achieving the mutual best response for cooperation in the context of known strategies and under certain reward structures. It embodies a state of expectations, strategies, and motivations. In repeated games, Nash equilibrium came to be viewed not merely as a static solution concept, but as the expected outcome of an adaptive learning process. As rational agents repeatedly interact, observe the behavior of others, and adjust their strategies in response to accumulated experience, equilibrium behavior is expected to emerge and, under favorable conditions, eventually converge. Consequently, convergence to equilibrium has often been treated as a hallmark of successful learning and

has served as a proxy for rationality, stability, and theoretical consistency within models of strategic adaptation (13).

Beginning in the 1950s, the equilibrium concept became the central organizing principle for understanding learning and adaptation in game theory. Early models such as fictitious play (3) addressed the initial intuitions of agents making gradual updates to beliefs about opponents' strategies, and best responding to the empirical distribution of previous plays. Evolutionary dynamics (14) and reinforcement learning (*Reinforcement Learning: An Introduction*, 1998) eventually expanded this discussion with models to incorporate antecedents of successful strategies into future strategies. The central premise was that repeated interaction and feedback would gradually shape behavior, with strategic adjustments occurring in proportion to past successes and failures, ultimately guiding the system toward a stable equilibrium (15).

Emphasis on convergence is not only a mathematical convenience, but it represents a deep philosophical impulse in the discipline. Equilibrium suggests some kind of predictive closing - a single, internally consistent point from which future dynamics can be inferred. For many analytical models, convergence is a matter of tractability - it allows researchers to calculate long-run payoffs, stability properties, and welfare results (1). For studies based on simulation, convergence is an empirical signal that the agents' learning algorithms are functioning "correctly" (16).

But this convergence framework, while elegant, lacks considerable scope about what learning dynamics in repeated games can - and often do - look like. A number of empirical and computational studies, both in the laboratory and in the field (as well as simulations), have shown that agents exhibit cyclic, chaotic, or metastable styles of behavior that never settle into fixed equilibrium points. These alternate outcomes are typically and routinely dismissed as pathologies, even as numerical noise, model misspecifications, or failures in convergence-along the axis of learning rather than as possibly meaningful variations of adaptive rationality (17).

3.1.1 *Fictitious play, Q-learning, and replicator dynamics as canonical models*

Over the years, there has been an extensive development of mathematical frameworks that formalize how rational agents learn through repeated interaction. Of these, fictitious play, Q-learning (18), and replicator dynamics (19) are recognized as established models of learning, each with a unique interpretation of learning and rationality. However, they nonetheless share the same end goal, which is to converge to equilibrium. Overall, they provide the foundation to understand theorists' imaginations of adaptive behavior in strategic settings.

Fictitious play may be the original and most straightforward rational model of learning in games. First described by George Brown, it allows each player to assume that the empirical frequencies of the other player's past actions as the subjectively-held "belief," in terms of the other player's mixed strategy (20). Each player, at every round, acts in the best-reply to this

belief. More formally, if a_t^{-i} is the action of player *i* during round *t* and p_t^{-i} is the empirical distribution of the other players action in round t, than it will hold that, $a_{t+1}^i \in \arg\max_{a^i} E\, a^{-i} \sim p_t^{-i}\left[u_i(a^i, a^{-i})\right]$ (21).

In this framework, players are not necessarily optimizing over the long-run but are iteratively responding to the experiences they observe (7). In some instances - in particular, in zero-sum games or in potential games - fictitious play can be shown to converge to a Nash equilibrium. However, even in the mathematical case, non-cooperative environments can show cycling or a chaotic dynamic.

While fictitious play is deterministic and belief-based, Q-learning - that originates from reinforcement learning - represents a computationally motivated and reward based notion of adaptation (22). In Q-learning, agents maintain a table of Q(*s,a*) that represents the expected cumulative future rewards of taking action *a* in state *s* and then acting optimally thereafter. The update rule,

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

describes how agents adjust their expectations based on their experiences. In multi-agent settings, this learning process is interconnected: agents obtain rewards based on what other agents do as well as their learning over time (23). While single agent Q-learning is known to converge under certain conditions, the convergence of multi-agent Q-learning is not guaranteed. In particular, the non-stationary environment - agents are changing their policies as other agents are learning - creates a feedback loop that usually leads to oscillation, metastable behavior, or chaos. Regardless, Q-learning continues to be one of the dominant models of adaptation in repeated games and has been the basis of much of the contemporary research in multi-agent reinforcement learning (4).

Conversely, replicator dynamics arise from the literature on evolutionary game theory and interpret learning not as individual optimization, but as adaptation at the population level, meaning the proportion of individuals employing any given strategy will increase in proportion to its relative payoff advantage (14). The typical continuous-time replicator equation $\dot{x}_i = x_i\left[(A\,x)_i - x^T A\,x\right]$ describes how the frequency of strategy *i* evolves based on payoff matrix *A* with x_i being the proportion of the population playing strategy *i*. The replicator equation is similar to Q-learning or fictitious play in the sense that it assumes that players do not deliberate or foresee, only imitate the successful play of others (21). Nevertheless, the replicator dynamics captures many of the same properties of seeking equilibria. The stable equilibria of the replicator dynamics are the evolutionarily stable strategies (ESS) which are the evolutionary analogue of Nash equilibria (24). It is important to note that, like many linear dynamic systems, the replicator dynamics can yield limit cycles and chaotic orbits; even for many simple games. As such, learning and equilibrium are not comparable in dynamic

systems (1).

In sum, these various lines of modeling adaptation in repeated games offer many directions for an empirical science of adaptation (1). Fictitious play emphasizes belief formation and best-response reasoning, Q-learning operationalizes optimization of experience-driven rewards to payoffs, and replicator dynamics abstracts learning into aggregate imitation. Despite the obvious distinctions of the three models, they share a common measure of success, which is stability (25). When trajectories converge to fixed points, it is often interpreted in the research literature that rationality was achieved. However, when trajectories diverge, oscillate or fluctuate indefinitely - that is generally interpreted as either a misspecification of the model or as a transitory or ephemeral state (26).

However, new evidence from computational economics, behavioral experiments and artificial multi-agent systems indicates that divergence is not a failing of rationality, but a natural aspect of adaptive systems attempting to work with bounded information and finite computation (27). Continued instability may encode attempts to coordinate, partially learn, or evolving expectations, which are not part of the explanatory capacity of traditional equilibrium-based models (28).

3.1.2 *Convergence as the gold standard*

In the historical background of game theory and learning models, convergence to equilibrium has been viewed as the highest manifestation of success: the mathematical and conceptual destination to which all rational learning collectively aspires (1). This idealization of equilibrium has deep roots in the theoretical architecture of economics and the methodological habits of computer science. At equilibrium, all agents' expectations are consistently aligned; no player has the incentive to deviate profitably; and no further learning is needed. The system has reached a fixed point of mutually best responses - a state of perfect foresight and strategic alignment. For many theorists, equilibrium represents more than a convenient analytical endpoint; it signifies the completion of the learning process itself. Once equilibrium is reached, no further adaptation is required, and learning is effectively presumed to have ceased (24).

In mathematical terms, convergence is inherently pleasing. Equilibrium points are readily conceptualizable: they can be modeled as closed-form solutions, tested for stability, and subjected to comparative statics (29). When scholars can demonstrate that a particular algorithm or learning process converges to a Nash equilibrium, they receive something akin to a theoretical certification that their model "makes sense." The ability to demonstrate convergence reassures us that the processes by which agents learn behave consistently, rationally, and with bounds. This sense of formal control - of being able to specify, articulable and anticipate formal consequences, has rendered equilibrium both an ideal in theory and a benchmark when assessing learning algorithms in practice (30). Convergence means order, predictability, success, in economics, computer science, and the study of evolutionary biology (1).

From a methodological standpoint, this preference is supported by the abstract, notably, the simulation of repeated games typically entails stochastic learning models which generate long, noisy trajectories (31). Moreover, if there is no criterion for convergence, it is difficult to know what to make of the possible outcomes - when did the learning stop? What does success look like? Thus, the concept of convergence serves as a measurement device: the researchers can summarize an often complex dynamic in a single number - the distance from equilibrium (32). Even in cases where instability or oscillation exists the interpretation is that this transitory phenomenon will disappear with more iterations or smaller steps or better tuning of parameters (17). The assumption is that "true" rationality is in the limit.

From a philosophical perspective, equilibrium represents a strong ideal from the Enlightenment: that rational order exists despite apparent chaos. Here again, the equilibrium concept is not purely descriptive, but prescriptive. To be rational, in this mode of thinking, is to be consistent; to learn is to remove uncertainty; to converge is to arrive at truth (14). This intellectual inheritance is present even in current AI research, where convergence proofs are frequently constructed as approximating intelligence. An algorithm that does not converge is labeled as "unstable" or "divergent" which has normative, as well as descriptive, connotations (17). As a result, developers use the term convergence both literally and to create legitimacy for some moral or epistemic good.

Nevertheless, the consideration of convergence as the "gold standard" carries with it an implicit argument: that stability equals intelligence (23). However, this logic is now being questioned (33). In both empirical and computational cases, rational agents often fail to converge - not because they are irrational, but because of the instability of the environment, the nonlinearity of the feedback loops, and the incompleteness of knowledge. In fact, it does not mean that agents are failing to 'be rational', rather, non-convergence is an inevitable consequence of the process of adaptation and the constraints imposed on the agent. Nonetheless, agents are still learning, exploring, responding to feedback, and making adjustments, even if their joint behavior never achieves a stable convergence (19).

Consequently, rethinking convergence's privileged status is not about moving away from that rigor in game theory, but broadening its scope. It suggests that not reaching equilibrium is as important and informative as reaching it - that the patterns and regularities that exist in instability may reveal larger laws of adaptive behavior (34). In this light, convergence may be better understood not as a universal benchmark of rational learning, but as a methodological ideal whose relevance depends on the structure of the environment and the assumptions imposed on the agents (33). In many complex, bounded, and interactive settings, persistent adaptation may be a more realistic outcome than equilibrium attainment (35).

3.1.3 *Ignoring instability as an object of study*

By conceptualizing rationality in terms of

equilibrium, standard game theory implicitly removes adaptation from its own characterization of success (12). Once a system ceases to evolve (or stops changing), learning is assumed to be complete. However, in real adaptive systems - whether in economies, ecosystems, or autonomous AI networks - learning is often continuous. Agents will never face exactly complete and mutually predictable behavior due to feedback loops, delays in information, and computational possibilities. The result will not be chaos without structure, but structure without stability. To deny this structure is to deny the very dynamics of bounded rational agents in their environments (14).

The conceptual dilemma is not merely that convergence fails sometimes, but that the field lacks the language and structure through which to understand what non-convergence means. Current models can only classify behaviors as being converged (rational) or diverged (irrational) in a basic binary sense. Between these poles exists a vast space of structured instability - periodic orbits, quasi-cycles, metastable attractors - that remain invisible in any meaningful theoretical way under equilibrium assumptions (36). This blindspot has stalled the subfield from making sense of how learners behave when rationality is bounded.

To appreciate instability as an object of study will require us to rethink the very goals of learning theory. Rather than focusing on whether agents converge, we could ask how their deviations evolve, what structure is encoded in those deviations, and what information is provided about adaptive behavior under constraint (37). Only through elevating instability from anomaly to data will we create a richer and more realistic account of learning in repeated games.

To make sense of the history and durability of the convergence paradigm, we need to first return to the formal machinery of repeated games (38). This section will summarize the mathematical and conceptual framework for Nash equilibrium and the best-response reasoning as the foundation of rational agents interacting in a game, fictitious play, Q-learning and evolutionary dynamics as models of the adaptive-learners, and metrics of equilibrium deviation to provide an evaluative lens. Collectively, these concepts provide the main lexicon of learning theory-how we model rationality, how to define success, and how informal "understanding" is formalized in multi-agent systems (18). The purpose of this section is not to dismiss these frameworks, but to articulate their assumptions and strengths clearly enough to identify where their explanatory power begins to encounter important limitations in the analysis that follows (33).

3.1.4 *Thesis and conceptual direction*

We contend that instability in repeated games is not a failure of rationality. It is a reflection of the boundaries of rationality (39). All agents learning under constraints of finite information, finite computation, and evolving expectations cannot be assured that they will converge to equilibrium or that such convergence is optimal. The predominant paradigm of convergence, while elegant from a

mathematical perspective, equates rationality with stability, and non-convergence and non-optimality with rational error (33). This review challenges that assumption by reframing instability as a structured and potentially informative consequence of bounded-rational learning.

We posit that oscillations, cycles, and even chaotic trajectories in repeated games, are interpretable as information about the way agents are adapting under those constraints, rather than as noise (40). It is possible to analyze those behaviors by developing quantitative constructs through measures such as entropy, mutual information, and dynamical sensitivity, which operationalize information on the structure of the apparent disorder. And finally, by placing instability at the center of the inquiry, we propose a new research agenda for game theory - it is one that embeds behavioral fidelity to the bounded rational learning problems, alongside integrating dynamical analysis with information theory (2) to pursue an understanding of learning not as certainty but an operational-dynamic management of uncertainty (41).

3.2.1 *Nash equilibrium, best Response, and mixed strategy*

Game theory, by definition, provides a framework for analyzing interaction among agents capable of rationality; of whose outcomes depend not only on their actions but others combined with the environment. A game in its normal form is usually defined as $G = (N,\{A_i\}i \in N, \{u_i\}i \in N)$, where $N = \{1, 2,..., n\}$ is the set of players, A_i is the set of actions available to player i, and $u_i : A_i \times \ldots \times A_n \rightarrow \mathbb{R}$ is the payoff function that assigns a numerical utility to each player given the joint action profile $a = (a_1,. . . ., a_n)$. Knowing that each player seeks their own payoff $u_i(a)$, the outcome is directly correlational to the strategies chosen by others (12).

In classical game theory, rationality is captured by the "best response". A strategy $a_i^* \in A_i$ is considered a best response to the opponent's strategies a_{-i} if it maximizes the player's payoff given the opponent's actions, $a_i^* \in \arg\max_{a_i \in A_i} U_i(a_i, a_{-i})$ (12). This formulation indicates a static conception of rational adaptation: 'given what others are doing, I will do what is best for me'. Notably, however, this definition assumes a full knowledge of the actions and payoffs of others, which later learning models, such as Q-learning or fictitious play, will relax (42).

At Nash equilibrium, each player simultaneously plays the best response in accordance to everyone else. Formally, the strategy profile $a^* = (a'_1. . . , a_n)$ sets up a Nash equilibrium if for all players $i \in N$, $u_i(a'_i\ a)_{-i} \geq u_i(a_i,\ a^*_{-i}) \quad \forall a_i \in A_i$ (12). At equilibrium, no agent wants to unilaterally deviate. The system is self-consistent: expectations are aligned with beliefs, and no further skill or revision is required (1).

In many games, however, Nash equilibrium is not admitted in pure strategies, where each player can choose deterministic actions. To fix this, Nash implemented the idea of mixed strategies, where each player randomly chooses between their available actions. A mixed

strategy for player *i* is a probability distribution σ_i over A_i , where the set of all mixed strategies is denoted as $\Sigma = \Sigma_1 \times \ldots \times \Sigma_n$. The expected payoff is now $U_i(\sigma) = \mathbb{E}a \sim \sigma|u_i(a)$ (12). We need to remember that a mixed strategy Nash equilibrium is a profile $\sigma^* = (\sigma_1 \ldots \sigma_n)$ such that each player's mixed strategy maximizes their utility given the other's payoffs, $U_i(\sigma_i^*, \sigma_{-i}) \geq U_i(\sigma_i, \sigma_{-i}) \quad \forall \sigma_i \in \Sigma_i$ (12). The existence theorem establishing that every finite game has at least one mixed strategy equilibrium was transformative, as it made equilibrium analysis the basis of rational analysis of behavior (12).

Nash equilibrium can be seen as an end state of cognition, the place where all learning, experimentation, and uncertainty have disappeared (12). The players have common knowledge of rationality, exactly aligned expectations, and no further dynamics in the system remain. This static ideal represents what rationality "should" look like. However, and as will be discussed in later sections, this assumption that rationality should imply stasis presents a crucial blind spot (14). This is because players cannot always remain rational when constrained by computational, information, or time limits, consequently, their strategies may never stabilize into equilibrium (9).

3.2.2 *Fictitious play, Q-learning, No-regret learning, and evolutionary dynamics*

While classical game theory operates on the premise that strategies will remain fixed and outcomes static, learning models endow the analysis with time and the possibility of adaptation in strategies (1). Such models describe how agents, through repeated interactions, adjust their strategies in accordance with the results of the actions they have observed (19). In the continued research through the decades, from the behavioral formulations provided in the 1950's to the algorithmic models of the contemporary reinforcement learning literature, the nagging question nonetheless has not varied: Do these learning processes converge to Nash equilibrium? (21)

In the content that follows, we consider four canonical models of learning: fictitious play, Q-learning, no-regret learning and evolutionary dynamics, and the ways in which they formalise this question from complementary mathematical and conceptual points of view (43).

Fictitious play (FP) was introduced by Brown and was subsequently subjected to a rigorous analysis by Robinson (20). FP considers agents as Bayesian learners, whose empirical and possibly biased beliefs concerning the other agents' behavior lead to a plan based on such biases. It is assumed that each agent believes that others are playing stationary (but unknown) mixed strategies, and updates its beliefs in the light of observed frequencies of actions. Formally, if agent *i* observes the action taken by player *j* to be a_j^t at each round *t*, then its belief upon player *j*'s strategy after *T* rounds will be expressed by, $p_j^T(a_j) = \frac{1}{T}\sum_{t=1}^{T} 1\left[a_j^t = a_j\right]$ (12). At each iteration, player *i* best responds to this distribution, $a_i^{T+1} \in \underset{a_i \in A_i}{arg\,max}\, E\, a_{-i} \sim p_{-i}^T\left[u_i(a_i, a_{-i})\right]$ (12).

Fictitious play represents one of the earliest and most influential mathematical models of learning in games. In this framework, agents repeatedly form beliefs about their opponents' future actions based on observed past behavior and select best responses to those beliefs. Under suitable conditions—particularly in two-player zero-sum games and potential games—fictitious play converges to a Nash equilibrium. In more general classes of games, however, convergence is not guaranteed, and the dynamics may exhibit persistent oscillations or cycling behavior. These observations provided some of the earliest indications that equilibrium attainment depends critically on the structure of the game rather than arising automatically from repeated adaptation. Nevertheless, fictitious play remains a foundational model because it transformed equilibrium from a static solution concept into a dynamic process of strategic learning and adjustment (13).

Q-learning, interestingly, represents a shift to reward-based learning from belief-based learning. Agents learn through an action-value function: $Q(a_i,s)$ (instead of through an explicit model), which estimates the expected collective reward of taking action a_i in state s. These Q-values are updated through experience,

$$Q_{t+1}(s,a_i)=(1-\alpha)Q_t(s,a_i)+\alpha_t\left[r_t+\gamma \max_{a_i'} Q_t(s',a_i')\right],$$

where r_t is the reward received, γ is the discount factor, and α_t is the learning rate (42). In single-agent contexts, Q-learning is guaranteed to converge to the optimal value function under some conditions, while these guarantees do not hold in multi-agent contexts, where the environment itself is changing as opponents adapt. The resulting dynamics are often non-stationary and chaotic, and Q-values resort to infinite oscillation (44).

In repeated games, multi-agent Q-learning takes the classical equilibrium questions into the computational realm: can independent learners converge to a stable fixed point? In practice, convergence is found only in a restricted number of games (e.g. coordination games) while in competitive or mixed-motive games, agents display persistent non-equilibrium dynamics. Nevertheless, Q-learning continues to be a fundamental model because it most clearly represents bounded rationality (15): agents do not compute best responses, they learn them via interaction.

No-regret learning is another distinct paradigm related to Q-learning (33). Here, the goal is not to best respond to opponent agent behavior but to ensure that over time, the agent's performance is nearly as good as the best fixed strategy as set beforehand. The external regret after T rounds is defined (42) as,

$$R_i(T)=\max_{a_i\in A_i}\sum_{t=1}^{T} u_i(a_i,a_{-i}^t)-\sum_{t=1}^{T} u_i(a_i^t,a_{-i}^t)$$

An algorithm shows no regret if $R_i(T)/T \to 0$ as $T \to \infty$. Subject to all players using no-regret algorithms, the joint empirical distribution of play will converge to a set of coarse correlated equilibria, which are a generalization of a Nash equilibrium that allows for correlated randomness. This structure has a close connection to online learning and convex optimization, establishing ties between economic rationality and algorithmic efficiency (28). However, although the no-regret dynamics is mathematically attractive, it does not provide information about the temporal behavior of play: the play may oscillate or simply cycle forever, while the average play may converge asymptotically to equilibrium.

Evolutionary game theory models adaptation at the lower population level (45). The replicator dynamic describes how $x_i(a)$, the proportion of a population playing strategy a should evolve according to its relative payoff, $\dot{x}_i(a) = x_i(a)\left[u_i(a, x^{-i}) - \bar{u}_i(x)\right]$, where $\bar{u}_i(x)$ is the average payoff for the given population. Strategies that perform better than the averages of all strategies will increase in frequency while less successful strategies will decrease (43). In this formulation, the Nash equilibria correspond to stationary points, or states where strategies that remain active earn the same payoff (29).

Replicator dynamics demonstrates that equilibrium-seeking behavior can occur without cognition: evolution itself is performing distributed optimization (46). However, the dynamics also exhibits a rich array of non-equilibrium events in the form of cycles, limit cycles, and chaos, particularly in asymmetric or non-zero-sum games. These patterns correspond to instability in multi-agent reinforcement learning, suggesting deep relations between biological adaptation and computational learning (34).

Each of these models depicts an alternative way in which rationality could arise through adaptation, but they all measure progress toward that rationality by the same metric: distance from the Nash equilibrium (21). As subseqent sections will demonstrate, this constrains our attention to equilibrium and has caused us to miss an important insight: non-equilibrium behavior - oscillation, cycling, instability - may be its own structured and informative representation of bounded rationality (47).

3.2.3 *Convergence metrics and deviation measures*

Having reviewed the principal models of learning in repeated games, we now turn to a fundamental evaluative question: how do researchers determine whether a learning process has succeeded? In either classical or more modern frameworks, the consistent answer has been convergence. In this section, we will discuss the dominant paradigms of evaluation - i.e., metrics of convergence and deviation from equilibrium - and the implications of their construction on interpreting learning behavior (33).

Generally, the most direct way to access learning is to determine whether strategy profiles stabilize. A learning process produces a series of mixed strategies $\{x_t\}_{t=1}^{T}$, and convergence is said to occur if

$lim_{t\to\infty} ||x_{t+1} - x_t|| = 0$ (29). In empirical studies, it is not unusual to use either L_2-norm or total variation distance between the previous and current strategy distribution as a measure of convergence. In reinforcement learning, the similar criterion is for Q-values, value function and/or policy parameters to converge (48). Convergence is typically declared when changes in the relevant learning variables become negligibly small, with successive updates remaining below a specified threshold.

It is important to recognize that these approaches rest on a significant underlying assumption: stability is taken as evidence of successful learning and rational adaptation, whereas persistent instability is often interpreted as either a pathological feature of the model or a consequence of inadequate algorithmic design (1). However, as we will argue later, instability could be alternatively viewed as adaptive exploration and/or equilibrium multiplicity which are characteristics of bounded rationality (39).

In this paradigm, there are two primary measures used as competition metrics: regret and exploitability. Using the no-regret learning framework, the performance of a strategy sequence $\{a_i^t\}$ is compared to the best strategy in hindsight (33). A learner characterized by sublinear cumulative regret has a relation of "rationality" to the equilibrium benchmark because, on average, the learner could not have performed significantly better by playing an arbitrary fixed action. This leads to convergence anywhere in the broad set of coarse correlated equilibria (CCE), not necessarily a Nash equilibrium (27). Formally, the average deviation from equilibrium can be represented by,

$$\epsilon_T = max_{a_i \in A_i} \frac{1}{T} \sum_{t=1}^{T} \left[u_i(a_i, a_{-i}^t) - u_i(a_i^t, a_{-i}^t) \right],$$

where $\varepsilon_T \to 0$ implies approximate equilibrium play (27). Similarly, in the domain of multi-agent reinforcement learning and algorithmic game theory, exploitability - the amount a unilateral deviator could profit - serves as a practical measure of convergence (5). Algorithms such as Counterfactual Regret Minimization (CFR) and Nash-Q minimize exploitability explicitly to demonstrate success.

Another class of evaluation measures considers the statistical properties of the sequences that are played, rather than convergence pointwise (31). Researchers measure variance in trajectories of payoffs, autocorrelation between successive actions, or entropy of the joint action distribution. In replicator dynamics and other continuous-time models, whether equilibria are locally stable is determined by Lyapunov functions or potential functions. A reduction in these measures signifies progress towards equilibrium, while variability means that there is inefficiency or instability (29).

The prevalence of convergence indicators represents not a necessity of regression to the empirical, but of conceptual continuity. Game theory, which has emerged from economics,

was originally built around a predictive ideal called equilibrium (12). Behaviors that oscillate, bifurcate, or express chaos are often excluded from discussion, not because these styles of behavior are "not structured," but rather because their behavior cannot be effectively summarized by given evaluative metrics (45). Convergence-based evaluative metrics have afforded mathematical certainty and some experimental clarity, while concurrently undermining those clear assumptions conceptually (49). In fact, they have created a disrupted definition of learning closely targeted at the removal, or non-presence, of variability, rather than understanding variability. The next sections will argue that non-equilibrium, instead of being a nuisance analytically, is still an analytical, yet information-rich outcome of bounded rational adaptive actions (47).

3.3.1 *Evidence of non-convergence*

While the theories suggest eventual convergence, there is not a strong correspondence between empirical/ computational evidence and theory. Across behavioral experiments, market simulations, and multi-agent reinforcement learning systems, adaptive agents can often be observed exhibiting oscillatory, chaotic, or metastable behavior that does not become stable equilibrium (17). This section will review evidence-focusing on studies involving human learning (e.g., matching pennies and market prediction games) and in artificial systems (e.g., self-play agents reorganizing endlessly in suboptimal loops). The aim here is to establish a tension; that under conditions when agents should be converging to an equilibrium, the evidence suggests that learning systems are often non-converging and dynamic. The section will not offer explanations; rather it will establish the empirical puzzle that will motivate the theoretical re-evaluation that will follow (49).

3.3.2 *Studies proving metastable,chaotic, cyclic behaviors in repeated games*

Despite equilibrium being the standard reference point in game theory, a considerable amount of theoretical, computational, and empirical research shows that learning dynamics often do not converge to Nash equilibria. The first and clearest counterexample appears in the fictitious play literature, where a simple two-player 3 × 3 game (a non-zero-sum Rock–Paper–Scissors) in which fictitious play continues indefinitely, fails to converge to a stationary mixed strategy (50).

Reinforcement-learning algorithms - including Q-learning and related temporal-difference methods - have shown the same behavior in computational experiments, where in some cases the behavior was more chaotic. Sato, Akiyama, and Farmer (51) studied learning in Rock–Paper–Scissors and similar simple two-player games, showing that reinforcement learning can yield Hamiltonian or chaotic trajectories in the learning path: depending on parameters and initial conditions, it could evolve to convergence, to stability in cycles, or be sensitive and chaotic (34). This shows that lack of convergence is not just an artifact of belief-based models, but arises in reward-based adaptation as well. Sato and Crutchfield (52) built on this insight by deriving coupled

replicator equations as a dynamical-systems reduction of multi-agent reinforcement learning, thereby directly linking individual learning rules to population-level continuous dynamics (29).

Subsequent research further advanced these observations, focusing on more general learning rules and larger systems. Galla and Farmer (53) and related works showed that the vast class of adaptive learning rules (e.g., experience-weighted attraction / EWA) exhibit a wide variety of attractors - fixed points, limit cycles, and chaos - even in the case of modestly sized games, and the qualitative aspects are sensitive to learning parameters and payoff structure (17). Using statistical-physics-type techniques, Sanders, Farmer, and Galla (54) provided evidence that as games get more complicated (many players, many strategies), the range of parameters that support stable fixed points diminishes (39).

More recently, a series of impossibility and "no-go" results has reshaped the theoretical landscape of learning in games. Milionis et al. (55) proved an impossibility theorem showing that there exist classes of games for which no deterministic learning dynamics can simultaneously satisfy a set of desirable properties while guaranteeing convergence to a Nash equilibrium from every initial condition. Related work by Milionis, Papadimitriou, Piliouras, and collaborators further demonstrates that, for a positive-measure subset of games, no learning dynamics can reliably attain an ε-Nash equilibrium for arbitrarily small values of ε (49).

Alongside these theoretical investigations and computational fractions, regret-based and experimental literatures indicate that average or asymptotic convergence does not imply that temporal trajectories are fast or slow, or that they do not remain highly non-stationary. Hart & Mas-Colell (56)'s regret-matching dynamics guarantee that the empirical distributions of this dynamics converge to correlated equilibrium concepts, but they do leave the path of convergence open: players' instantaneous strategies can still oscillate or vary while time-averages converge (22,27). Regret-like oscillation or adaptation-typical behaviour are routinely observed in human subject experimental studies (matching pennies, auction and coordination-type experiments).

On a general level, these studies demonstrate a cross-methodological regularity: non-convergence is common, often structured (e.g., systematically hopping back in the direction of the previous strategy, making behaviour correlated with population dynamics) and in a number of cases, dependent on the parameters of the model (8). Cases include belief updating leading to fictitious play, reward-driven adaptation leading to Q-learning and adjoining coupled replicator dynamics, population eschewing leading to replicator equations, and learning via imitation for high-dimensional strategies leading to replicator dynamics. For our purposes in this review, these results highlight a re-framing: non-convergence must not be simply regarded as a failure of a learning rule, but as an informative signature of bounded, context-sensitive adaptation; this is precisely what we can capture and quantify

with our dynamical systems and information-theoretic tools in the coming sections.

3.3.3 *Prior experimental evidence from human learning and market dynamics*

Experimental economics and behavioral game theory have long sought to ascertain whether human learners, when repeatedly faced with strategic situations, systematically converge to equilibrium play or deviate from equilibrium in systematic ways (6). The collected evidence across several decades--from controlled laboratory studies to market level experiments--reveals that human agents exhibit, on average over time, persistent cycles, drifts, and metastability (8); similar to what was indicated beforehand.

The original laboratory work on repeated 2 × 2 zero-sum games such as Matching Pennies, provided some of the first empirical systematic deviations from Nash equilibrium (40). Although theory predicts that players should randomize over two strategies at equilibrium, empirical studies have shown that human choice behavior often exhibits persistent and structured oscillations even after extensive repeated interaction and experience (40). More direct evidence emerged from eye-tracking and response-time studies that sought to investigate the cognitive processes underlying strategic decision-making. By examining where participants directed their attention, how long they deliberated, and the sequence of information they considered, researchers were able to infer aspects of the reasoning processes guiding their choices. These studies suggested that individuals typically employ limited-depth reasoning and adapt their behavior in response to the recent history of play rather than computing fully rational best responses. Human decision-making therefore appeared to be adaptive, reactive, and cognitively constrained, rather than the product of exhaustive strategic optimization.

Evidence from coordination and market-entry games further suggests that human learning often exhibits path dependence and bounded adaptation rather than full convergence to equilibrium. A prominent example is provided by the “beauty contest” or “guess two-thirds of the average” experiments (57). In these studies, participants rarely engage in the infinite sequence of recursive reasoning required to reach the Nash equilibrium. Instead, they appear to anchor their decisions at a finite depth of strategic reasoning, resulting in systematic departures from the equilibrium prediction. As a consequence, aggregate behavior frequently converges to intermediate, metastable outcomes that reflect the cognitive limitations and shared expectations of the participants rather than the theoretical equilibrium itself. For example, repeated play often produces average responses that stabilize around values such as 30 or 40 rather than converging to the equilibrium value of zero. These findings suggest that learning trajectories can remain strongly influenced by initial beliefs, reasoning depth, and historical patterns of play, even after extensive experience.

Evidence from market environments provides a particularly compelling illustration of persistent instability outside highly controlled laboratory settings. Experimental studies of two-sided double auctions and asset markets have

repeatedly documented the emergence of price bubbles, crashes, and prolonged departures from equilibrium, even when participants are provided with substantial information about underlying asset values (45). Rather than converging smoothly toward equilibrium prices, traders often exhibit overshooting, momentum-driven behavior, and delayed adjustment to changing conditions. Expectations can become self-reinforcing: rising prices encourage additional buying, which further elevates prices, while falling prices can trigger the opposite dynamic. As a result, collective behavior may generate persistent fluctuations and cyclical patterns that cannot be fully explained by equilibrium-based models of rational adjustment.

High-frequency trading experiments and agent-based financial market simulations (Hommes, 2013; Galla & Farmer, 2018) also demonstrate that instability into these micro markets does not result from noise or irrationality - instead, it results from feedback sensitivity in adaptive expectation models (35). As traders increase their learning rates or weigh recent information more heavily, markets will undergo a bifurcation transition out of stable equilibria into oscillations, and eventually into fully chaotic regimes (17). Support for the empirical trajectory described above aligns well with the theoretical set of bifurcations displayed by Q-learning and EWA dynamics. Ultimately, bounded rationality + high adaptivity = instability.

A consistent pattern emerges from the studies reviewed above: real learners, whether human or algorithmic, rarely settle permanently into static equilibrium behavior. Instead, they continue to balance exploration of new strategies with exploitation of previously successful ones, generating persistent adaptation even after extended periods of interaction. Recent studies in behavioral reinforcement learning suggest that subjects often maintain stochastic patterns of play over hundreds of rounds rather than converging to deterministic strategies (58). In this context, what is frequently characterized as "noise" may serve an adaptive function, preventing agents from becoming locked into potentially suboptimal behaviors and enabling continued responsiveness to changing or non-stationary opponents (7).

3.3.4 *Machine learning oscillation and divergence in multi-agent RL*

As reinforcement learning (RL) transitioned from single-agent environments to multi-agent systems, practitioners quickly learned that the theoretical guarantees and empirical mannerisms learned in single-agent learning would not hold (16). In single-agent settings, the environment is (typically) stationary, and the agent's updates asymptotically converge to an optimal policy; in multi-agent reinforcement learning (MARL), the environment is endogenously non-stationary because every other agent is changing its policy simultaneously (46). This process of coupled adaptation creates feedback loops that produce oscillation, cycling, and divergence for both agents simultaneously - behaviors that ML researchers have observed in games, simulated markets, and complex competitive tasks.

A common empirical observation is policy

wobble, or limit-cycle behavior, in otherwise simple competitive games. Simply put, independent policy-gradient or Q-learning agents training against each other in zero-sum settings often do not converge; they cycle around mixed-strategy equilibria or orbit quasi-periodically rather than converge (59). These oscillations are statistically obvious from both the extraordinarily observable policy simplex plots, as well as observable in the rise of variance of returns and non-decaying norms for policy distance. In practice, this can be interpreted as agents that will ‘exploit’ each other for some period of time before they are, in-turn, ‘exploited’.

The ML community has provided theoretical explanations for this phenomenon by finding rotational components in gradient dynamics and showing that typical optimization techniques (e.g. simultaneous gradient descent/ascent) produce non-conservative flows (30). In the context of differentiable game formulations, the gradient field can display significant curl components that provide rotational motion instead of descending towards a minimum; i.e., this is a structural source of cycling in learning (48). Studies of adversarial training (especially in the GAN literature) have illustrated how saddle-like objectives and opposing directions of optimization can lead to oscillatory or even chaotic training trajectories, leading to intervention to stabilize the training in practice (59).

The MARL literature has described even more forms of instability than simple oscillation (4). Agents may become stuck in metastable regimes-long stretches of near-stationary actions followed by a sudden shift in the regime-which makes potential evaluation and deployment more difficult. There are also fragile equilibria where imperceptible perturbations in initialization or learning rates can cause the long-run outcome to switch from convergence to being chaotic; hyperparameter sensitivity is therefore a practical indicator of multi-agent instability (40). Catastrophic forgetting and cyclic overfitting to the currently popular behaviors of opposing agents are further instabilities: an agent that learns quickly will exploit a currently weak opposing agent policy while subsequently forgetting how to respond when the opposing agent has also learned how to adapt (5).

In response to these challenges, the machine learning (ML) community has offered various mitigation strategies that illustrate the depth of the instability problem (4). In terms of architecture, the stabilization of agent updates using extra information during learning is possible with centralized training and decentralized execution (60); in terms of algorithms, fictitious self-play, population-based training, and opponent-aware learning (e.g., second-order or meta-gradient that accounts for an opponent's learn rate) reduce non-stationarity to steer learning dynamics towards more stable attractors (5). However such methods can trade one form of instability for another (e.g., they can create a scale of regime shift, or have to have the correct population design), reiterating that instability is systemic, and not merely a byproduct of naive algorithms (49).

Together, the MARL literature provides three salient points for our review. First, instability in learning is commonplace in realistic, interactive ML scenarios - not rarefied or aberrant in nature, and not restricted to pathological toy games (23). Second, the nature of the instability is structured: rotational gradient flows, parameter sensitivity, metastability, and cyclic exploitation are repeated phenomena, with interpretable explanations of their causes (47). Third, engineering fixes (centralization, modeling opponents, diversity in populations) reduce but do not eliminate instability, because they address only the non-stationarity symptoms of instability without redefining the meaning of instability. Overall, these observations reinforce the paper's central argument: instead of assuming divergence is a failure, we should examine non-convergent MARL trajectory in a quantitatively based and theoretically motivated systematic analysis of all the potential information that such trajectories contain (2).

3.4.1 *Theoretical gap: rationality and its boundaries*

The persistence of non-convergence suggests that the models developed thus far may rely on assumptions that are overly idealized to adequately capture real learning processes. Most notably, the model that is utilized most prevalently relies on unbounded rationality, which suggests that agents engage in some type of perfect best response with complete information, optimized without any cost. This section engages with this assumption, tracing the idea of bounded rationality from Herbert Simon’s (61) initial insights to contemporary developments in computational and behavioral science (15). By recognizing that learning occurs under cognitive, informational, and temporal constraints, we can reinterpret non-convergence not as a deficiency to be corrected, but as a natural consequence of rational adaptation in bounded environments (9).

3.4.2 *The limits of perfect foresight*

Classical game theory is built on two assumptions that jointly guarantee consistency in equilibrium concepts: agents are perfectly rational, and agents have unlimited computation time and foresight (12). These assumptions have been shown to be mathematically elegant, though increasingly strained by the behavioral and algorithmic literature discussed earlier in the paper. They provide an attractive formal framework for existence and stability proofs for equilibria, but they also specify an imaginary universe that no human being (or agent with modern computational abilities) lives in (14).

The first assumption - unbounded computation assumes that every agent evaluates all possible states, strategies, and value outcomes before taking action (12). For a typical finite game, this assumption means that an agent can compute best responses instantaneously over an exponentially large strategy space and, by extension, anticipate the complete iterative reasoning process of each opponent (57). This abstraction seems to enable closed-form equilibria and comparative statics. However, it does so at the expense of obscuring the algorithmic limitations underlying decision processes. When actual learners are

implemented as algorithms; i.e., a Q-learner, or with human subjects, there are finite limits to memory and attentional constraints, as every update takes time and draws from available information (62). Agents have to approximate their state value functions, compress their state space, and employ heuristics (41). These cognitive or computational bottlenecks fundamentally shift the dynamics of adaptation and what might be perceived as "noise" in the world of classic game-theory may really reflect the cost of computation.

The second assumption, perfect foresight, posits that agents not only understand the entire structure of the game but also have perfect foresight over other agents' reasoning steps and future actions (12). Perfect foresight makes equilibrium analysis tractable from a perspective of omniscience, but it systematically neglects the feedback that constitutes learning in the real world (19). Perfect foresight implies that agents never need to explore, misestimate, or reconsider their expectations, which contradicts both actual human behavior and the learning patterns observed by an adaptive multi-agent reinforcement learner. Perfect foresight in reality would require all players to operate under uncertainty about each other's internal models, learning rates, and utility (objective functions) (63). Players may utilize partial histories to run a multi-step probabilistic forecast, but they would then not be acting from any perspective of omniscience. This bounded character of players produces the oscillations, path dependencies, and metastable states.

Classical equilibrium analysis therefore describes an idealized end state rather than one that agents necessarily attain in practice (14). Equilibrium reasoning assumes that learning and adaptation have effectively ceased, that agents have formed mutually consistent expectations, and that no further strategic adjustment can improve outcomes. Yet both empirical observations and computational complexity results suggest that equilibrium attainment may be difficult, slow, or even infeasible in many games. If the informational and computational demands required for equilibrium reasoning exceed the capabilities of actual agents—whether human or artificial—then persistent adaptation and non-convergent behavior may arise not as failures of rationality, but as natural consequences of rational decision-making under constraint (33).

3.4.3 *The concept of bounded rationality*

Bounded rationality, a concept coined by Herbert A. Simon in the mid-20th century, refers to a critique of the 'Olympian rationality' presumed to characterize classical economics and game theory (14). Simon argued that the archetype of the perfectly rational agent—capable of unlimited computation, complete information processing, and flawless memory—is a purely theoretical construct with little resemblance to how humans or organizations actually make decisions. Rather than taking actions to maximize a global objective or utility function over an infinite space of potential options, agents face constraints related to the capacity of information, time, and computation and must, therefore, rely upon simpler decision rules to make 'good enough' choices instead of 'best' choices (37).

Simon's theory of decision-making introduced the concept of *satisficing*, a behavioral rule under which decision-makers establish aspiration levels and select the first option that meets or exceeds those thresholds rather than searching indefinitely for an optimal solution (37). In doing so, Simon recast rationality as a process constrained by finite resources rather than as an outcome requiring perfect optimization. Agents were viewed neither as flawless maximizers nor as irrational actors, but as adaptive decision-makers navigating complex environments through iterative feedback, learning, and adjustment within limited cognitive and temporal horizons (14). Crucially, bounded rationality was not equivalent to irrationality; rather, it represented rational behavior operating within the practical constraints of perception, memory, information, and computation (39).

In the decades that followed Simon's work, bounded rationality transitioned from a psychological observation into a formal modeling framework across economics, cognitive science, and computer science. Initial models expanded the bounded rationality concept to include procedural rationality, emphasizing not only the choices made, but the algorithmic process by which decisions are made. This avenue of research naturally coincided with the emergence of artificial intelligence and machine learning as decision rules are implemented as computational procedures. Models of limited information processing (37) constrained formal decisions by limiting how much information an agent could acquire or compute in each decision processing cycle, laying the foundation for modern information-theoretic and algorithmic models of bounded rationality (20).

Subsequent refinements delineated even more the different types of bounds shaping decision making. Cognitive bounds reflect the limits of human memory and attention, which constrain the complexity of sustainable mental models an agent can use in complex environments (62). Temporal bounds acknowledge that decisions must often be made under time constraints, therefore limiting complete optimization of decisions (14). Computational bounds, which are perhaps more directly applicable to learning dynamics, acknowledge the algorithmic intractability of reasoning and decoding equilibria in high-dimensional or multi-agent decisions where equilibrium nature makes reasoning based on past experiences extremely difficult (unless the problem can afford computational resources that reduce information processing demands). Together, these three types of bounds delineate a feasible decision region in which agents search for satisfactory strategies rather than explicit optimization (often through adaptive or heuristic decision making) (41).

The notion of bounded rationality is one important place where computational learning theory and behavioral game theory converge (18). In modern formulations, bounded rationality is seen, not as a deficiency, but as a fundamental feature of adaptive systems (35). For instance, reinforcement learners instantiate bounded rationality through their learning rates, exploration parameters, and finite memory buffers that limit how they approximate optimal policies. In evolutionary

and replicator dynamics, stochastic choice rules and bounded sampling of populations characterize bounded rationality in similar ways (29).

In sum, bounded rationality provides both the conceptual and mathematical framework for understanding why learning systems—whether biological, economic, or artificial—may not converge to the equilibrium outcomes predicted by classical theories of rationality. By emphasizing adaptive search under conditions of limited information, finite computation, and imperfect foresight, Simon transformed rationality from a static ideal into a dynamic process of continual adjustment and approximation (41).

3.4.4 *The Ad Hoc implementation of bounded rationality*

Despite its central role in economics, psychology, and artificial intelligence, bounded rationality is often acknowledged conceptually yet incorporated into formal models in a largely ad hoc manner (39). Scholars will often acknowledge that agents are unable to compute perfect equilibria or foresee infinite future states; however, they do not formally model, or describe, these constraints and instead introduce terms that describe noise, random perturbation (44), or quite simply heuristics, as a "proxy" to describe cognition and computational constraint (62).

The most common version of this phenomenon manifests in the addition of stochasticity to often deterministic learning dynamics (40). In both reinforcement learning and evolutionary game theory, researchers use bounded rationality as a way to "soften" the best-response function and represent the bounded rationality agents as being stochastic - that is they choose actions probabilistically, taking greater rewards/actions with greater frequency, but not deterministically. A representation of this is formalized in terms of Logit (or Boltzman) choice rules, where the inverse temperature, β, parameterizes the trade-off between exploration and exploitation, wherein higher β approximates perfect rationality (greedy selection in an approximate sense), while lower β allows for stochastic deviations from optimal policy (48).

Despite this modeling of uncertainty in decision-making, there is not an attribution to the structural origins of bounded rationality, principally that the agents could be perfect rational agents but are simply ‘noisy’ (39). Furthermore, the conceptual jump between noisy and cognitive is left unstated.

Similarly, in Q-learning and in no-regret frameworks, bounded rationality is incorporated through learning rates (α) or exploration probabilities (ε) in ε-greedy policies (42). These hyperparameters control the rate at which agents revise their beliefs or depart from exploitative play. While they clearly affect convergence and stability, they are usually chosen empirically, rather than as being derived from theories relating to limits to information processing or computational cost (62). Consequently the same α or ε can be taken as cognitive hesitation, or memory decay, or random search, all of which are reasonable, but none of which are linked formally to a theory of bounded rational decision making.

In evolutionary and replicator models meanwhile, bounded rationality usually arises in the form of a mutation or experimentation term, small probabilities that agents change strategies independently of payoffs received (63). These trembles or noise-floors are supposed to prevent a population being trapped in a sub-optimal equilibrium and enable diversity to be preserved in strategy space. Yet once again it is the stochastic features that are introduced for stability or realism's sake, devoid of any connection with cognitive, algorithmic or other restrictions on information processing. The randomness can, of course, be taken as an approximate indicator of ignorance or misperception, but does not elucidate one or other of the questions of why such deviations should have occurred, or how closely they relate to limitations in the depth of reasoning.

Even more sophisticated modern approaches, such as Experience-Weighted Attraction (EWA) learning and population-based MARL models, retain this basic character (7). EWA models, for example, assume memory decay parameters (φ) and coefficients governing intensity of choice (β) in such a manner that bounded memory and inaccurate decision processes can be approximated. Indeed, these variables generate the empirical learning curves which appear in human experiments but still remain phenomenological (8). These parameters fit data rather than constituting a theory of bounded rationality as a structural characteristic of computation under constraint.

Yet the ad hoc treatment of bounded rationality raises a broader methodological concern: the distinction between modeling uncertainty and modeling constraint (39). While stochasticity, noise, and probabilistic choice rules can often reproduce behavioral patterns associated with bounded rationality, they do not necessarily capture the underlying mechanisms that give rise to those patterns. A principled account of bounded rationality should emerge from explicit limitations on cognition, information, computation, time, or recursive reasoning depth (9), rather than from arbitrary perturbations imposed upon otherwise fully rational models. Without such foundations, randomness risks serving as a convenient proxy for boundedness rather than a genuine representation of its causes.

This distinction is crucial. Many contemporary models that invoke bounded rationality do not fully distinguish between genuine cognitive constraints and externally imposed stochasticity. As a result, departures from equilibrium are often treated as calibration artifacts, noise, or residual error rather than as emergent consequences of constrained adaptation (49). In doing so, these models risk reintroducing the very assumption they seek to relax: that rationality is fundamentally perfect except when perturbed. The challenge—and the opportunity—lies in moving from ad hoc representations of boundedness to principled ones, in which limitations on information, computation, memory, and reasoning are treated as intrinsic components of the learning dynamics themselves (34). Under such a framework, instability ceases to be a nuisance parameter or modeling residual and instead becomes a measurable and interpretable manifestation of how boundedly rational systems adapt to their environments (47).

3.5.1 *Instability as information: formal definitions and framework*

With the boundaries of rationality now clarified, we may reconsider the nature of instability. Rather than viewing instability as failure, or as noise, we now explore the hypothesis that instability itself may encode structure; patterns in which bounded agents adapt, explore, and respond to uncertainty (47). Using notions from the theory of information and from dynamical systems, we introduce the concept of "structured instability" and indicate quantitative measures (entropy rate, Lyapunov exponents, mutual information) for measuring how the learning process evolves over time. This section delineates the theoretical pivot of the paper, from why convergence fails to what instability tells us (36).

3.5.2 *Defining "structured instability"*

Definition 1 *(Structured Instability / Structured Chaos).* A learning dynamic x_t in a repeated game G exhibits *structured instability* if and only if the following five conditions hold simultaneously:

(i) Non-convergence: $\liminf_{t\to\infty}|x_{t+1} - x_t| > 0$. The trajectory does not settle to a fixed point or periodic orbit with period 1.

(ii) Boundedness: There exists a compact set K contained in Δ^n (the probability simplex) such that x_t is in K for all $t \geq t_0$, for some finite t_0. The trajectory remains confined to a bounded region of strategy space.

(iii) Positive finite complexity: $0 < h_mu < H_max$, where h_mu is the entropy rate of the discretized trajectory, and $H_max = \log_2(|A|)$ is the maximum entropy over the action set A. This distinguishes structured instability from both pure noise ($h_mu \sim H_max$) and deterministic periodic orbits (h_mu approximately 0).

(iv) Sensitivity: The maximal Lyapunov exponent satisfies $\lambda_{max} > 0$ (but bounded). Small perturbations in initial conditions lead to exponential divergence of nearby trajectories, yet the divergence remains bounded by condition (ii).

(v) Inter-agent coupling: The time-averaged mutual information $I(X;Y) > 0$ between the strategy sequences of different agents. The agents' trajectories are statistically dependent, reflecting genuine co-adaptive behavior rather than independent random walks.

Conditions (i) and (ii) distinguish structured instability from both convergence, which violates condition (i), and unbounded divergence, which violates condition (ii). Condition (iii) separates structured instability from both pure noise and trivial periodic behavior, while conditions (iv) and (v) ensure that the dynamics exhibit sensitive dependence on initial conditions together with meaningful inter-agent coupling. Collectively, these criteria identify learning dynamics that remain bounded, adaptive, and informationally structured despite the absence of equilibrium convergence.

The transfer experiments suggest that exploration may function as a persistent adaptive resource rather than merely a transient feature of early learning. In adversarial environments, broad policy support appears to be retained and transferred across tasks, influencing subsequent adaptation even after the original learning context has changed. This contrasts with convergent environments, where exploratory diversity progressively contracts as

equilibrium is approached.

More generally, the transfer results are consistent with the hypothesis that structured instability reflects the preservation of organizational structure within policy space rather than unrestricted stochastic exploration. The observed decline in entropy therefore does not necessarily imply dissipative convergence toward a fixed point. Instead, it may indicate increasing organization within a constrained region of policy space, allowing agents to remain adaptive and responsive while avoiding complete collapse into static behavior. Further geometric analyses would be required to determine whether such organization corresponds to low-dimensional manifolds, rotational dynamics, or other forms of structured adaptive behavior.

3.5.3 *Quantifying structured instability*

If structured instability is viewed as patterned adaptation rather than random fluctuation, it should be measurable in systematic ways (47). Traditional convergence measures (i.e. distance to Nash equilibrium, regret minimization) cannot capture the richness of ongoing, cyclical, or metastable dynamics, but we can borrow insight from information theory and nonlinear dynamical systems to analyze structured instability more rigorously, as these active areas of study are intentionally structured around order in apparent disorder (36).

The Lyapunov exponent, denoted with a γ, is a tool from game and chaos theory that quantifies how small differences in initial conditions evolve over time (13). When $\gamma < 0$, $\gamma = 0$ and $\gamma > 0$, trajectories converge to a stable fixed point, the system exhibits neutral stability and the system displays signs of chaos respectively. A positive, but bounded, Lyapunov exponent in the context of a repeated game or multi-agent learning indicates a sensitivity to initial conditions, yet is self-constraining (40). In the short term, there is unpredictability in agents' behaviors, yet they are changing in bounded regions of the strategy space - characteristic of structured instability.

For example, various studies analyzing replicator dynamics in rock–paper–scissors and multi-agent Q-learning found consistent positive Lyapunov spectra indicating that the learning paths were chaotic, yet bounded - representing a “structured” or non-random attractor (29). While Lyapunov exponents measure sensitivity, the entropy rate h_μ quantifies the information generation of a process (47). In other words, it quantifies how unpredictable the next state is given the past. Low, high and intermediate entropy rates respectively imply repetitive, predictable behavior; random or disorderly activity and structured instability, where the behavior is partially predictable. In a repeated game, an intermediate entropy rate indicates a system that is still learning - producing new patterns without collapsing into complete noise. This balance between predictability, yet innovation, can be interpreted as an optimal cognitive regime for bounded rationality flags, maximizing potential to adapt while preventing chaos and inefficiency.

In practice, it may be possible to estimate the entropy rate of agent strategy distributions

through time-series compression, or transition probability matrices, to depict the dynamic richness of their trajectories (26).

Structured instability is also not just about individual adaptation, but also reflects mutual feedback among agents. The mutual information $I(X;Y)$ between agents' strategies signifies how much one agent's behavior informs or predicts another's. If $I(X;Y) \approx 0$, agents act independently. On the other hand, if $I(X;Y)$ is high, agents are locked in rigid patterns. This approach enables us to differentiate between structured instability (highly coordinated with variability) and pure randomness (no coordination) or convergence (stable coordination) (32). Consequently, in practice, the calculation of time-varying mutual information across strategy updates can, in fact, provide glimpses of adaptive synchronies, meaning agents coordinated together temporarily, before dis-coordinating again - indicative of a structurally unstable system (2).

These metrics—Lyapunov exponents (sensitivity), entropy rate (complexity), and mutual information (coordination)—provide a quantitative framework for characterizing instability in repeated games. Rather than treating all departures from equilibrium as evidence of failure, they allow non-convergent behavior to be analyzed in terms of its dynamical and informational properties. Through such measures, instability can be differentiated into distinct regimes, including chaotic dynamics, metastable behavior, periodic structure, and varying degrees of informational coupling between agents (47).

More broadly, these metrics suggest the possibility of developing a standardized framework for evaluating adaptive behavior beyond equilibrium attainment. Future work could explore combining dynamical and information-theoretic measures into a composite index of adaptive complexity, flexibility, or bounded-rational performance. Under such an approach, bounded rationality would no longer be viewed solely as a limitation on optimization, but as a measurable property of how agents process information, adapt to constraints, and respond to changing environments. Such a perspective may help bridge behavioral game theory, machine learning, and dynamical systems science within a common quantitative framework (18).

3.5.4 *Measuring learning behavior over time*

The utilization of dynamical and information-theoretic metrics in repeated games facilitates the examination of learning as an evolving process, rather than as an end-state (64). Each of the metrics discussed previously provides a different temporal dimension of adaptation - how agents engage in change, coordination, and complexity over time. When assessed over time, we can now understand not if learning converges, rather how learning unfolds, thereby identifying the transitions between exploration, coordination, and stabilization that are features of boundedly rational behavior (47).

The Time-Resolved Lyapunov Exponent (TLE), when treated as a function of time rather than as a single asymptotic quantity, reveals how sensitivity evolves throughout the learning process. Early in training, relatively large positive Lyapunov exponents often

indicate heightened sensitivity to strategic perturbations, reflecting active exploration and adaptation in the absence of stable expectations. As learning proceeds, sensitivity typically decreases but remains positive, suggesting a regime of bounded responsiveness in which agents continue to react to one another while operating within increasingly organized regions of strategy space (32). Consequently, the time-resolved Lyapunov spectrum provides a useful characterization of how adaptive systems evolve from highly exploratory dynamics toward more structured, yet still non-convergent, patterns of behavior. Rather than signaling a transition from chaos to order, the temporal evolution of Lyapunov exponents captures changes in the degree and organization of adaptive sensitivity over time (65).

The entropy rate enhances this perspective by operationalizing behavioral significance over time (50). The entropy rate of the distribution of strategies over sliding windows is a means of understanding the system's persistent ability to produce novelty. High entropy in early learning reflects random search behavior; however, as the system becomes predictable, a decline in entropy is expected. Further, in situations of bounded rationality, entropy will not be eliminated - although it will stabilize at some intermediate constraint associated with predictable adaptation capacity (47). This phenomenon of intermediate entropy suggests that learning should be viewed not as convergence but, rather, as a self-adjusting process in which agents are developing regularities and maintaining flexibility and adaptive capacity to change (66).

Mutual information introduces an explicitly interaction-based perspective on learning dynamics (10). The mutual information $I_t(X;Y)$ between the strategy trajectories of two agents quantifies the extent to which knowledge of one agent's behavior reduces uncertainty about the behavior of the other. As learning progresses, mutual information often increases as agents become increasingly responsive to one another's actions and begin to form expectations about future behavior. In many settings, however, this dependence fluctuates over time as agents continually adapt to changing strategies and expectations (32). Such fluctuations reveal alternating periods of stronger and weaker strategic coupling, suggesting that co-adaptation is an ongoing and dynamic process rather than a monotonic progression toward equilibrium. Consequently, mutual information provides a quantitative measure of inter-agent responsiveness and coordination, offering insight into the informational structure underlying non-convergent learning dynamics and structured instability (47).

By jointly tracking sensitivity λ_t , complexity $h_\mu(t)$, and coordination $I_t(X;Y)$, we can characterize the evolution of adaptive flexibility in repeated games. These measures distinguish structured adaptation from both random fluctuations and simple convergence, allowing non-equilibrium behavior to be analyzed in terms of its informational and dynamical properties. In this framework, instability is no longer treated as an inconvenient deviation from equilibrium, but as a measurable expression of sustained adaptation under constraint (35).

This perspective also has practical implications. If structured instability reflects the preservation of adaptive flexibility, then measures of sensitivity, complexity, and coordination may help identify learning regimes that retain responsiveness to novel environments rather than collapsing into rigid equilibrium behavior. As demonstrated in the transfer experiments presented later, agents trained in adversarial settings frequently preserve broader exploratory support and adapt more effectively to new tasks than agents trained in convergent environments. These results suggest that metrics of structured instability may provide useful indicators of adaptive capacity, although establishing such a relationship directly remains an important direction for future work.

3.5.5 *Operationalizing structured instability*

Definition 2 *(Conditions for Boundedness).* The learning dynamics remain bounded if the following assumptions hold:

Assumption A1 (Bounded payoffs): There exists $M > 0$ such that $|u_i(a)| \leq M$ for all players i and for all action profiles a in A. This is satisfied by any finite game with finite payoffs, which includes all standard matrix games.

Assumption A2 (Simplex constraint): Strategy updates preserve the probability simplex: $x_i(t)$ is in $\Delta^{|Ai|}$ for all t, where $\Delta^k = p \in \mathbb{R}^k : p_j \geq 0$, sum $p_j = 1\}$. Both Q-learning with Boltzmann policies and replicator dynamics naturally preserve this constraint: Boltzmann softmax maps R^k to the interior of Δ^k, and the replicator equation preserves the simplex as an invariant set.

Assumption A3 (Decaying or bounded step-sizes): For Q-learning, the learning rate α_t is in (0, 1) and satisfies either: (a) the Robbins-Monro conditions sum $\alpha_t = \infty$ and sum $(\alpha_t)^2 = \infty$ (guaranteeing asymptotic convergence in single-agent settings), or (b) $\alpha_t = \alpha$ for a constant alpha in (0, 1), in which case Q-values remain bounded by $|Q(s,a)| \leq \frac{M}{1-\gamma}$ via the contraction argument, where gamma is the discount factor. In our simulations, we use $\alpha = 0.1$ and $\gamma = 0.95$, yielding $|Q| \leq 20M$.

Assumption A4 (Boltzmann exploration with tau > 0): The Boltzmann temperature parameter $\tau > 0$ ensures that policies are interior to the simplex, preventing boundary singularities. Specifically, for any Q-vector, the Boltzmann policy $\pi(a) = \frac{\exp(Q(a)/\tau)}{\sum_{a'} \exp(Q(a')/\tau)}$ satisfies $\pi(a) > 0$ for all a. If $\tau \to 0$, the policy concentrates on the greedy action, but boundedness still holds on the interior of Δ. In our simulations, tau starts at 1.0 and decays with rate 0.9998, remaining positive throughout.

Assumption A5 (Lipschitz dynamics): The replicator map is $F(x) = x_i[(Ax)_i - x^T Ax]$ Lipschitz continuous on the simplex Δ, since payoffs are bounded (A1) and x in Δ is compact. This ensures short-time existence and uniqueness of solutions by the Picard-Lindelof theorem, with the simplex as an invariant set. The stochastic perturbation term (noise_scale = 0.005 in our simulations) is bounded and does not break Lipschitz continuity.

Proposition 1 (Existence of Compact Attractor). *Under Assumptions A1-A5, the joint strategy trajectory (x_t, y_t) remains in $\Delta^{|A|} \times \Delta^{|B|}$ for all t, which is compact. By the Birkhoff theorem, the omega-limit set $\omega(x_t)$ is nonempty,*

compact, and invariant. When $\omega(x_t)$ is not a fixed point, it is either a limit cycle, quasi-periodic orbit, or a strange attractor, that is, structured instability in the sense of **Definition 1**.

Justification: Bounded payoffs (A1) bound the drift of Q-values and replicator fitness; simplex constraints (A2) confine trajectories to a compact set; step-size conditions (A3) prevent Q-value explosion; Boltzmann policies (A4) keep strategies in the interior of the simplex; Lipschitz continuity (A5) prevents finite-time blowup. Together these guarantee that non-convergent trajectories remain bounded, making structured instability the only alternative to convergence. The key insight is that the strategy simplex is compact and invariant under both Q-learning (via Boltzmann projection) and replicator dynamics (by construction), so the omega-limit set of any trajectory is necessarily nonempty, compact, and invariant. Non-convergence then implies the existence of a nontrivial attractor.

4. Results

4.1 Empirical demonstration: extracting structured chaos across game classes

4.1.1 *Simulation design and protocol*

We implement two canonical learning algorithms, multi-agent Q-learning with Boltzmann exploration and continuous-time replicator dynamics with small stochastic perturbation, across eight games spanning three structural classes. Each simulation runs for T = 10,000 iterations (sufficient to observe both transient and asymptotic behavior), with parameters chosen to reflect realistic bounded rationality: Q-learning uses $\alpha = 0.1$, $\gamma = 0.95$, initial temperature $\tau = 1.0$ with decay rate 0.9998; replicator dynamics use $dt = 0.01$ with noise scale 0.005. All simulations use the same random seed (23) for reproducibility.

For each game, we compute the following diagnostic suite: (1) strategy trajectory plots for both agents under both learning rules; (2) entropy time-series with sliding-window smoothing; (3) maximal Lyapunov exponent estimation via Rosenstein et al. (67) nearest-neighbor divergence method; (4) transfer entropy between agents in both directions; (5) KL divergence from uniform over time; (6) PCA of joint strategy evolution; (7) frequency spectra via FFT; (8) eigenvalue analysis of the Jacobian of the replicator dynamics at quasi-stationary points; (9) Poincare return maps; and (10) surrogate data significance testing via temporal shuffling with autocorrelation-based test statistics.

4.1.2 *Games analyzed*

4.1.2.1 Zero-sum games: (a) Rock-Paper-Scissors (RPS), the canonical cycling game producing limit cycles and heteroclinic orbits; (b) Shapley's 3 × 3 game, the classical example of persistent cycling in which fictitious play fails to converge; (c) Matching Pennies with perturbations, producing chaotic responses under Q-learning variants. Potential games: (d) Stag Hunt with exploration, a coordination game revealing partial convergence and oscillatory patterns near the risk-dominant equilibrium; (e) 3 × 3 Coordination Game, a multi-equilibrium setting where exploration noise determines equilibrium selection.

4.1.2.2 General-sum games: (f) Battle of the Sexes, an asymmetric coordination game; (g) Minority Game (3-action), a widely accepted example of structured chaos in adaptive market settings; (h) Prisoner's Dilemma, the classic social dilemma.

4.1.3 *Zero-Sum games: persistent cycling and structured chaos*

4.1.3.1 Rock-Paper-Scissors (RPS). Under Q-learning, RPS produces the prototypical structured instability signature. The strategy trajectories oscillate persistently across all three actions without convergence (Figure 1, top-left). The entropy of the Q-learning policy stabilizes at ~1.10 bits (out of a maximum of 1.58 bits for a 3-action game), confirming the intermediate entropy regime predicted by Definition 1. The maximal Lyapunov exponent is positive (λ = 0.031), indicating sensitive dependence, yet trajectories remain bounded within the simplex. Transfer entropy is non-zero in both directions ($TE(A\rightarrow B)$ = 0.009, $TE(B\rightarrow A)$ = 0.014 bits), confirming inter-agent coupling. The Jacobian eigenvalues at the interior fixed point are purely imaginary (~ ± 0.579i), confirming the conservative cycling structure. The surrogate significance test rejects the null hypothesis of temporal randomness with $p < 0.001$. Under replicator dynamics, the entropy stabilizes at 1.58 bits (near maximum) as the dynamics produce near-Hamiltonian orbits around the interior fixed point.

Entropy rate stabilization as 'order within chaos.' A key empirical finding is that even though Q-learning in RPS does not converge, the entropy rate stabilizes at ~1.26 bits/step (Figure 1, bottom-right). This means agents organize around a predictable cycling pattern despite stochasticity: the per-step surprise generated by the system reaches a steady state, indicating that the dynamics have settled into a structured attractor rather than random drift. This stabilization of the entropy rate is the empirical signature of what we term structured chaos.

4.1.3.2 Shapley's Game. Shapley's 3 × 3 game (figure 2) exhibits the strongest structured instability signature among all games tested. The maximal Lyapunov exponent under Q-learning is the highest observed (λ = 0.044), and transfer entropy is nearly symmetric between agents ($TE(A\rightarrow B)$ = 0.021, $TE(B\rightarrow A)$ = 0.022 bits), indicating balanced mutual adaptation. The entropy stabilizes at 0.92 bits, lower than RPS, reflecting a tighter cycling pattern. The PCA analysis reveals that 38% and 34% of variance are captured in the first two principal components, indicating a low-dimensional attractor structure. The Jacobian eigenvalues are real and negative at the measured point, but the global dynamics are chaotic, demonstrating that local linearization at a single point is insufficient to characterize the full dynamical behavior.

4.1.3.3 Matching Pennies (perturbed). The 2 × 2 Matching Pennies game produces particularly clear dynamical signatures (Figure 3). Under Q-learning, the strategy entropy is lower (0.45 bits out of max 1.0), reflecting the tendency for Boltzmann exploration to concentrate on the currently winning action. The Lyapunov exponent is positive (λ = 0.045). The Jacobian eigenvalues of the replicator dynamics are

purely imaginary, confirming the precisely Hamiltonian nature of the dynamics. This means the replicator dynamics in Matching Pennies produce exact limit cycles with zero dissipation, the theoretical ideal of structured instability. The frequency spectrum shows a sharp dominant peak, indicating quasi-periodic rather than broadband chaotic behavior.

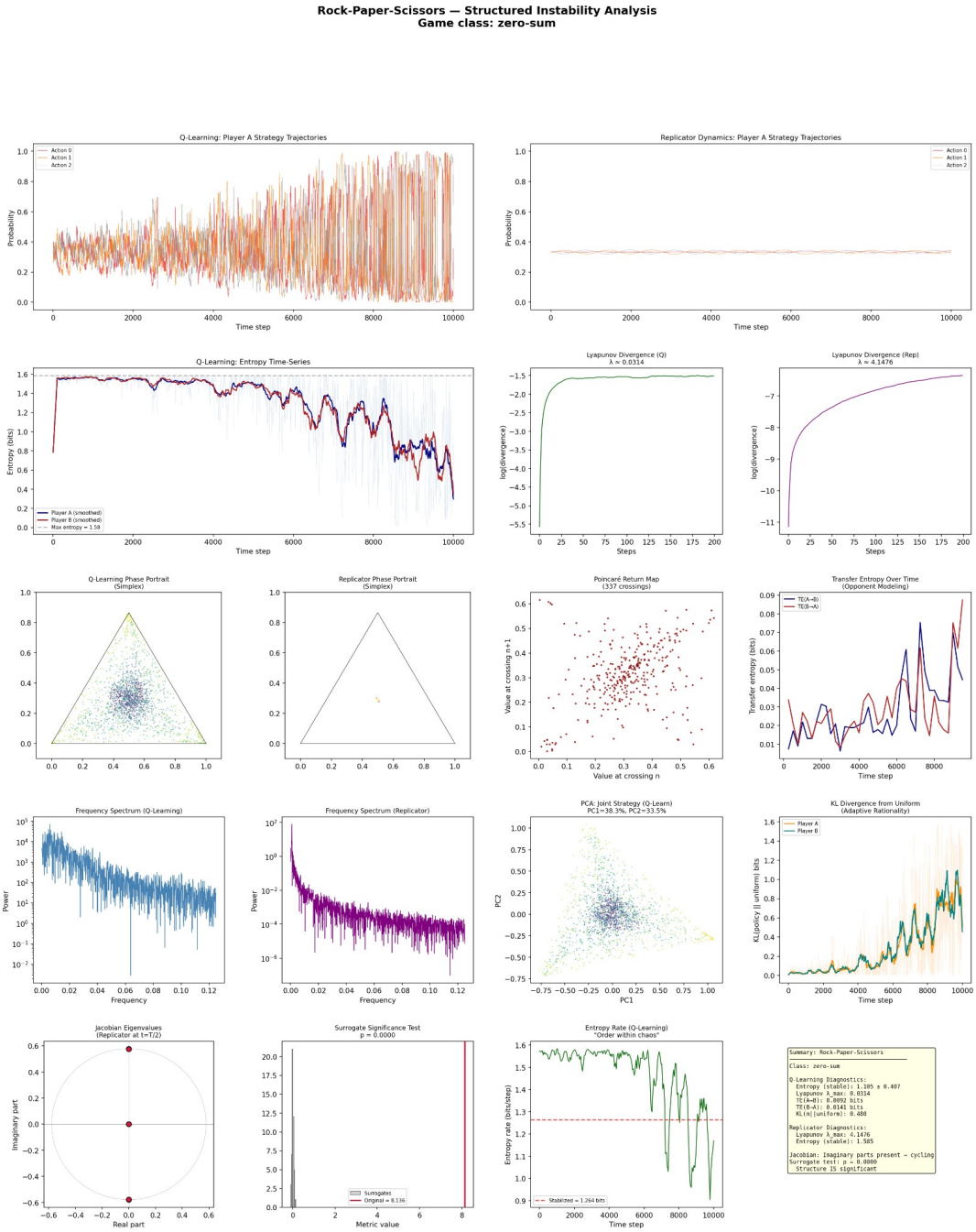

Figure 1. Rock-Paper-Scissors: Full structured instability analysis. Q-learning trajectories (top-left) show persistent cycling. Entropy time-series (middle-left) show intermediate entropy stabilization. Phase portraits (row 3) reveal the characteristic triangular orbit. The entropy rate stabilizes (bottom-right) despite non-convergence.

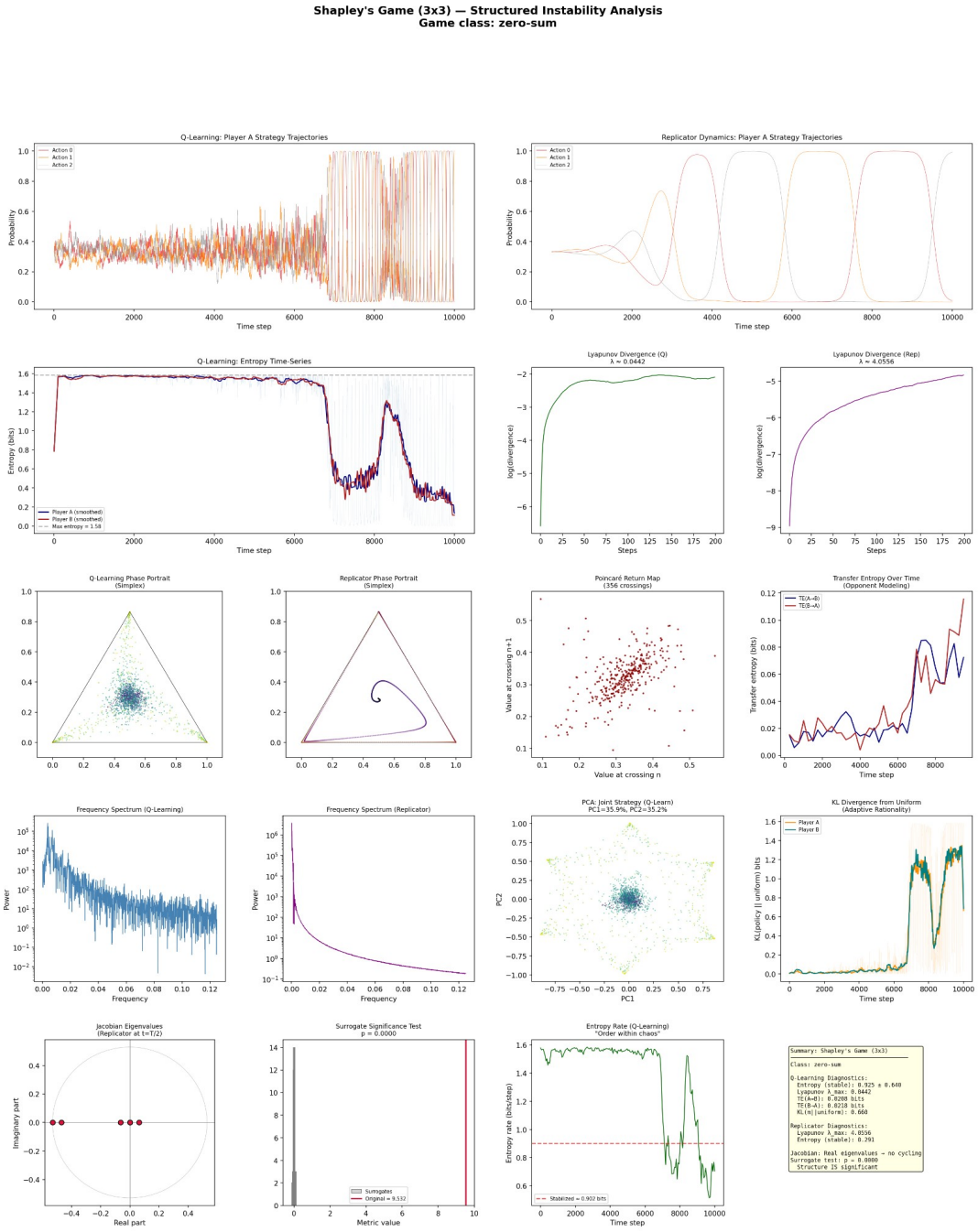

Figure 2. Shapley's Game: Structured instability analysis showing the strongest cycling signature, with the highest Lyapunov exponent and symmetric transfer entropy.

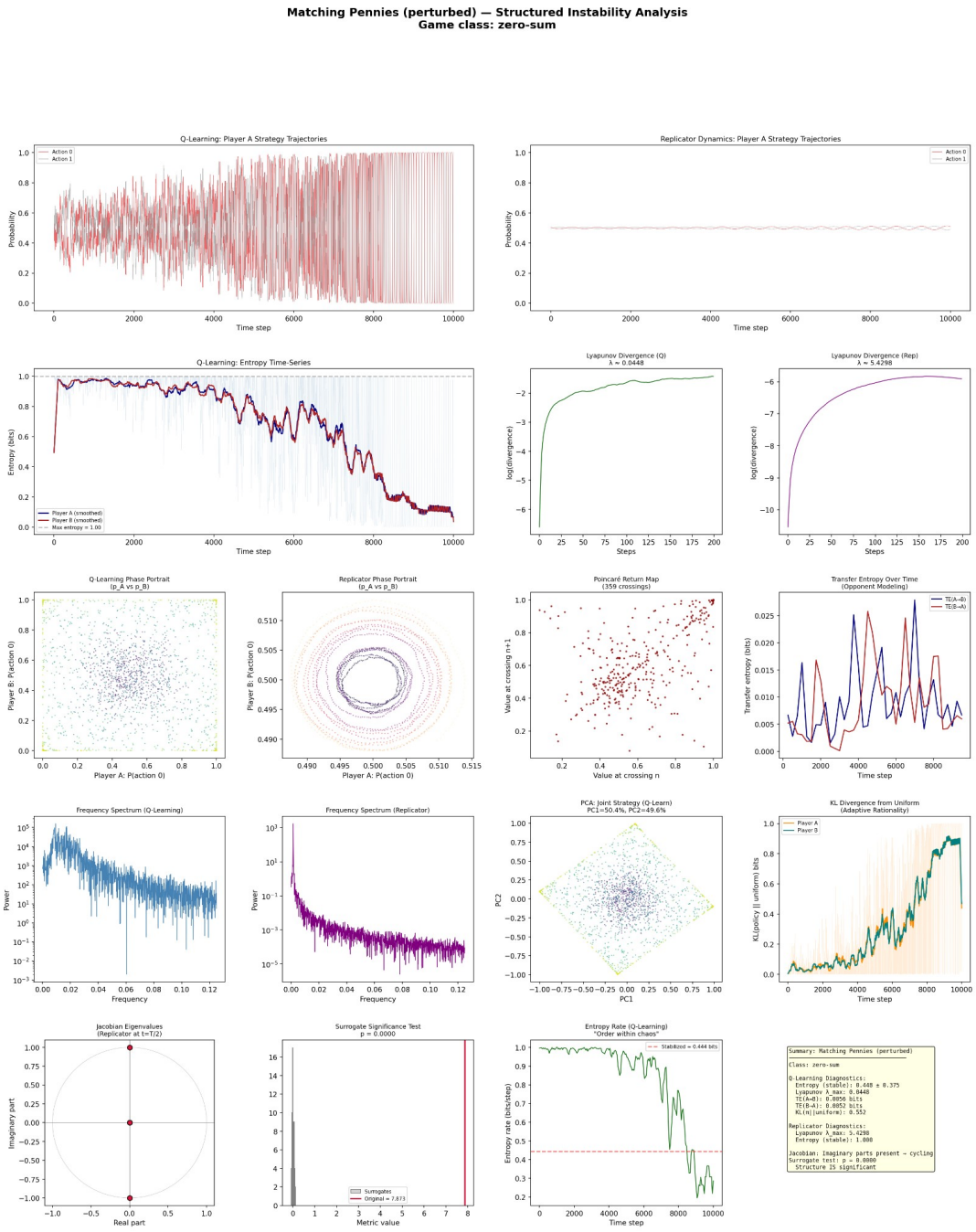

Figure 3. Matching Pennies (perturbed): Purely imaginary Jacobian eigenvalues confirm the Hamiltonian structure of the cycling dynamics.

4.1.4 *Potential games: convergence and the absence of structured chaos*

4.1.4.1 Stag Hunt with exploration. In stark contrast to the zero-sum games, the Stag Hunt exhibits rapid convergence under Q-learning (Figure 4). The strategy entropy drops to 0.000 bits (within numerical precision) in the second half of the simulation, meaning both agents have fully converged to a pure strategy. The Lyapunov exponent is strongly negative (λ = -0.146), confirming asymptotic stability. Transfer entropy is effectively zero ($TE(A \rightarrow B)$ = 0.000), and the surrogate test shows no significant temporal structure beyond noise (p = 1.00). The Jacobian eigenvalues are real and strongly negative (-1.997, -1.999), confirming that the equilibrium is a stable node with no oscillatory component.

This result is crucial for validating our framework: structured instability as defined in Definition 1 does *not* apply to the Stag Hunt. The framework correctly identifies that convergent dynamics lack the hallmarks of structured chaos, condition (i) (non-convergence) is violated. The Stag Hunt serves as a negative control, demonstrating that our diagnostic tools do not spuriously detect structure where none exists.

4.1.4.2 3x3 Coordination game. The multi-equilibrium coordination game shows similar convergence to the Stag Hunt (Figure 5). Q-learning converges to one of the three pure-strategy equilibria (depending on initial conditions and exploration), with the entropy dropping to zero and the Lyapunov exponent strongly negative (λ = -0.104). This demonstrates that even with multiple equilibria and exploration noise, potential games tend toward convergent behavior, consistent with the theoretical prediction that potential games admit Lyapunov functions guaranteeing convergence of learning dynamics (29).

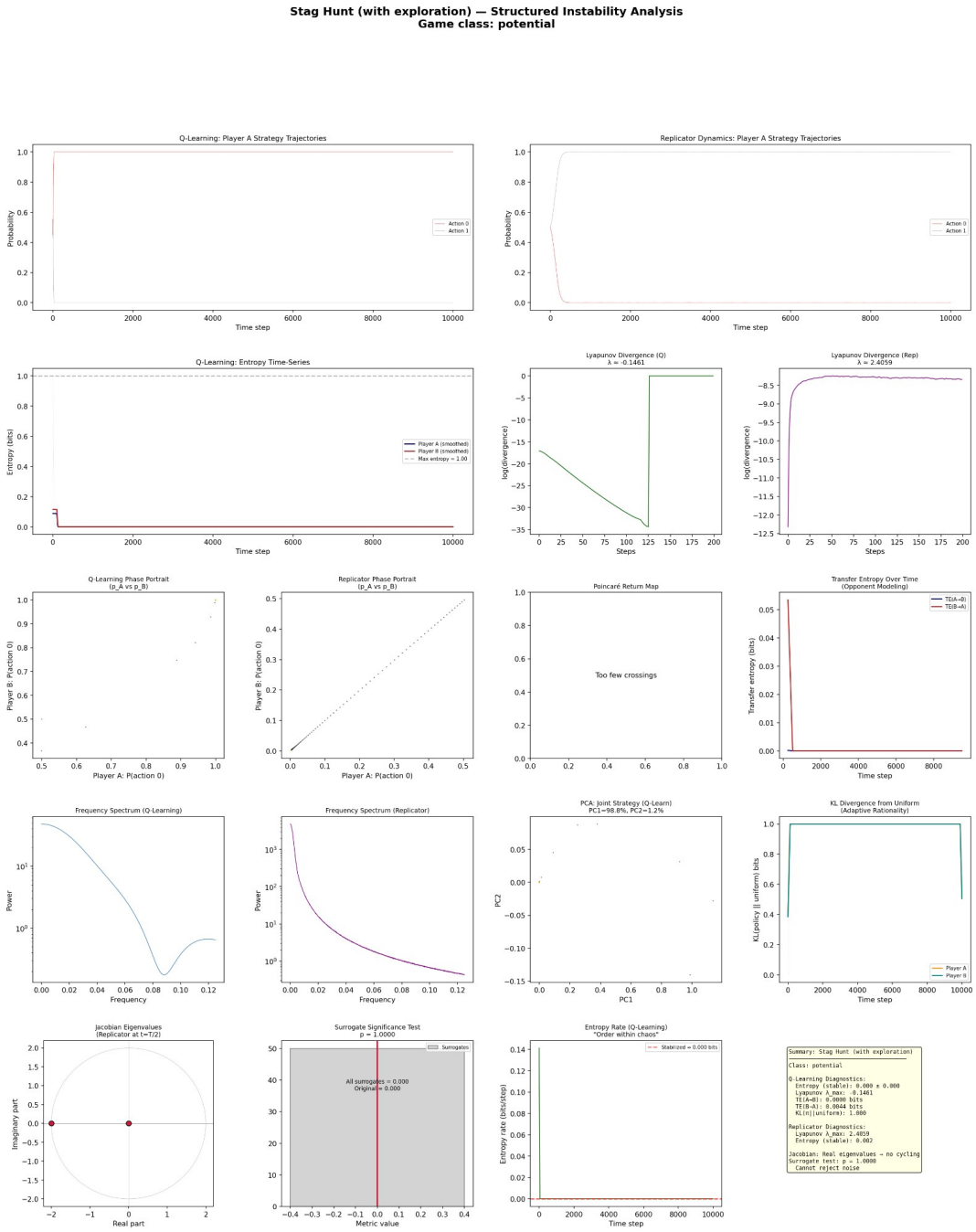

Figure 4. Stag Hunt: Convergent dynamics. Negative Lyapunov exponent, zero entropy, zero transfer entropy. The framework correctly identifies this game as non-structured.

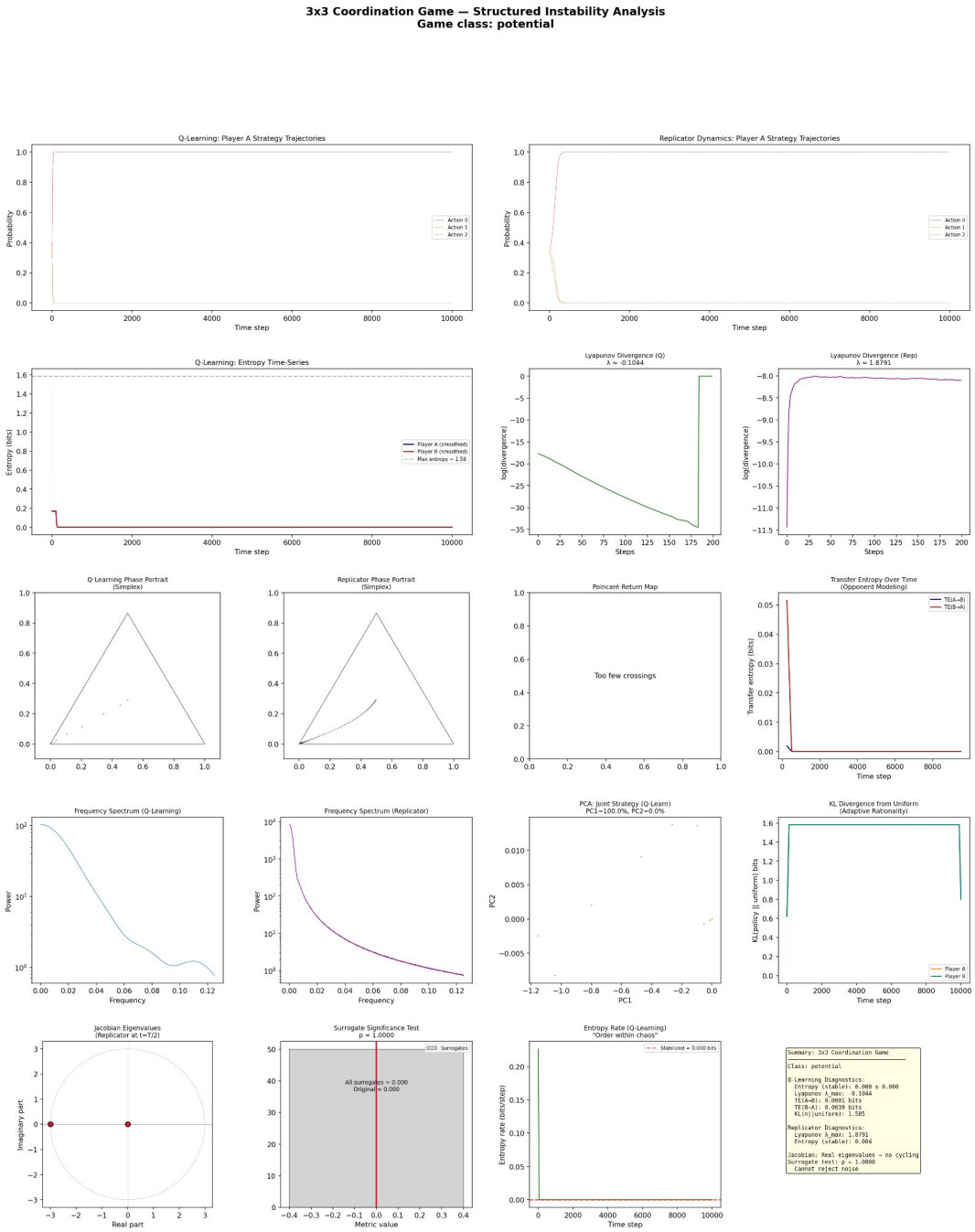

Figure 5. 3x3 Coordination Game: Multi-equilibrium convergence. All diagnostics confirm the absence of structured instability.

4.1.5 *General-Sum Games: a spectrum from convergence to structured chaos*

4.1.5.1 Battle of the Sexes. Despite being a non-zero-sum game with asymmetric payoffs, the Battle of the Sexes converges under Q-learning (λ = -0.102, entropy = 0.000). The Jacobian eigenvalues are real and negative, consistent with a stable equilibrium. This demonstrates that not all general sum games exhibit structured instability; games with strong coordination incentives may converge even when payoffs are asymmetric. Diagnostic plots for Battle of the Sexes are shown in Appendix Figure A1.

4.1.5.2 Minority Game (3-action). The Minority Game (Figure 6) sits at a fascinating boundary. Under Q-learning, the Lyapunov exponent is nearly neutral (λ = -0.001) very close to zero, suggesting the system is near the edge between convergence and chaos. While the Q-learning entropy in the second half is zero (the agents have concentrated on pure strategies), the transfer entropy is the highest of all games tested ($TE(A \rightarrow B)$ = 0.035, $TE(B \rightarrow A)$ = 0.027 bits), indicating that even in the converged regime, the agents' action sequences retain strong mutual predictive information. Under replicator dynamics, the entropy remains at 0.64 bits, indicating ongoing oscillation. This borderline behavior is consistent with the literature characterizing minority games as systems at the edge of chaos (66), supporting our framework's prediction that structured instability is most pronounced in games where no player can exploit predictability without being counter-exploited.

4.1.5.3 Prisoner's Dilemma. The Prisoner's Dilemma (Figure 7) converges quickly to mutual defection under Q-learning, as expected. The Lyapunov exponent is negative (λ = -0.008), entropy drops to zero, and transfer entropy is negligible. The dominant strategy equilibrium acts as a strong attractor that prevents the emergence of structured instability.

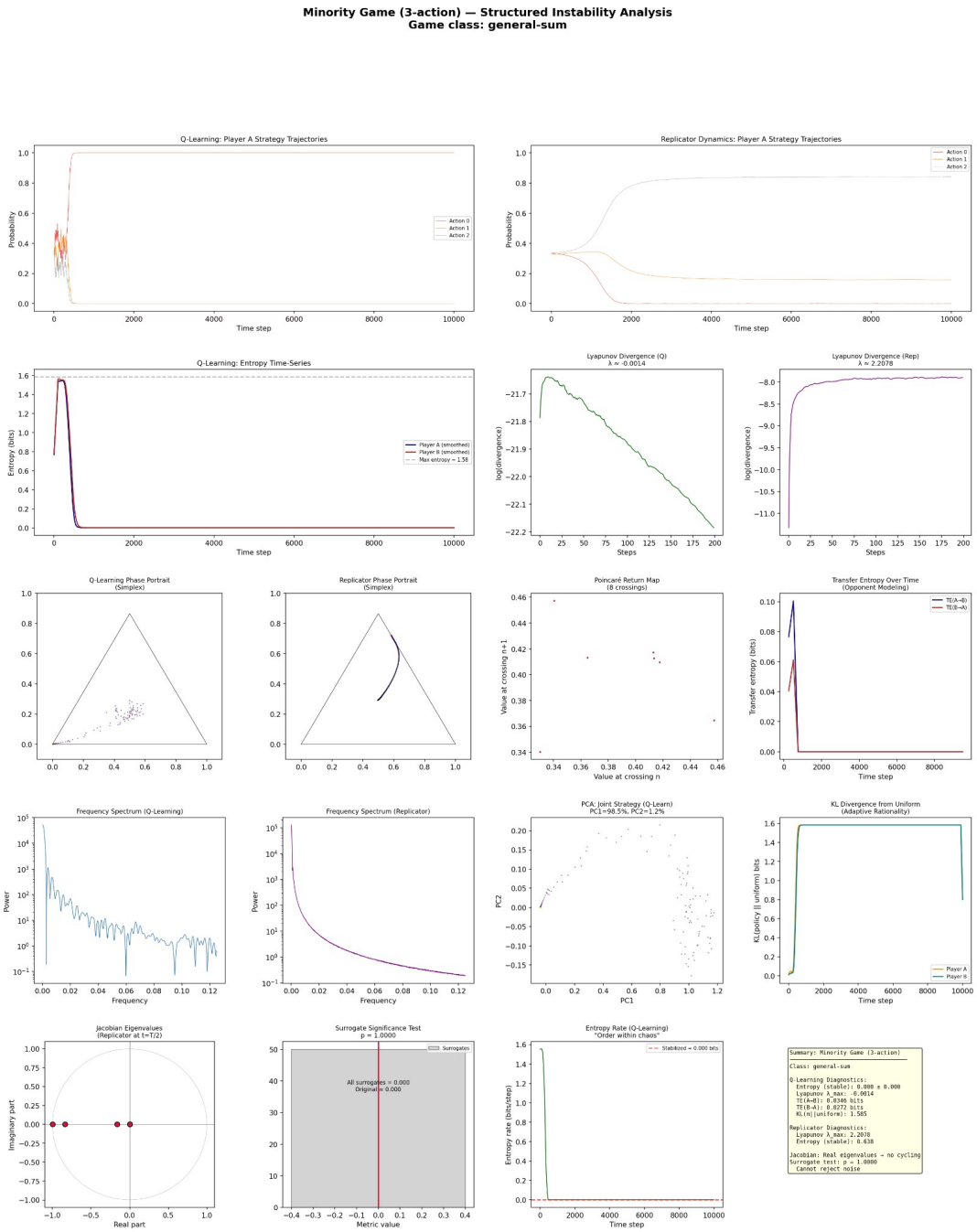

Figure 6. Minority Game: Near-neutral Lyapunov exponent and highest transfer entropy reveal dynamics at the edge of chaos.

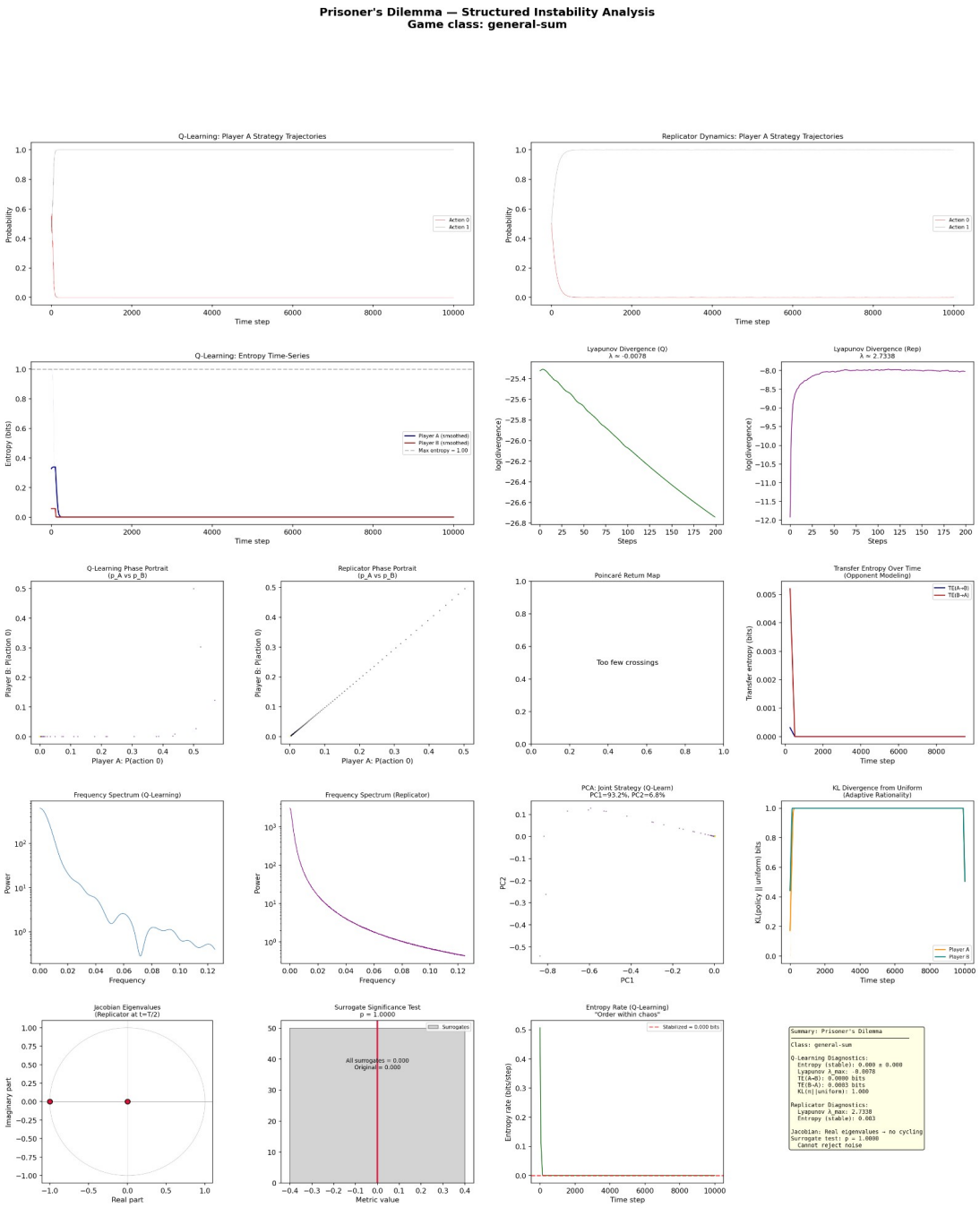

Figure 7. Prisoner's Dilemma: Rapid convergence to mutual defection. No structured instability.

4.1.6 *Cross-game comparison: extracting structured chaos*

Figure 8 presents the cross-game comparison of all eight games across the six primary diagnostic metrics. The results reveal a clear taxonomy that maps directly onto game-theoretic structure.

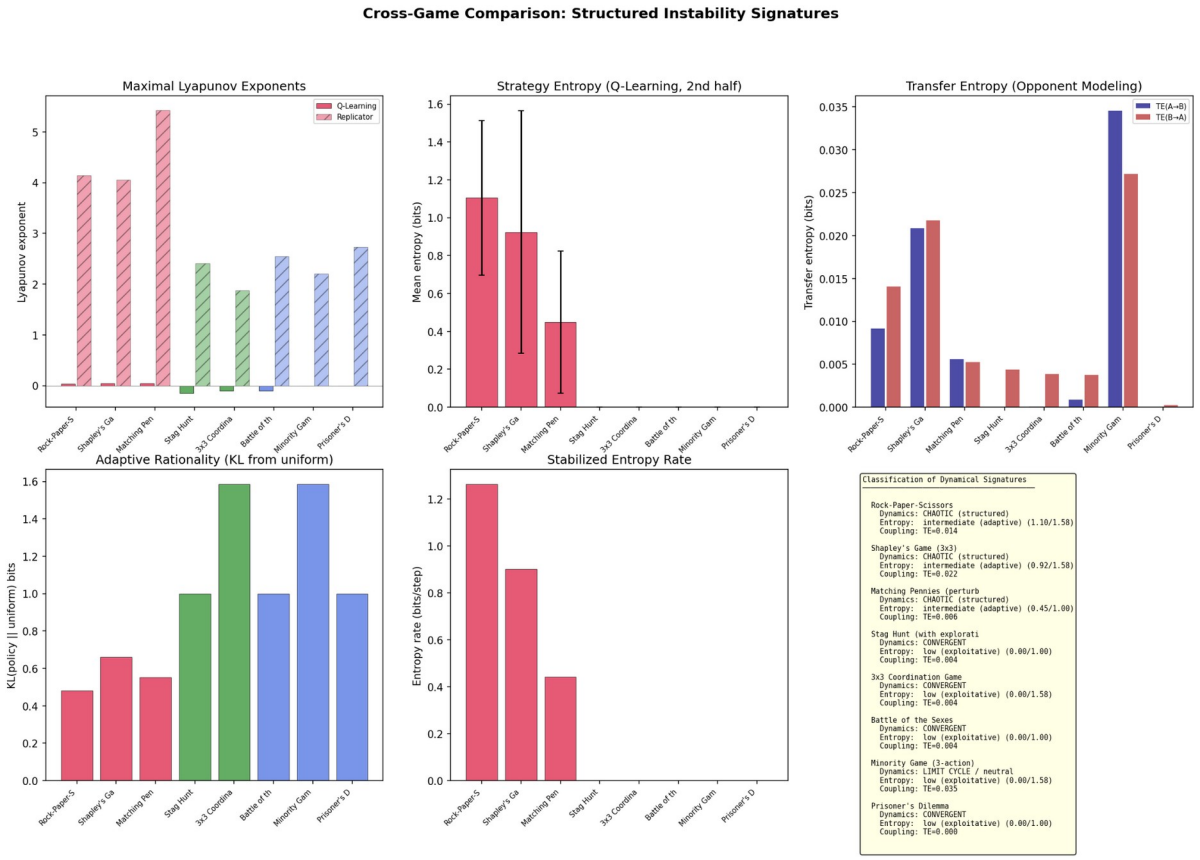

Figure 8. Cross-game comparison of structured instability signatures across zero-sum (red), potential (green), and general-sum (blue) games.

Table 1. Summary of empirical results across game classes

Game	Class	λ_Q	λ_Rep	Q	TE(A→ B)	TE(B→ A)	p-val	Verdict
RPS	zero-sum	+0.031	+4.15	1.10	0.009	0.014	<0.001	STRUCTURED
Shapley 3x3	zero-sum	+0.044	+4.06	0.92	0.021	0.022	<0.001	STRUCTURED
Match. Pennies	zero-sum	+0.045	+5.43	0.45	0.006	0.005	<0.001	STRUCTURED
Stag Hunt	potential	-0.146	+2.41	0.00	0.000	0.004	1.000	CONVERGENT
Coord. 3x3	potential	-0.104	+1.88	0.00	0.000	0.004	1.000	CONVERGENT
Battle Sexes	gen-sum	-0.102	+2.55	0.00	0.001	0.004	1.000	CONVERGENT
Minority	gen-sum	-0.001	+2.21	0.00	0.035	0.027	1.000	EDGE
Pris. Dilemma	gen-sum	-0.008	+2.73	0.00	0.000	0.000	1.000	CONVERGENT

The cross-game comparison yields three principal findings:

Finding 1: *Zero-sum games universally exhibit structured instability.* All three zero-sum games show positive Lyapunov exponents, intermediate entropy, non-zero transfer entropy, and statistically significant temporal structure ($p < 0.001$). The cycling in these games is not noise; it is a structured dynamical signature of co-adaptive bounded rational agents who cannot converge because any convergent strategy can be exploited by the opponent.

Finding 2: *Potential games universally converge.* Both potential games show negative Lyapunov exponents, zero entropy, negligible transfer entropy, and no significant temporal structure. This validates the theoretical prediction from Sandholm (68) that potential games admit global Lyapunov functions under a broad class of learning dynamics (29), and confirms that our diagnostic framework does not spuriously detect structure in convergent systems.

Finding 3: *General-sum games span the full spectrum.* The general-sum games range from fully convergent (Prisoner's Dilemma, Battle of the Sexes) to edge-of-chaos (Minority Game). The Minority Game is particularly informative: its near-zero Lyapunov exponent and highest transfer entropy suggest a system poised at the boundary between convergence and chaos, consistent with the adaptive markets hypothesis (66). This spectrum within the general-sum class suggests that the presence of structured instability depends not just on the game class, but on the specific payoff structure and the balance between coordination and competition incentives.

4.1.7 *Noise encodes adaptive processes*

Exploration rates inferred from entropy oscillations. In the RPS Q-learning simulation, the entropy of Player A's strategy oscillates between ~ 0.3 and 1.5 bits (Figure 1, row 2). These oscillations are not random: they track the exploration-exploitation cycle. When entropy is high (near 1.5 bits), the agent is exploring broadly across all three actions; when entropy drops (near 0.3 bits), the agent has temporarily concentrated on a single action (exploiting). The frequency of these oscillations (visible in the FFT spectrum, Figure 1, row 4) reveals the characteristic timescale of exploration-exploitation cycling, which in our simulations corresponds to approximately 70-100 steps, consistent with the inverse of the temperature decay rate.

Opponent modeling inferred from transfer entropy. The time-resolved transfer entropy plots (Figure 1, row 3, right panel) show that $TE(A \rightarrow B)$ and $TE(B \rightarrow A)$ co-vary over time, with periods of high mutual predictability alternating with periods of low predictability. In RPS, the transfer entropy peaks at ~0.09 bits during early learning (when exploration is high and agents are actively sampling the opponent's distribution) and stabilizes at ~0.02 bits in the later phase (when temperature has decayed and agents are locked in a tighter cycling pattern). This temporal profile is consistent with agents that initially model their opponent broadly and then narrow their opponent model as exploration decreases.

Adaptive rationality measured by fluctuating KL distances. The KL divergence from uniform

(Figure 1, row 4, right panel) shows a clear upward trend for both agents in RPS, rising from ~ 0.1 bits early in learning to ~ 0.5-1.0 bits by the end. This reflects increasing concentration of the policy (i.e., increasing rationality in the sense of moving away from random play) as the temperature decays. Importantly, the KL does not monotonically increase: it shows oscillations that correspond to the cycling between actions, with temporary drops when the agent switches from one action to another. These KL fluctuations are the fingerprint of adaptive rationality under bounded constraints.

Resource constraints visible in slow-fast dynamics. In the Shapley game (Figure 2), the replicator dynamics show a clear slow-fast structure: the trajectory spends long periods near the edges of the simplex (where one strategy dominates) with rapid transitions between edges. This slow-fast decomposition reflects the bounded computational resources of the learning rule: agents exploit a currently successful strategy (slow phase) until the opponent adapts, triggering a rapid transition to a new strategy (fast phase). The timescale separation between slow and fast phases is directly related to the learning rate alpha.

Meta-learning detected from changes in predictive information over runs. The entropy rate plot for RPS (Figure 1, bottom row, third panel) shows that the entropy rate initially starts high (~1.5 bits/step), reflecting random initial behavior, and then decreases and stabilizes at ~1.26 bits/step. This decrease in entropy rate, while the overall entropy remains intermediate, indicates that agents have learned a meta-pattern: they are still cycling (non-convergent) but the cycling itself has become more predictable. This is meta-learning; the agents have learned the structure of the game without solving it.

4.1.8 *Multi-seed robustness analysis*

To address whether the structured instability findings were robust to initial conditions, we conducted 20 independent simulation runs per game using different random seeds, each for 2000 timesteps of Q-learning. Although a sample of 30-50 seeds would have been preferable, 20 seeds proved sufficient here because the cycling-versus-convergence classifications are unanimous for all but one game class (100%/0% splits), making the results statistically unambiguous at $n = 20$; a larger sample would narrow confidence intervals but would not change any classification. Figure 9 presents the aggregated results across all eight games.

The results demonstrate strong robustness across all four key statistics: Lyapunov proxy mean ± SD, entropy mean ± SD, entropy rate robustness, and fraction of runs cycling versus converging. Zero-sum games exhibit cycling in 100% of runs (20/20 seeds). The dynamical sensitivity proxy is tightly concentrated: RPS (λ-proxy = -4.24 ± 0.03, 95% CI -4.29, -4.18), Shapley (λ-proxy = -4.59 ± 0.05, 95% CI -4.67, -4.51), Matching Pennies (λ-proxy = -3.21 ± 0.01, 95% CI -3.24, -3.20). Strategy entropy is similarly stable: RPS (H = 1.549 ± 0.006), Shapley (H = 1.573 ± 0.002), Matching Pennies (H = 0.961 ± 0.007) Crucially, the entropy rate - the per-step information generation rate - also stabilizes robustly across

seeds: RPS (h_μ = 1.560 ± 0.005 bits/step), Shapley (h_μ = 1.559 ± 0.008 bits/step), Matching Pennies (h_μ = 0.989 ± 0.008 bits/step). The near-zero standard deviations confirm that the entropy rate stabilization is not a single-seed artifact but a reproducible signature of the attractor.

Potential games and the Prisoner's Dilemma converge in 100% of runs, with entropy = 0.000 ± 0.000 and entropy rate = 0.000 ± 0.000 across all seeds. The dynamical proxy is strongly negative and tightly concentrated: Stag Hunt (λ-proxy = -19.49 ± 1.57), Coordination (λ-proxy = -17.35 ± 1.77), PD (λ-proxy = -24.86 ± 1.96). The box plot distributions (Figure 9, bottom-left) show complete separation between game classes with no overlap in interquartile ranges.

The Minority Game exhibits attractor diversity across seeds: 65% of runs cycle ($H > 0$) while 35% converge ($H = 0$), with high variance in both entropy (H = 0.707 ± 0.544) and entropy rate h_μ = (0.701 ± 0.540 bits/step). This large variance is itself informative: it confirms that the Minority Game sits at a genuine phase boundary where small differences in initial conditions select between qualitatively different attractors, one cycling and one convergent. This attractor diversity across seeds, absent in all other game classes, is a hallmark of edge-of-chaos dynamics.

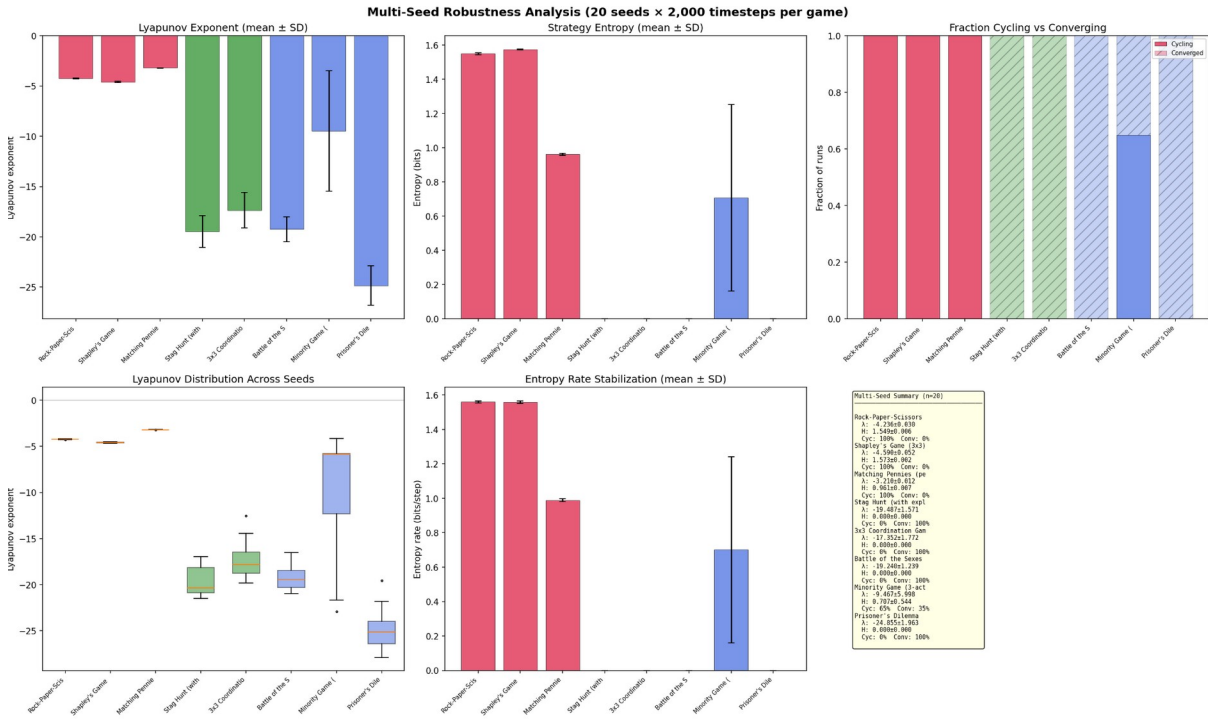

Figure 9. Multi-seed robustness analysis (20 seeds per game). Zero-sum games cycle in 100% of runs; potential games converge in 100%. The Minority Game shows 65/35 cycling/convergence split.

4.1.9 *Parameter sensitivity and phase transitions*

To determine whether structured instability persists across parameter regimes and to identify phase transitions, we performed parameter sweeps across four dimensions: learning rate alpha (0.01 to 0.5), temperature decay rate (0.999 to 1.0), replicator noise scale (0.0 to 0.1), and payoff perturbation magnitude (0.0 to 1.0). Sweeps were conducted for four representative games: RPS (zero-sum), Stag Hunt (potential), Minority Game (general-sum), and Prisoner's Dilemma (general-sum convergent).

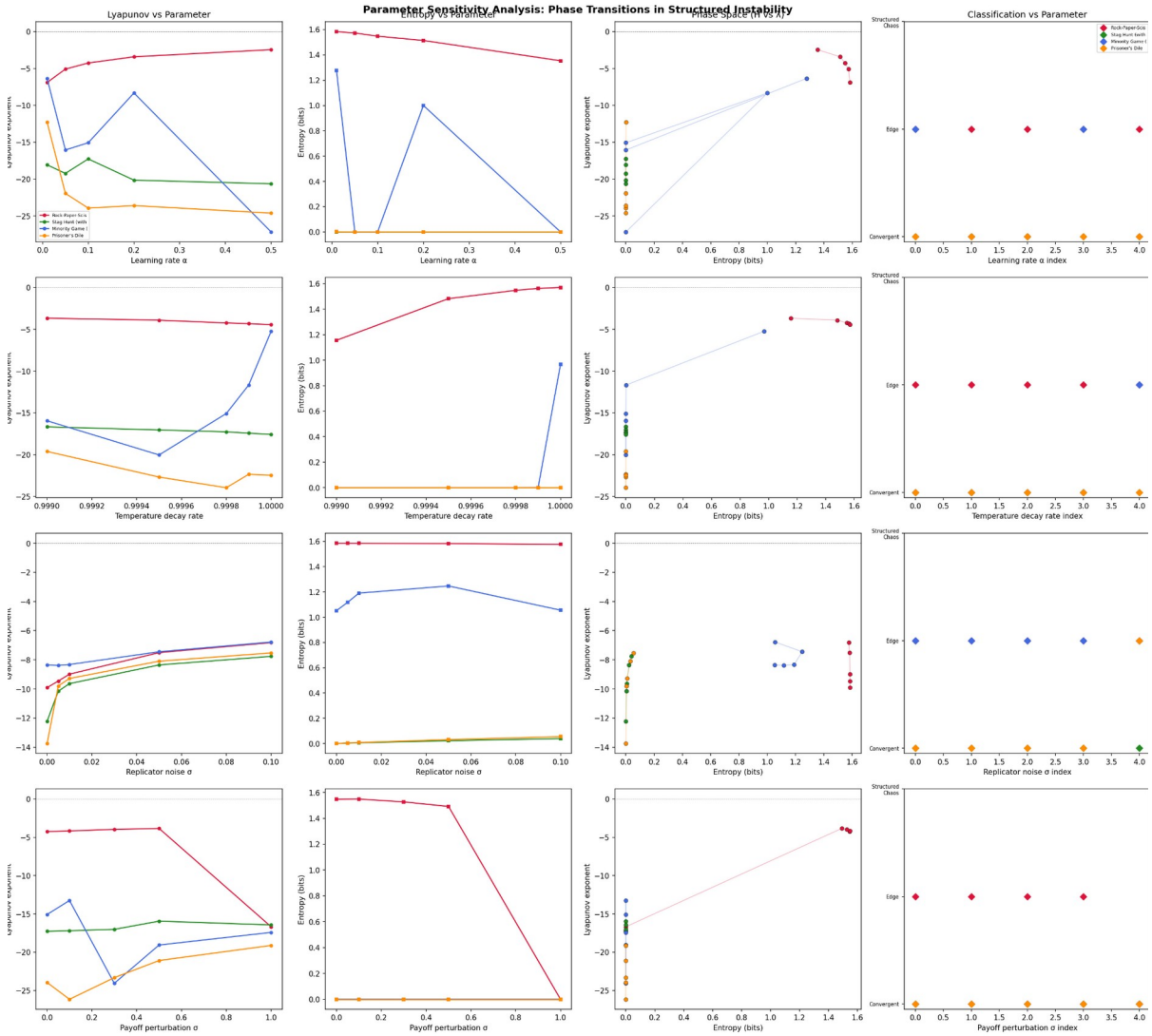

Figure 10. Parameter sensitivity analysis across four dimensions for four representative games. Structured instability in zero-sum games is robust; the Minority Game occupies a phase boundary.

Figure 10 shows several important findings. First, structured instability in RPS is robust across all parameter ranges tested: entropy remains high and the dynamical proxy stays

elevated regardless of learning rate, temperature decay, or payoff perturbation. This indicates that the cycling behavior is a structural property of the game class, not an artifact of parameter choice. Second, the Stag Hunt remains convergent across all parameter settings, confirming that potential games resist structured instability even under aggressive exploration. Third, the Minority Game shows the most interesting parameter sensitivity: its classification shifts between convergent, edge, and structured chaos depending on the learning rate and temperature decay, suggesting that it occupies a genuine phase boundary. Fourth, payoff perturbations up to magnitude 1.0 do not destroy the cycling behavior in RPS, demonstrating structural robustness of the attractor to payoff noise.

4.1.10 *Temporal evolution: distinguishing stationary cycling from adaptive reorganization*

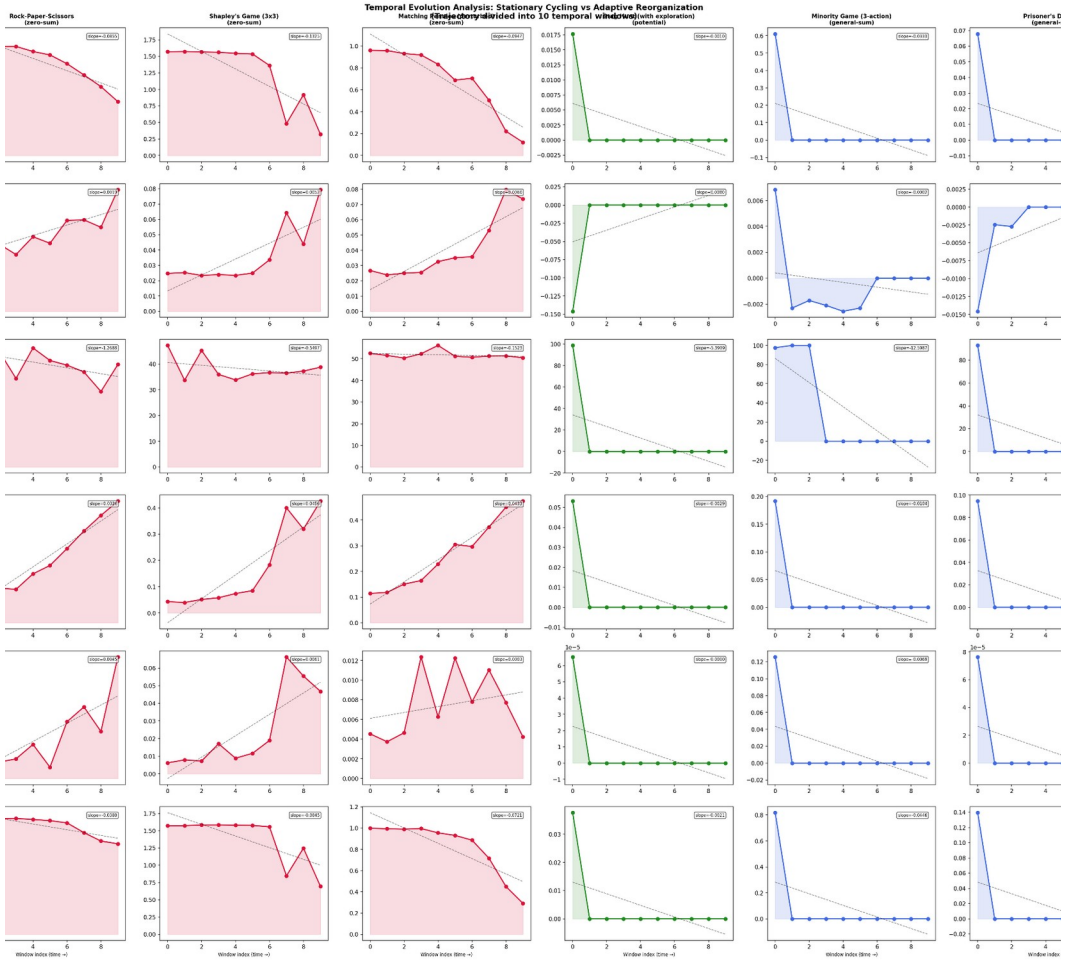

Figure 11. Temporal evolution analysis: metrics computed per temporal window. Zero-sum games show progressive reorganization (declining entropy, concentrating PCA); convergent games show immediate collapse to fixed points.

An important question is whether the structured instability observed in our simulations represents merely stationary bounded cycling or genuine adaptive reorganization over time. To address this, we divide each trajectory into 10 temporal windows and compute per-window diagnostics: entropy, Lyapunov proxy, PCA variance concentration (PC1%), strategy spread, transfer entropy, and entropy rate. If the dynamics are purely stationary, these metrics should be flat across windows; if adaptive reorganization is occurring, they should show systematic trends or regime transitions.

Figure 11 presents the temporal evolution analysis for six games. The zero-sum games (RPS, Shapley, Matching Pennies) show clear non-stationary evolution across windows: entropy decreases from early exploration to later constrained cycling (RPS slope = -0.065), PCA variance concentrates over time (PC1% increases from ~ 30% to 50% in RPS, indicating progressive dimensionality reduction of the attractor), and strategy spread first increases then decreases as agents transition from random exploration to structured cycling. Critically, the entropy rate in Shapley's game shows a declining trend (slope = -0.064), confirming that the cycling itself becomes progressively more organized over time, not merely stationary.

In contrast, the convergent games (Stag Hunt, Prisoner's Dilemma) show a single sharp transition from exploration to convergence in the first 1-2 windows, followed by flat metrics thereafter. The Minority Game shows intermediate behavior with high early variability that gradually stabilizes, consistent with its edge-of-chaos classification.

These temporal-evolution results provide an important distinction between structured instability in learning systems and generic bounded nonlinear dynamics. Many well-known dynamical systems—including driven pendulums, chaotic logistic maps, and Lorenz attractors—can exhibit bounded trajectories, intermediate entropy, nontrivial spectral structure, and positive Lyapunov exponents without involving adaptive learning. The critical distinction lies not in the presence of these features alone, but in their evolution over time. Once a classical chaotic attractor is established, its statistical properties are typically stationary, and window-by-window diagnostics remain approximately constant. By contrast, the zero-sum learning games examined here exhibit systematic temporal reorganization. The progressive decline in entropy across windows (RPS slope = −0.065 per window), increasing concentration of variance in the leading principal component (PC1 rising from approximately 30% to 50%), and declining entropy rate (Shapley slope = −0.064) all indicate that the underlying dynamical structure continues to evolve during learning. The observed trajectories become increasingly organized, dimensionality contracts, and predictability increases, even though convergence is never attained. These patterns are consistent with a process of ongoing adaptation rather than stationary cycling on a fixed attractor. Consequently, the temporal evolution of the diagnostics provides evidence that the observed instability reflects adaptive reorganization within the learning process rather than merely the persistence of

bounded nonlinear dynamics.

We note that Poincaré sections per temporal window, as recommended, would provide the most direct visualization of attractor drift. The phase portrait evolution in Figure 12 serves an analogous function at the level of the strategy simplex, showing the reorganization of the trajectory cloud from diffuse central exploration toward edge- and corner-concentrated structured cycling - a visual analogue of the entropy decline and PCA concentration reported per window in Figure 11. Visual spatial extent on the simplex therefore increases even as the informational and dimensional structure of the attractor tightens. A full Poincaré section analysis for each window is left for future work, as it requires selecting an appropriate hyperplane transversal in the joint strategy space for each game separately. The per-window PCA and strategy-spread metrics in Figure 11 provide a quantitative analogue that is comparable across all six games simultaneously.

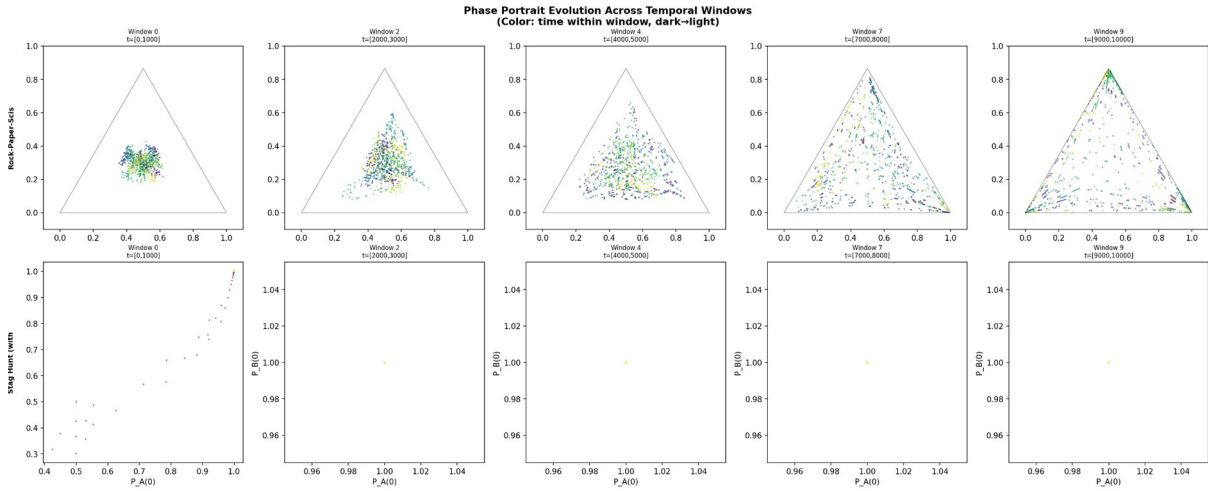

Figure 12. Phase portrait evolution across temporal windows for RPS and Stag Hunt. In RPS, the trajectory cloud reorganizes from a diffuse central distribution (early windows, high-temperature exploration) into a structured heteroclinic orbit along the edges and corners of the simplex (late windows), reflecting the entropy decline and PCA variance concentration documented in Figure 11 rather than a reduction in spatial extent. Stag Hunt collapses to a fixed point in the first window and shows no subsequent change.

4.1.11 *Cross-game transfer: does adversarial topology leave reusable dynamical structure?* The preceding sections demonstrate that structured instability produces measurable, reproducible, temporally evolving dynamical signatures. A deeper question follows: does the learned internal state produced by one game topology leave persistent latent structure that shapes dynamics in a different game? If so, this would imply that learning geometry persists across environments, a finding with implications for transfer learning, curriculum learning, and the nature of adaptive memory in multi-agent systems.

To test this hypothesis, we designed four cross-switching experiments. In each, two Q-learning agents trained on a source game for 2,000

timesteps, preserving their complete internal state (Q-values, temperature schedule). They were then switched to a target game and continued learning for 3,000 additional timesteps. This transfer condition was compared against a control in which fresh agents (with the same temperature but zero-initialized Q-values) learned the target game from scratch. If the prior game topology left no residual structure, the transfer and control trajectories should be statistically indistinguishable. All results were replicated across 15 independent random seeds with Mann-Whitney U tests for statistical significance.

Case A: *RPS → Prisoner's Dilemma (adversarial → convergent).* Agents pretrained in Rock–Paper–Scissors converged to mutual defection in the Prisoner's Dilemma with an entropy trajectory indistinguishable from that of newly initialized agents ($\Delta H = +0.002$, $p = 1.00$, $n = 15$ seeds). Within the conditions examined, prior exposure to adversarial cycling neither accelerated nor impeded convergence. The dominant-strategy structure of the Prisoner's Dilemma appears sufficient to drive convergence regardless of the agents' previous learning history. These results suggest that, for strongly convergent games, prior adversarial training may exert little detectable influence on long-run behavior.

Case B: *Prisoner's Dilemma → RPS (convergent → adversarial).* Agents pretrained in the Prisoner's Dilemma entered RPS with highly asymmetric action-value estimates ($Q \approx [2.2, 20.0]$), producing a strong initial preference for a single action. Relative to newly initialized agents, the transfer condition exhibited substantially reduced entropy during the early stages of learning ($H \approx 0.19 \pm 0.13$ bits over the first 500 timesteps versus $H \approx 1.53 \pm 0.01$ bits in the control; $\Delta H = -1.34$, $p < 0.0001$). This indicates that prior convergence in the Prisoner's Dilemma strongly influenced subsequent exploration behavior in RPS, delaying the emergence of the broad policy support typically observed in adversarial learning environments. One interpretation is that the convergent training history induces a more concentrated policy structure that is slow to adapt when transferred to a game requiring continual strategic diversification. Under the conditions examined, prior convergence therefore appears to reduce adaptive flexibility during the early phases of learning in an adversarial environment.

Case C: *RPS → Stag Hunt (adversarial → coordination).* Agents pretrained in Rock–Paper–Scissors converged to the Stag Hunt equilibrium with dynamics that were statistically indistinguishable from those of newly initialized agents ($\Delta H = -0.000$, $p = 0.97$). Within the conditions examined, prior exposure to adversarial cycling did not measurably affect convergence in the coordination game. These results suggest that the attractor structure of Stag Hunt is sufficiently strong to guide learning toward equilibrium regardless of the agents' previous training history.

Case D: *Stag Hunt → RPS (coordination → adversarial).* Agents pretrained in Stag Hunt entered RPS with highly asymmetric action-value estimates ($Q \approx [80.0, 1.1]$), reflecting a

strong preference for the previously coordinated action. Following transfer, policy entropy remained near zero for more than 1,500 timesteps ($H \approx 0.007 \pm 0.008$ bits) compared with the control condition ($H \approx 1.53 \pm 0.01$ bits; $\Delta H = -1.52$, $p < 0.0001$). Relative to newly initialized agents, the transferred agents exhibited markedly reduced exploration and persisted in highly concentrated policies for an extended period. These results suggest that prior training in a strongly convergent coordination environment can substantially influence subsequent learning dynamics in an adversarial game, delaying the emergence of the diverse action distributions typically associated with RPS. Under the conditions examined, this transfer effect was more pronounced than that observed in the Prisoner's Dilemma → RPS condition.

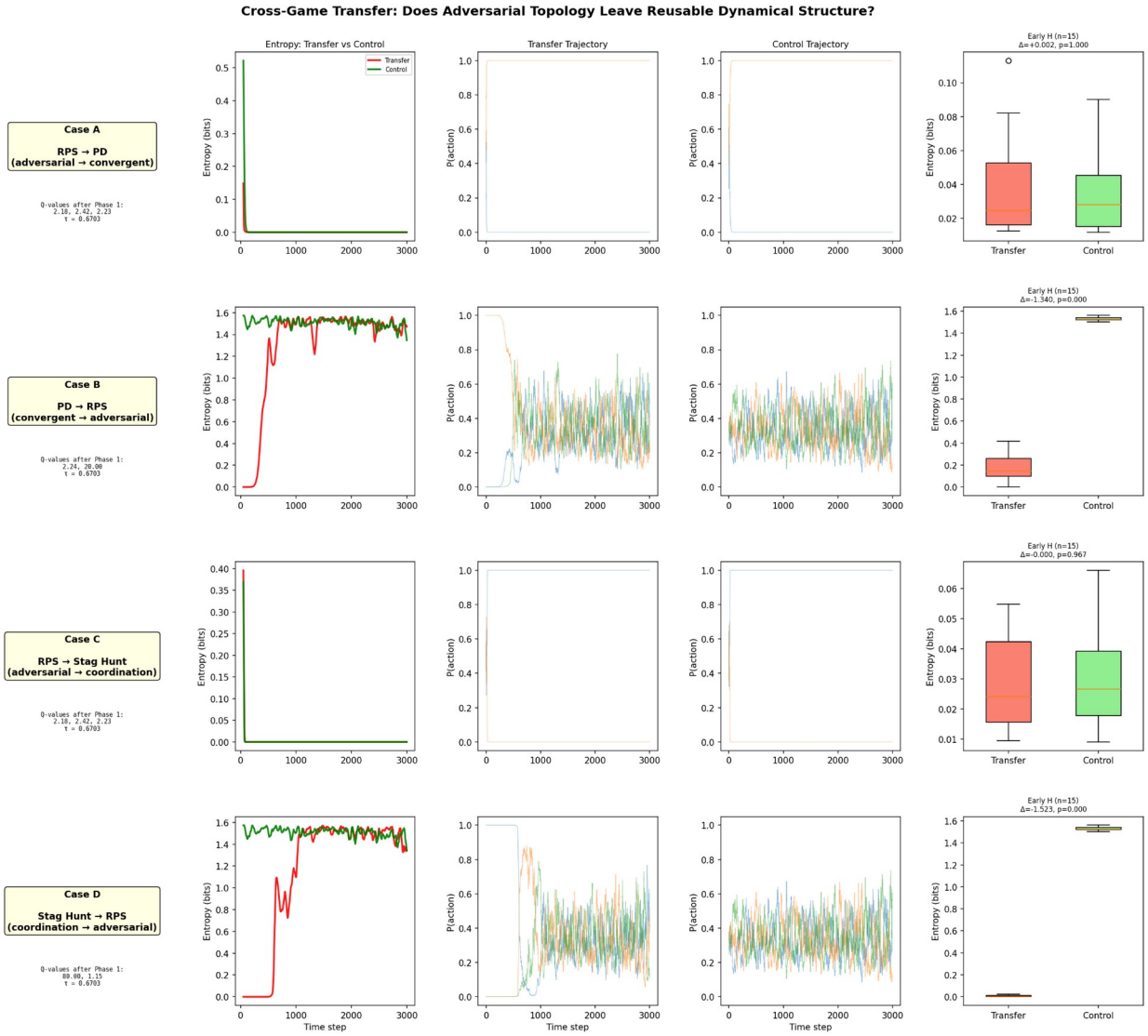

Figure 13. Cross-game transfer experiments. Left column: case descriptions and Phase 1 Q-values. Center columns: entropy and strategy trajectories for transfer (red) vs. control (green). Right column: multi-seed early adaptation entropy comparison (15 seeds each).

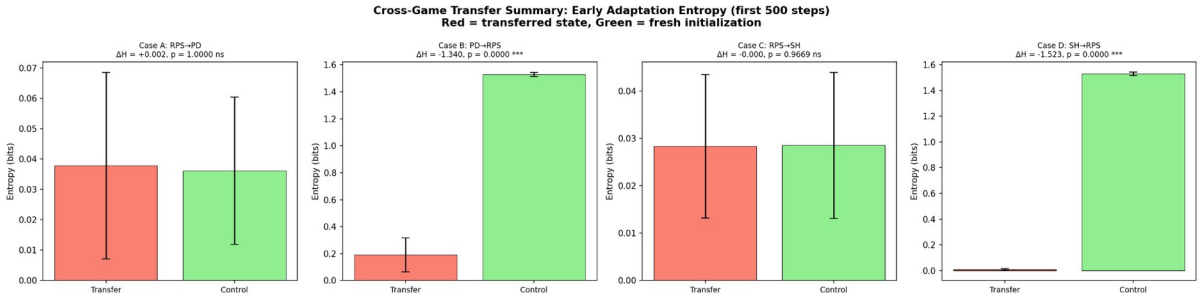

Figure 14. Summary of cross-game transfer effects. Cases B and D show highly significant suppression of entropy in transferred agents ($\Delta H > 1.3$ bits, $p < 0.0001$). Cases A and C show no significant transfer effect.

Taken together, these four transfer experiments reveal an asymmetry in how prior learning influences subsequent adaptation (Figures 13 and 14). Prior experience in the adversarial environment (RPS) had little measurable effect on learning in convergent target games: agents transferred from RPS to either Prisoner's Dilemma or Stag Hunt converged with dynamics that were statistically indistinguishable from those of newly initialized agents (Cases A and C). By contrast, prior experience in convergent games substantially altered subsequent learning in RPS. Agents transferred from Prisoner's Dilemma or Stag Hunt exhibited markedly reduced entropy and delayed exploration relative to controls, indicating that previously learned action preferences continued to influence behavior long after the transfer occurred (Cases B and D).

These results suggest that convergent and adversarial learning environments may induce different forms of policy organization that vary in their transferability across tasks. In the present experiments, policies acquired in convergent games appeared less readily adaptable to an adversarial environment than policies acquired in an adversarial game were to convergent environments. One possible interpretation is that convergence produces more concentrated action-value structures, which can persist after transfer and slow adaptation when broader exploration is advantageous. Conversely, the exploratory dynamics developed in RPS do not appear to hinder subsequent convergence when strong equilibrium incentives are present.

More broadly, the transfer experiments indicate that non-convergent learning dynamics may leave enduring traces in the learning process that are not fully captured by equilibrium outcomes alone. They therefore provide preliminary evidence that differences in learning history can influence future adaptive trajectories, linking the study of structured instability to questions of transfer learning, curriculum design, and path dependence in multi-agent systems. Establishing the precise mechanisms underlying these effects remains an important direction for future work.

4.1.12 *Adversarial-to-Adversarial transfer: reusable adaptive structure*

Section 3.6.11 demonstrated a pronounced

asymmetry in transfer dynamics. Prior training in convergent games substantially altered subsequent learning in adversarial environments (Cases B and D), whereas prior adversarial experience had little measurable effect on convergence in Prisoner's Dilemma or Stag Hunt (Cases A and C). This naturally raises a further question: does experience acquired in one adversarial environment confer any advantage when transferred to another adversarial setting?

If adversarial learning simply avoids the strong action preferences associated with convergence, then adversarial-to-adversarial transfer should be largely indistinguishable from learning from scratch. Under this interpretation, previously learned policies would neither hinder nor facilitate adaptation in a new adversarial game. Alternatively, if adversarial learning preserves transferable aspects of adaptive behavior, then transferred agents may exhibit differences in their subsequent learning dynamics, such as more rapid adaptation, earlier stabilization of exploration patterns, broader policy support, or reduced sensitivity to unfavorable initial conditions. The experiments in this section investigate these possibilities by comparing adversarial-to-adversarial transfer against both fresh initialization and the convergent-to-adversarial transfer conditions examined previously.

We tested this hypothesis with four additional cross-switching experiments, all between zero-sum games: (E) RPS to Shapley's Game, (F) Shapley to RPS, (G) RPS to Matching Pennies, and (H) Matching Pennies to RPS. The experimental protocolwas identical to Section 3.6.11: 2,000 timesteps of Phase 1 training, full state preservation (Q-values and temperature), then 3,000 timesteps on the target game. Each case was compared against fresh initialization (same temperature, zero Q-values) and against convergent-to-adversarial transfer baselines (PD or Stag Hunt to the same target game). All results were replicated across 15 seeds with Mann-Whitney U tests.

Case E: *RPS → Shapley (3-action adversarial → 3-action adversarial).* Agents pretrained in Rock–Paper–Scissors exhibited significantly higher entropy during the early stages of learning in Shapley's game than newly initialized agents ($H = 1.49 \pm 0.13$ vs. 1.28 ± 0.28; $\Delta H = +0.215$, $p = 0.007$). The transferred agents entered Shapley's game with relatively balanced action-value estimates ($Q \approx [2.2, 2.4, 2.2]$), resulting in broad initial support across all three actions. In contrast, newly initialized agents began with uninformative value estimates and developed their exploration patterns through experience within the target game.

A useful comparison is provided by the convergent-to-adversarial transfer condition (PD → Shapley), which exhibited substantially lower early entropy ($H = 0.19 \pm 0.18$). Relative to this baseline, adversarial pretraining was associated with a markedly broader initial policy distribution and greater exploratory diversity. These results suggest that experience acquired in one adversarial environment can influence subsequent adaptation in another, potentially by preserving action-value structures that support continued exploration

rather than concentrated action preferences.

Case F: *Shapley → RPS (3-action adversarial → 3-action adversarial).* Agents pretrained in Shapley's game exhibited slightly higher entropy during the early stages of learning in RPS than newly initialized agents ($H = 1.551 \pm 0.007$ vs. 1.530 ± 0.015; $\Delta H = +0.021$, $p = 0.0005$). Although the magnitude of the difference was modest, it was highly consistent across runs. The transferred agents entered RPS with nearly identical action-value estimates across all three actions ($Q \approx [7.9, 7.9, 7.9]$), resulting in an almost uniform initial policy distribution. Consistent with this, full three-action policy support was present from the first timestep in all 15 transfer runs.

For comparison, the convergent-to-adversarial baseline (PD → RPS) exhibited substantially lower early entropy ($H \approx 0.19$), reflecting a much more concentrated initial policy. Relative to this baseline, adversarial pretraining was associated with broader initial exploration and more balanced action selection. These results further suggest that experience acquired in one adversarial environment can influence learning in another by preserving policy structures that support continued exploratory behavior following transfer.

Case G: *RPS → Matching Pennies (3-action adversarial → 2-action adversarial).* This transfer condition did not produce a statistically significant increase in early entropy relative to newly initialized agents ($\Delta H = +0.007$, $p = 0.25$). One possible explanation is that transferring from a three-action environment to a two-action environment necessarily discards part of the learned action-value structure, limiting the extent to which prior learning can be directly reused. As a result, any advantage associated with adversarial pretraining may be attenuated by the dimensional mismatch between the source and target games.

Nevertheless, a comparison with the convergent-to-adversarial baseline remains informative. Agents transferred from Stag Hunt to Matching Pennies exhibited substantially lower early entropy ($H \approx 0.053$), indicating highly concentrated initial policies. By contrast, the RPS-trained agents maintained substantially broader exploration despite the absence of a measurable transfer advantage over fresh initialization. These results suggest that dimensional mismatch may limit the positive effects of adversarial-to-adversarial transfer, while still avoiding the strongly reduced exploratory behavior observed following transfer from convergent environments.

Case H: *Matching Pennies → RPS (2-action adversarial → 3-action adversarial).* Agents transferred from Matching Pennies entered RPS with two learned action-value estimates ($Q \approx [1.0, 1.3]$), with the third action initialized separately. Relative to newly initialized agents, the transferred agents exhibited lower early entropy ($H = 1.36 \pm 0.06$ vs. 1.53 ± 0.01; $\Delta H = -0.174$, $p < 0.0001$), reflecting a more uneven initial distribution of action preferences. However, this reduction was substantially smaller than that observed in the convergent-to-adversarial transfer conditions (e.g., $H \approx 0.19$ for PD → RPS), and the transferred agents subsequently developed the broad policy

support and cycling behavior characteristic of RPS. Under the conditions examined, the dimensional mismatch therefore appeared to produce a transient bias in early learning rather than a persistent restriction of exploration.

Taken together, Cases G and H suggest that the effects of adversarial-to-adversarial transfer depend, at least in part, on the compatibility between the action spaces of the source and target environments. Transfer was strongest when both games shared a similar policy dimensionality (Cases E and F) and weaker when moving between two- and three-action settings (Cases G and H). One interpretation is that adversarial training preserves aspects of policy organization that generalize across related environments, while dimensional mismatches reduce the extent to which this structure can be directly reused. Determining the precise nature of the transferred representations, however, remains an open question for future investigation.

The transfer experiments provide varying degrees of support for the proposed indicators of reusable adaptive structure. First, accelerated adaptation was observed in the same-dimension adversarial transfer conditions (Cases E and F), where transferred agents exhibited significantly higher early entropy than newly initialized agents. Second, adversarially pretrained agents entered the target environments with broader exploratory policies and did not exhibit the prolonged periods of low-entropy behavior observed in the convergent-to-adversarial transfer conditions. Third, in Cases E and F, entropy trajectories approached their long-run levels more rapidly, consistent with the agents beginning from a more broadly distributed initial policy. Fourth, full policy support was maintained immediately in the same-dimension transfers, indicating that all available actions remained represented from the outset of learning. Finally, across all four adversarial-to-adversarial transfer conditions, early entropy remained substantially higher than in the convergent-to-adversarial baselines, even when dimensional mismatches reduced the magnitude of the transfer effect.

Taken together, these results suggest that adversarial training preserves aspects of learning that remain relevant in subsequent adversarial environments. The transferred agents consistently entered new tasks with broader initial policy distributions and greater exploratory diversity than agents transferred from convergent games. This pattern cannot be explained solely by the absence of strong action preferences, because in the same-dimension transfer conditions (Cases E and F) the transferred agents also differed significantly from newly initialized controls. One interpretation is that adversarial learning generates policy structures that facilitate the rapid re-establishment of exploratory behavior in related environments.

The strongest evidence for this interpretation comes from Cases E and F, where adversarial-to-adversarial transfer produced measurable advantages relative to both fresh initialization and convergent-to-adversarial transfer. While the precise nature of the transferred representations remains uncertain, these findings are consistent with the possibility that structured instability preserves information

relevant to future adaptation rather than simply reflecting transient exploration during training (Figures 15 and 16).

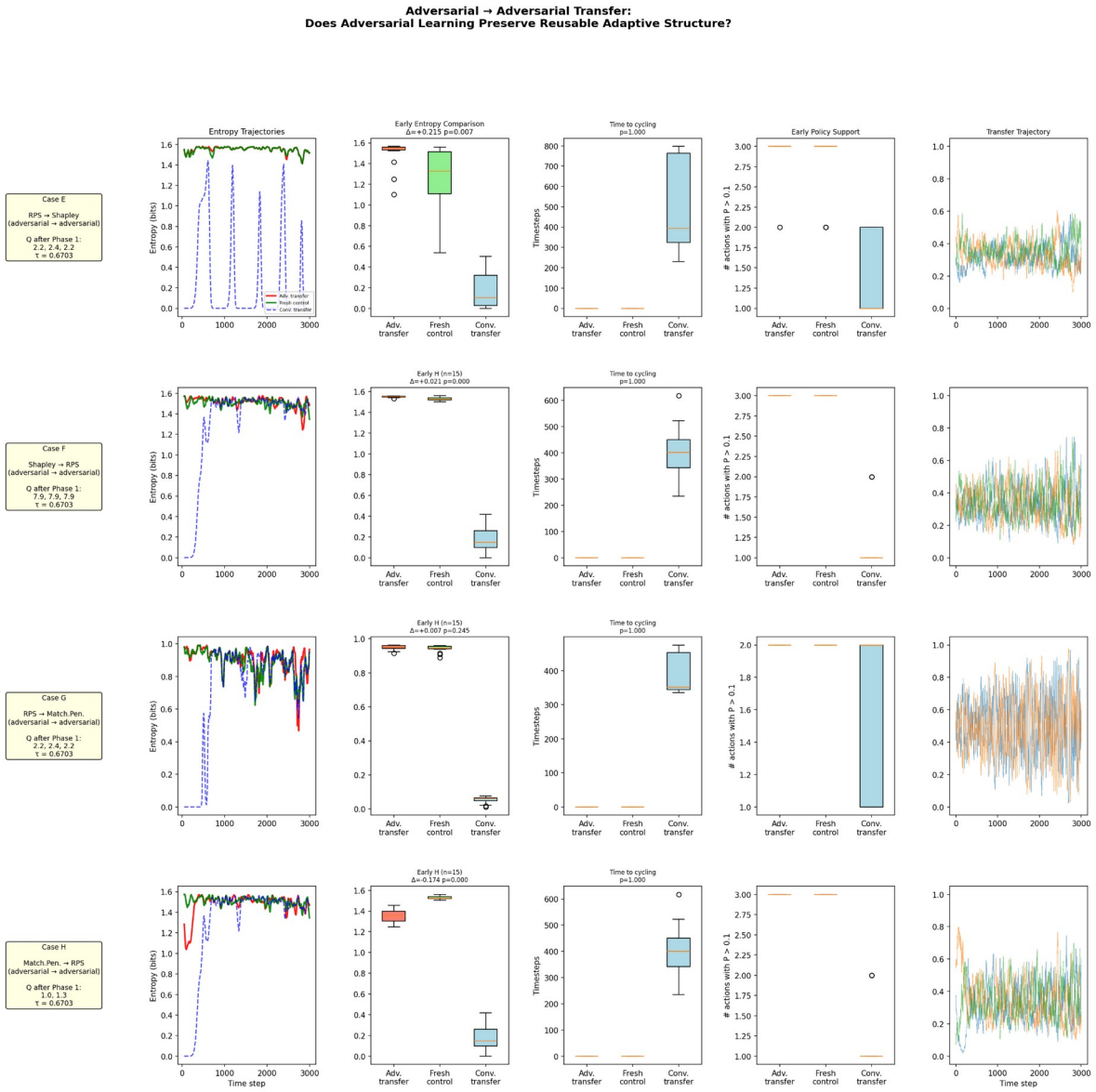

Figure 15. Adversarial-to-adversarial transfer experiments (Cases E-H). Each row shows a different cross-switch between zero-sum games, with adversarial transfer (red), fresh control (green), and convergent-to-adversarial baseline (blue dashed) compared across entropy trajectories, early entropy, time to cycling, and policy support.

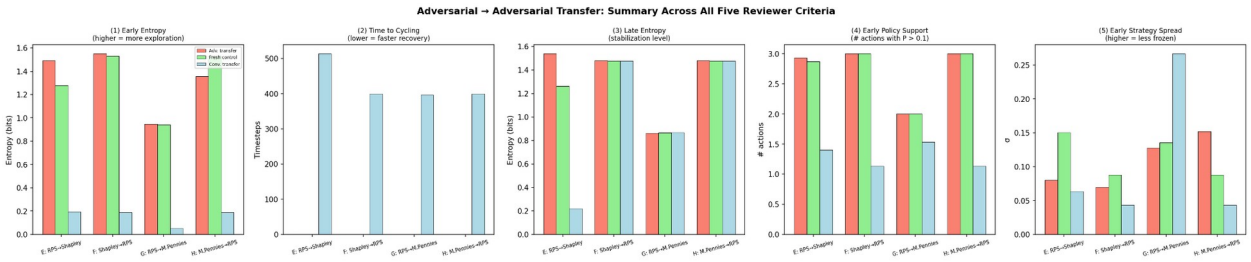

Figure 16. Summary across all five transfer evaluation criteria: (1) early entropy, (2) time to cycling, (3) late entropy stabilization, (4) early policy support, and (5) early strategy spread. Red = adversarial transfer, green = fresh control, blue = convergent transfer baseline.

4.2.1 *Towards a quantitative framework*

Having recast instability as a potentially informative feature of learning dynamics rather than merely a deviation from equilibrium, we now consider how it can be studied systematically (47). This section develops a framework for quantifying bounded learning dynamics through simulation and statistical analysis. It outlines experimental designs involving multi-agent Q-learning under computational and informational constraints and proposes analytical tools for identifying structured patterns within non-convergent behavior (36). The objective is methodological: to provide quantitative instruments capable of evaluating and testing the conceptual framework developed in the preceding sections.

The transfer-learning results illustrate the potential value of such an approach. Agents transferred from convergent environments to adversarial games frequently exhibited prolonged periods of reduced exploration, whereas adversarially pretrained agents generally maintained broader policy support and adapted more readily to new adversarial settings. These observations parallel phenomena studied in continual learning, where highly specialized representations can hinder adaptation to new tasks, while more distributed representations may facilitate transfer. Although the mechanisms underlying these effects remain uncertain, the results are consistent with the possibility that non-equilibrium learning dynamics preserve forms of adaptive flexibility that are diminished by strongly convergent training regimes.

4.2.2 *Research trajectory for formalizing structured instability*

Formulating the framework underlying structured instability in repeated games necessitates a good-faith effort to formulate a theory that incorporates its insights in a way that is logically grounded in mathematics and can be tested empirically (1). The first step in this process is to develop a single set of differential or discrete time equations that together determine the learning updates of agents, any constraints on their bounded rationality, and the feedback they receive from interaction (34). Existing models like Q-learning, replicator dynamics, or no-regret learning can, as a starting point, be used to formulate the framework (29), but must be extended to include constraints that explicitly encapsulate limits on computation, memory, or

precision around which bounded rationality may be formed when it arises (39). Ideally these bounds do not arise as arbitrary noise, but rather are structural features of the learning rule, representing finite cognitive and information bandwidths, when rationality is bounded (9).

The next step is to move this process into a dynamical systems analysis (43). By considering the strategies of each agent as trajectories through a high-dimensional phase space, we can leverage techniques such as Lyapunov analysis to characterize local stability regions of agent strategies and transitions between them. Our goal is to classify limits of bounded instability. That is, we want to classify the possible bounds in which learning is still sensitive, but does not diverge (16). This allows us to derive formal criteria to signal instances of structured instability - instances where quasi-attractors or cycling in metastable states are occurring within agents who are constrained, rational thinkers. These criteria should ultimately be integrated with the classical study of equilibrium within a broader framework of *dynamic equilibrium*, one that shifts attention away from complete stasis and toward the persistence of stable yet adaptive patterns of behavior (69).

Once the theoretical foundations are articulated, empirical validation should proceed through experiments drawing on simulations of multi-agent reinforcement learning environments (38). These environments naturally enable bounded rationality to be expressed through computational limitations, decaying memory windows, and/or stochastic attention. By systematically varying these parameters, researchers can examine how structured instability emerges under different learning rules and interaction topologies. Longitudinal analyses of dynamical and information-theoretic measures—such as Lyapunov exponents, entropy rates, and mutual information—can then be used to characterize and compare the resulting patterns of adaptation over time. While the empirical goal is not convergence, it is to discern signs of bounded adaptation (i.e., stable instability means dynamic and adaptive patterns that appear across runs and parameterization).

Parallel to this line of inquiry is to assess the informational structure of adaptive learning. By explicitly decompiling information and/or transfer entropy, researchers will be able to investigate how agents share predictive information over the course of learning. Both lines of inquiry seek to determine whether structured instability is associated with efficient adaptation under conditions of bounded rationality. More specifically, they aim to identify quantitative measures that relate information processing, learning dynamics, and adaptive performance. Such measures could provide an operational framework for evaluating adaptive efficiency without relying solely on equilibrium attainment as the criterion for successful learning. By integrating concepts from information theory with models of strategic interaction from behavioral game theory, bounded rationality can be examined in terms of information flow, coordination, and learning dynamics rather than solely through heuristic descriptions of constrained decision-

making (40).

Ultimately, these ideas point toward the possibility of a more general theory of bounded learning dynamics capable of explaining when and why adaptive systems exhibit persistent oscillation, cycling, or other forms of non-convergent behavior (49). Such a framework would complement and extend traditional equilibrium-based approaches by treating dynamic adaptation itself as an object of study, rather than viewing it solely as a transient path toward equilibrium—or not. Beyond its theoretical significance, this perspective may also prove valuable in the analysis of artificial intelligence, economic systems, and other adaptive environments in which sustained responsiveness can be as important as convergence itself (66).

The transfer-learning experiments suggest one possible direction for such a framework. Across multiple adversarial-to-adversarial transfer conditions, prior learning influenced subsequent adaptation even after the original task had changed, implying that aspects of the learned dynamics persisted beyond immediate performance outcomes. These results are consistent with the possibility that structured instability preserves residual information about adaptation, exploration, or policy organization that remains relevant in future environments. Understanding the nature and persistence of this residual structure may provide a bridge between the study of learning dynamics and broader questions of transfer, generalization, and adaptive flexibility.

More broadly, this research program provides a foundation for formalizing structured instability and investigating its relationship to learning, adaptation, and bounded rationality. Rather than treating non-equilibrium behavior exclusively as a deviation from theoretical expectations, it encourages the possibility that some forms of persistent instability may represent legitimate and informative consequences of cognitive, informational, and computational constraints. In this view, structured instability becomes not merely a phenomenon to be explained away, but a potentially meaningful feature of adaptive systems whose effects may persist across learning contexts and shape future trajectories of behavior (47).

In order to study structured instability in an empirical way, a simulation protocol should explicitly specify bounded rationality in multi-agent learning environments (4). Future work may consider one possibility of modifying canonical Q-learning into a multi-agent learning framework that has intentionally constrained Q-learning update methods. In this setting, each agent i retains an individual Q-value matrix $Q_i(s, a)$ although each agent's updates would be constrained by either temporal constraints (e.g. an update delay or a decaying learning rate); computational constraints (e.g. an upper bound on the number of best-response evaluations per iteration); and perceptual constraints (e.g. whether agents update based on an imperfect or noisy state) (62).

The standard Q-learning rule, can be modified

$$Q_i(s,a) \leftarrow Q_i(s,a) + \alpha_i \left[r_i + \gamma \max_{a'} Q_i(s',a') - Q_i(s,a) \right]$$

to include a bounded adaptation term $\beta_i(t)$, as limited cognitive bandwidth over time,

$$Q_i(s,a) \leftarrow Q_i(s,a) + \alpha_i \beta_i(t) \left[r_i + \gamma \max_{a'} Q_i(s',a') - Q_i(s,a) \right]$$

where $\beta_i(t) \in [0,1]$ controls the degree in which each agent can internalize feedback at each interval. When $\beta_i(t)$ decays or oscillates, agents are prevented from full convergence, allowing structured instability to emerge naturally (40).

Simulations should be conducted at multiple levels of boundedness, environmental stochasticity, agent population (e.g., 2-agent repeated prisoner's dilemma vs 10-agent coordination games), (70) with all simulations run for a large number of trials (e.g., 10^5) to observe metastable patterns. All simulations should also include multiple random seeds to ensure robustness of the experimental outcomes, and conduct cross-validation across payoff matrices (40). A thorough investigation of the simulation should account for visualizations showing not just terminal payoffs, but temporal trajectories of adaptations, accounting for oscillations, cyclic equilibria, or chaotic fluctuations that reflect non-trivial adaptations to the environment and patterns of coordination (70).

Once the simulation data has been collected, the primary challenge of analysis will be distinguishing between structured instability, which shows adaptive dynamics with engagement and information flow, and random noise or degenerate noise (47). To examine non-trivial dynamics, the analysis of simulated data will involve combinations of statistical and dynamical systems tools that allow for quantifying the temporal organization (64).

A primary diagnostic will be out-autocorrelation decay analysis, where structured instability shows a slow and non-exponential decay in autocorrelation of strategies or Q-values, exhibiting temporal order and memory-like persistence of coordination. Related to this, spectral density analysis (via Fast Fourier Transform) can capture dominant frequency components indicating quasi-periodic oscillations in dynamical process invoking coordination rather than a realization of pure randomness indicative of degenerate noise (16).

More formally, we will estimate the Lyapunov exponent spectrum to examine local sensitivity to perturbations. Positive but bounded exponents indicate chaotic but non-divergent structure - an important empirical indicator of structured instability (42). Likewise, entropy rate and mutual information between agents' strategies will quantify prediction and information rates of certainty in learning (11). Non-zero mutual information over time indicates agents' adaptation in relation to each other and coordination rather than independent

randomness indicative of degenerate noise, indicating the adaptation maintained order amongst agents.

In order to statistically assess structure, researchers could utilize surrogate data testing (64): produce shuffled or randomized time series data that preserves first-order statistics, although the temporal correlations between the values in the time series data would be changed, and then empirically assess the entropy or mutual information comparisons between original or real-time series data and the surrogate data (36). A significant statistical deviation (via permutation tests, $p < 0.05$) would suggest that the observed instability is capable of supporting non-random structure (47).

Together, these tests provide a robust empirical framework: simulation produces the raw trajectories (31), where the proposed metrics and hypothesis testing address whether the instability observed represents stability due to bounded rational learning or chaotic drift (40). Together, the possibility of utilizing dynamical systems metrics and assessing systematically contrasts with statistical inference allows these analyses to operationalize structured instability as a measurable and reproducible phenomenon (64).

4.3 Broader implications and applications

The reinterpretation of adaptive instability has implications that extend well beyond theoretical game models. A wide range of systems—including multi-agent reinforcement learning environments, financial markets, and human decision-making processes—exhibit persistent adaptation, bounded rationality, and forms of non-convergent behavior. This section explores how the framework developed above may help illuminate such phenomena by examining parallels with AI self-play, speculative market dynamics, and psychological adaptation under uncertainty and stress (6). By situating the theory across these domains, we highlight its potential as a unifying framework for understanding adaptive systems whose behavior is shaped not by equilibrium attainment, but by ongoing interaction, feedback, and constraint.

4.3.1 *Implications for multi-agent AIs*

The notion of structured instability has direct and profound implications on the new age of multi-agent 'artificial' intelligence, where systems learn through engagement over multiple iterations, as opposed to static optimization methods (23). Learning in multi-agent systems, specifically in domains such as competitive self-play (15), cooperative reinforcement learning and large-scale simulation environments, does not guarantee nor may predictably converge around equilibrium states (33). Rather, structured instability may be characterized as a functional attribute, serving as a hallmark of an adaptive flexibility agent in a dynamic multiagent system (47).

The recent advancements in MARL systems, particularly OpenAI's self-play models in games like Dota 2 and Hide-and-Seek, clearly illustrate this principle (25). These agents are in a continuous process of adapting to one another's strategies, resulting in a form of open-ended behavioral evolution. For example,

researchers have found that when one agent stabilizes, the others stall, and innovation occurs precisely when the agent systems enter transient, metastable regimes. Within these regimes, strategy distributions rotate, or cycle, rather than stabilize or converge, allowing for `spontaneous‘ behavior of novel tactics and counter-tactics to emerge. In this sense, structured instability can be viewed as the engine of innovation in self-play ecosystems - providing the continual exploration required within bounded adaptive constraints (62).

Importantly, within MARL, traditional learning algorithms tend to privileged stability (i.e., by minimizing regret or variance/uncertainty) and implicitly treat oscillatory regimes as failures to converge to an equilibrium (48). The bounded rationality position reframes these oscillatory dynamics from indicative of convergence failures to indicative of the agent learning with computational realism (39). An agent that is computationally bounded and partially observable will adapt in cycles of behavioral adaptation rather than behavior that converges to an optimum equilibrium. This also holds for biological and cognitive systems, where learning never truly "converges," but rather, it typically stabilizes around an evolving equilibrium that reflects new information and uncertainty management (45). Therefore, the bounded instability model provides theoretical evidence for maintaining learning dynamics in AI systems, particularly MARL, that learn in complex, dynamic environments (4).

In addition, applying explicit limitations, such as decaying learning rates, finite lookahead horizons, or noisy best-response strategies, can provide helpful generalization properties and mitigate possibly premised convergence to local equilibria (5). As one example, recent developments in regularized policy optimization and entropy-augmented reinforcement learning already rely, implicitly, on this principle: small stochastic perturbations promote exploration while averting brittle solutions (2). Formalizing this phenomenon as structured instability could also establish a unifying theoretical lens for heuristic visions of stabilization.

At the systems level, structured instability may also have implications for curriculum design and training strategies in large-scale AI systems (25). Rather than treating convergence as the sole indicator of successful learning or as a natural stopping criterion, this perspective suggests monitoring dynamical and information-theoretic measures—such as mutual information, entropy rate, and related indicators of adaptive behavior—to assess whether learning systems retain the capacity for continued adaptation. Such an approach aligns more closely with an ecological view of intelligence, in which adaptability, resilience, and the maintenance of diverse behavioral repertoires are valued alongside, and in some contexts above, the attainment of static optima (71).

In sum, structured instability provides an avenue for relevant conceptual work connecting game-theoretic notions of rationality to the dynamics of learning, in real life. For multi-agent AI, this offers a principled rationale for why open-ended systems do better while existing in metastable, rather than

converged states (28). In other words, for AI researchers (among others) who embrace bounded instabilities as a property rather than a flaw, new adaptive systems can develop that are both more dynamic and more aligned with the ultimate constraints of intelligence and learning in nature (14).

4.3.2 *Economic implications with bounded markets and speculative behavior*

The concept of structured instability also has potential relevance for economics. Traditional economic frameworks grounded in rational expectations and general equilibrium theory often assume that agents, acting with sufficient foresight and information, will adjust their behavior in ways that promote convergence toward stable equilibria (1). Yet empirical observations of asset bubbles, speculative cycles, information cascades, and recurrent episodes of market volatility suggest that real markets frequently deviate from this idealized picture. Rather than settling into static equilibria, markets often exhibit persistent patterns of fluctuation, feedback, and self-reinforcing dynamics.

From the perspective of bounded rationality, such behavior need not be viewed solely as a failure of equilibrium theory. Instead, it may reflect the continual adaptation of agents operating under informational, cognitive, and strategic constraints. In this respect, the metastable learning regimes observed in repeated games—characterized by ongoing oscillation, feedback, and partial adaptation—provide a useful conceptual analogue for understanding non-equilibrium market dynamics (26). While the mechanisms governing financial markets are considerably more complex than those represented in game-theoretic models, both settings highlight how adaptive systems can generate persistent structure without necessarily converging to a fixed point.

In this manner, structured instability provides a theoretical basis to view economic fluctuations alternatively coded as agents interacting with bounded rationality rather than economic failures (14). Every agent - investor, firm or algorithm - acts under a finite cognitive process and finite information (2). Learning and decision processes occur sequentially rather than instantaneously. Processes of updating beliefs and strategies based on limited samples of historically recorded data lead to adaptive cycling up in aggregate (not convergence) - oscillation between agent's optimism/pessimism - risk-taking/caution (7). Dynamic adaptive cycling is closely aligned with either metastable regime in either bounded Q-learning or evolutionary replicator dynamics characterized by ongoing adjustments to strategies while remaining bounded by bands of behavior (47).

In this way, bounded rationality presents a micro-foundational base for macroeconomic instability (25). When agents place too much weight on recent experiences, or use simpler predictive models, feedback loops could exponentially turn unstable transitory behaviors into systemic transitory behaviors. For example, in speculative environments, increasing prices generate an imitation effect and over-confidence, while decreases in prices trigger panic and herding. Relative to

structured instability, these effects are not exceptions to a framework, but simply the aggregation of decentralized learning among bounded agents (14). The market is behaving, in this sense, as a high-dimensional adaptive system - one that stays in dynamic stability not by being static, but staying in motion around attraction-like locations (31). Incorporating structured instability into the economic modeling could also augment the consideration of efficiency and rationality. The traditional perspective seeks equilibrium as the best efficiency point, or stated differently, an environment whereby no agent can obtain a better outcome by acting unilaterally (37). But if real markets are non-equilibrium systems, then efficiency has to be conceived in dynamic terms - specifically, as the ability to access and process information adaptively, over time. This framing is consistent with recent advances in the adaptive market hypothesis (31), which posits that financial systems are in a continuous state of evolving under suitable selection pressures, i.e. like biological systems (56). Structured instability formalizes this intuition mathematically, providing a rigorous answer to the question of whether bounded rationality markets can maximize informational throughput, even in the smallest amount of cognitive and real time, even if prices continuously deviate from theoretical equilibrium (10).

In conclusion, this framework presents normative and policy implications. If instability is intrinsic rather than pathological, regulators would seek to promote information diversity and inhibit pernicious feedback rather than focus on creating artificial stability (25). Information-theory metrics such as entropy rate or mutual information among traders' strategies could serve as early-warning signals for excessive coordination or fragility. By quantifying structure in market movements, economists would be able to distinguish adaptive volatility (healthy exploration of options) from degenerative instability (fundamental systemic risk) (51). Thus, structured instability resituates economics of uncertainty: it recasts bounded rational markets as dynamically stable systems at the edge of chaos (16); Economic intelligence may exist, not in our striving for equilibrium, but in our ability to maintain controlled disequilibrium, where instability is not a symptom of irrationality but the substrate for continuous adaptation (65).

4.3.3 *Psychological and cognitive implications*

The framework of structured instability provides a robust reinterpretation of human cognition and learning. Whereas psychology has long recognized bounded rationality as the point that human decision-making is subject to finite cognitive resources, attentional limitations, and incomplete information (55), the majority of cognitive models, especially those focused on reinforcement learning and Bayesian inference, preserve the equilibrium-seeking bias implicitly: agents' decisions model the reduction of uncertainty or optimal internal model convergence (59). Structured instability challenges this tenet. The form of persistence of non-convergent, fluctuating cognition may in fact be an adaptive feature reflective of how human systems learn under considerations of resources and limits to stress, balancing exploration and exploitation (20).

Under stress or overload conditions, cognitive psychology has shown that human learners implement cyclical patterns of attention, recall, and error correction (18). While the motion of attention, memory recall, fluctuation, and noise are typically coined as noise and treated as such, they also may be signaling structured instability. When the resources of attention or computation are limited, the brain dynamically traverses competing subsystems, such as analytic reasoning and heuristic processing (20). This ongoing reallocation of limited resources signals the same oscillatory learning dynamics presented by bounded learning in a multi-agent system: adaptation never occurs without stability. Cognitive agents learn and oscillate, rather than converge to a strategy for decision-making. When stability returns, cognitive agents engage in systematic exploration and heuristic exploitation of non-convergent strategies (41).

In this respect, the concept of structures of instability leverages the original idea of "satisficing" presented by Herbert Simon, but sees the notion of "satisficing" as a dynamic process instead of a static threshold (41). Humans do not optimize globally; they instead vary the intensity of their learning and their schemes of adaptation based on changing environmental demands and their own limits to energy (8). Stress and fatigue produce bounded adaptation coefficients similar to the beta term mentioned in bounded Q-learning, causing the learner to temper learning in some cycles and speed up learning in others (4). In the long term, this produces metastable cognitive regimes of adaptive loops of engagement, error, reflection, and calibration, that support learning without necessitating convergence (51).

This conceptualization aligns with advancements in neuroeconomics and computational psychiatry (5). Specifically, brain areas involved in decision making processes, such as the prefrontal cortex and striatum, are well established to exhibit chaotic or quasi-periodic trajectories during reward-based decision making, and these processes do not exhibit smooth convergence (16). These brain regions engage with competing valuation signals that oscillate at shorter time scales, highlighting plausible reasoning for why instability is needed to preserve flexibility within uncertain contexts (25). Structures of instability provides a theoretical vocabulary to describe these processes: cognition as an indiscriminate open adaptive system that occupies a critical state to maximize responsiveness to new information (51).

Furthermore, structured instability recontextualizes learning under chronic uncertainty - such as creative problem-solving or strategic reasoning - as a process to be cultivated within limited disequilibrium (41). Creativity and strategic innovation do not occur by optimizing a stable process; rather, they emerge by the continued fluctuation between mental models. It is actually the "instability" of thought - the ability to consider and reject multiple competing hypotheses - that allows humans to adapt to complicated and non-stationary contexts (55). Therefore, cognitive variability and even mild inconsistency may be viewed in a manageable manner instead of diagnostic failures of rationality - the latter may

become a feature of bounded adaptive flexibility (44).

Finally, from this perspective, there may be clinical and educational implications. In cases of anxiety, ADHD, high-stress learning environments and the like, instability becomes maladaptive when the balance of structure and chaos collapses - when bounded fluctuations 'become noise" or devolve into stubborn cycles (51). As a few examples of future work, language may be used to quantify the structure of cognitive instability (e.g., running a rate of entropy for behavioral responses or mutual information among neural circuits) to make distinctions between adaptive cognitive flexibility and pathological volatility (either direction) (13). Educational systems could also utilize structured instability by creating a designed learning schedule guided by challenge and consolidation schedules that inherently replicate a bounded rhythm of cognitive adaptation (65).

4.4 Limitations

While the present results demonstrate that non-convergent learning trajectories can exhibit bounded structure and measurable informational signatures, they do not yet establish whether such dynamics correspond to progressive adaptive reorganization over longer timescales. An important unresolved question is whether structured instability merely reflects stationary bounded cycling, or whether it can also support higher-order adaptive reorganization analogous to phase transitions observed in phenomena such as grokking in deep neural networks. Distinguishing persistent non-equilibrium dynamics from genuinely evolving representational regimes remains an important direction for future work.

Several more specific limitations deserve mention. First, the multi-seed analysis uses 20 seeds per game rather than the preferable 30-50 seeds; while the unanimous 100%/0% cycling classifications and tight confidence intervals make additional seeds unlikely to alter the conclusions, a larger-scale replication would strengthen claims about attractor diversity in the Minority Game, where the 65/35 split warrants finer characterization of the two attractor classes. Second, the temporal windowed analysis demonstrates attractor evolution via per-window PCA and entropy metrics, but does not include full Poincaré sections per window; computing Poincaré sections requires selecting a transversal hyperplane specific to each game’s geometry, and is left for future work. Third, these observations are based on Q-learning with a decaying Boltzmann temperature; whether qualitatively similar adaptive reorganization occurs under constant-temperature or policy-gradient learning rules remains an open empirical question. Finally, all simulations use two-player matrix games; extending the framework to population-based learning and *n*-player games with richer strategy spaces is a necessary step toward establishing generality. These results suggest that adaptive learning systems may exhibit meaningful structure in regimes between strict equilibrium and pure randomness.

5. Conclusion

The concluding section integrates the paper's argument, which states that while equilibrium

convergence is a necessary component of the theory, it is not a sufficient explanation of bounded learners: that is, there must be a new notion of rationality which is viewed as dynamic rather than static, as adaptation occurring not by convergence but as structured within constraint. The conclusion gives some suggestions for future research in the area of combining information dynamics and bounded rationality, suggesting that the results may provide a more realistic and inclusive theory of adaptive flexibility both natural and artificial.

5.1 Reframing instability as structured chaos

The central thesis of this essay is that instability in repeated games—and, by extension, in adaptive learning systems—should not automatically be interpreted as evidence of failure (35). The traditional emphasis on equilibrium convergence as the hallmark of successful adaptation rests on an idealized conception of rationality, one that assumes agents possess sufficient information, computational capacity, and foresight to identify and maintain equilibrium behavior (55). Once these assumptions are relaxed, a different picture emerges. Real agents must operate under conditions of finite computation, incomplete information, and limited time, and their behavior reflects these constraints (44).

From this perspective, persistent instability need not be viewed as a defect to be eliminated. Rather, it may represent a natural consequence of bounded rational adaptation. The evidence reviewed in this essay suggests that many learning systems—whether human, artificial, or economic—do not simply progress toward static equilibrium states. Instead, they often exhibit ongoing processes of adjustment, feedback, and reorganization that remain structured without fully converging. Such behavior may be better understood as a form of bounded adaptive order, in which learning is expressed through continual reconfiguration rather than the attainment of permanent stasis (51).

Structured instability occupies an intermediate regime between equilibrium convergence and unconstrained chaos—a domain in which sensitivity, feedback, and bounded rationality interact to produce organized yet non-convergent dynamics (16). Such behavior may be viewed as the dynamical signature of systems that remain responsive to their environments. These systems continue to adjust, recalibrate, and reorganize over time, not because learning has failed, but because adaptation occurs under conditions of ongoing environmental change, incomplete information, and finite computational resources (31). Rather than converging to a permanently fixed state, they sustain a capacity for continued responsiveness, preserving flexibility in the face of uncertainty and changing conditions.

This reconsideration also changes how learning is evaluated (34). Rather than asking solely whether an agent converges, it encourages us to ask whether the agent maintains adaptive responsiveness under constraint: whether its fluctuations contain meaningful information (13), whether its cycles reflect continued exploration and adaptation, and whether its non-equilibrium dynamics remain organized rather than degenerating into randomness or instability (51). From this perspective,

successful learning need not be identified exclusively with convergence. In many environments, the capacity to remain responsive, flexible, and adaptable may be as important as the attainment of a fixed equilibrium. Convergence therefore becomes one possible outcome of learning rather than its only measure, while structured instability emerges as a potentially important mode of adaptation in complex environments characterized by uncertainty, feedback, and change (42).

5.2 The new world for game theory and informational dynamics

If bounded rationality produces structured chaos, then the next frontier is to merge the analytical precision of game theory with the quantitative techniques of information dynamics (1). Classical game theory presents elegant equilibrium concepts, but is a predominantly static endeavor (37). Information theory provides a powerful set of ways to measure uncertainty, entropy and mutual dependence (among others) but does not provide a behavioral foundation (52). True synthesis would provide a way of combining these efforts and establishing the examination of adaptive systems not as equilibrial payoffs but how they emerge as a consideration of their informational structure over time (51).

This synthesis begins by reconsidering the object of analysis itself. Rather than focusing exclusively on whether a learning process converges to a Nash equilibrium, researchers might instead ask whether it maintains informational coherence over time—that is, whether agent behavior continues to exhibit structured patterns of information exchange, predictive dependence, and adaptive responsiveness under constraint (30). From this perspective, information-theoretic measures such as entropy rate, transfer entropy, and mutual information provide a principled framework for characterizing the structure embedded within non-convergent dynamics. These measures make it possible to examine how information propagates through a system, how predictability evolves, and how coordination emerges and persists even in the absence of a fixed-point equilibrium (10).

From a modeling standpoint, future research should follow through with the simulation-based validation of these principles (46). Models like multi-agent Q-learning or replicator dynamics can be used to capture the information metrics in addition to traditional payoffs, while parameterizing agents with different bounds on cognition-e.g., limited memory, noisy perception, or limited computational steps (29). The resulting trajectories can be analyzed for informational ordering, and researchers can, through comparisons of simulations with alternative bounds to cognition, empirically map out how structured chaos emerges as a function of cognitive constraints. The endeavor would take game theory from being an equilibrium discipline to an informational dynamics discipline, where the empirical heft would move from static solving of outcomes to tracing the evolutionary trajectories of information bearing structures (65).

Empirical evaluation can also be extended to experimental economics and behavioral

science. Human-subject studies involving repeated strategic interactions—such as public goods games, matching pennies, coordination games, and related paradigms—provide opportunities to examine how individuals adapt under conditions of uncertainty, incomplete information, and time constraints. By analyzing behavioral time series using methods drawn from nonlinear dynamics and information theory, researchers can investigate whether human learning exhibits signatures consistent with structured instability, including bounded non-convergent dynamics, metastable patterns of adaptation, and persistent non-random fluctuations in strategy distributions (49). Such studies would provide a means of assessing whether the phenomena observed in computational learning systems also emerge in human adaptive behavior.

This line of inquiry points toward a broader synthesis that views strategic interaction not solely as a process of payoff optimization, but also as a process of information processing and adaptation. Within such a framework, equilibrium can be interpreted as one possible form of dynamical organization rather than the exclusive endpoint of learning; bounded rationality becomes a question of how agents adapt under informational and computational constraints; and stability encompasses not only fixed points, but also persistent and structured patterns of change. By combining the mathematical rigor of game theory with the analytical tools of information theory and dynamical systems, future research may be able to move beyond a simple opposition between order and disorder and instead investigate the rich intermediate regimes in which adaptive behavior is sustained (51).

More broadly, the transfer-learning results suggest the possibility of a tradeoff between exploitative convergence and future adaptive flexibility. In the experiments examined here, agents trained in strongly convergent environments often exhibited reduced exploratory behavior when transferred to adversarial settings, whereas adversarially trained agents generally retained broader policy support and adapted more readily to new adversarial tasks. One interpretation is that convergence can compress behavioral diversity in ways that facilitate short-term stability while limiting responsiveness to subsequent environmental change. Conversely, non-convergent learning dynamics may preserve residual information about exploration and adaptation that remains useful after transfer. Although the mechanisms underlying these effects remain to be established, the results are consistent with the possibility that structured instability contributes to the maintenance of adaptive capacity rather than representing merely transient noise in the learning process.

6. References

1. Fudenberg, D., & Levine, D. K. (1998). *The theory of learning in games*. The MIT Press.

2. Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex*

Systems, 11(1), 17–41. https://doi.org/10.1142/S0219525908001465

3. Berger, U. (2006). *Brown's original fictitious play*. *Journal of Economic Theory, 135*(1), 572–578. https://doi.org/10.1016/j.jet.2005.12.010

4. Bloembergen, D., Tuyls, K., Hennes, D., & Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research, 53*, 659–697. https://doi.org/10.1613/jair.4818

5. Bowling, M., & Veloso, M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence, 136*(2), 215–250. https://doi.org/10.1016/S0004-3702(02)00121-2

6. Camerer, C. F., Loewenstein, G., & Rabin, M. (Eds.). (2004). *Advances in behavioral economics*. Princeton University Press.

7. Camerer, C. F., & Ho, T.-H. (1999). Experience-weighted attraction learning in normal form games. *Econometrica, 67*(4), 827–874. https://doi.org/10.1111/1468-0262.00054

8. Cheung, Y.-W., & Friedman, D. (1997). Individual learning in normal form games: Some laboratory results. Games and Economic Behavior, 19(1), 46–76. https://doi.org/10.1006/game.1997.0544

9. Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature, 34*(2), 669–700. http://www.jstor.org/stable/2729218

10. Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley-Interscience.

11. Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters, 85*(2), 461–464. https://doi.org/10.1103/PhysRevLett.85.461

12. Nash, J. (1951). Non-Cooperative Games. Annals of Mathematics, 54(2), 286–295. https://doi.org/10.2307/1969529

13. Wolf, A., Swift, J. B., Swinney, H. L., & Vastano, J. A. (1985). Determining Lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena, 16*(3), 285–317. https://doi.org/10.1016/0167-2789(85)90011-9

14. Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of*

Economics, 69(1), 99–118. https://doi.org/10.2307/1884852

15. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

16. Sprott, J. C. (2003). *Chaos and time-series analysis*. Oxford University Press.

17. Galla, T., & Farmer, J. D. (2013). Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences, 110*(4), 1232–1236. https://doi.org/10.1073/pnas.1109672110

18. Tuyls, K., & Nowé, A. (2005). Evolutionary game theory and multi-agent reinforcement learning. *The Knowledge Engineering Review, 20*(1), 63–90. https://doi.org/10.1017/S026988890500041X

19. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.

20. Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics, 54*(2), 296–301. https://doi.org/10.2307/1969530

21. Hofbauer, J., & Sandholm, W. H. (2002). On the global convergence of stochastic fictitious play. *Econometrica, 70*(6), 2265–2294. https://doi.org/10.1111/1468-0262.00359

22. Leslie, D. S., & Collins, E. J. (2005). Individual Q-learning in normal form games. *SIAM Journal on Control and Optimization, 44*(2), 495–514. https://doi.org/10.1137/S0363012904441800

23. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., et al. (2019). Machine behaviour. *Nature, 568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

24. Cressman, R. (2003). *Evolutionary dynamics and extensive form games*. MIT Press.

25. Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). *Emergent tool use from multi-agent autocurricula*. arXiv. https://arxiv.org/abs/1909.07528

26. Smith, V. L., Suchanek, G. L., & Williams, A. W. (1988). Bubbles, crashes, and endogenous expectations in experimental spot asset markets. *Econometrica, 56*(5), 1119–1151. https://doi.org/10.2307/1911361

27. Hart, S., & Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica, 68*(5), 1127–1150. https://doi.org/10.1111/1468-0262.00153

28. Sanders, J. B. T., Farmer, J. D., & Galla, T. (2018). The prevalence of chaotic dynamics in games with many players. *Scientific Reports, 8*(1), 4902. https://doi.org/10.1038/s41598-018-22013-5

29. Sandholm, W. H. (2010). *Population games and evolutionary dynamics*. MIT Press.

30. Mertikopoulos, P., Papadimitriou, C. H., & Piliouras, G. (2018). Cycles in adversarial regularized learning. In A. Czumaj (Ed.), *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 2018)* (pp. 2703–2717). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975031.172

31. Sanders, J., Farmer, J. D., & Galla, T. (2018). Stochastic dynamics of learning in games: Theoretical and numerical results. *Journal of Economic Dynamics and Control, 91*, 338–357. https://doi.org/10.1016/j.jedc.2018.04.002

32. Lizier, J. T., & Prokopenko, M. (2010). Differentiating information transfer and causal effect. *The European Physical Journal B, 73*(4), 605–615. https://doi.org/10.1140/epjb/e2010-00034-5

33. Milionis, J., Piliouras, G., & Papadimitriou, C. H. (2023). No-regret dynamics and the impossibility of convergence to Nash equilibrium. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems, 36* (pp. 58410–58431). Curran Associates, Inc.

34. Sato, Y., & Crutchfield, J. P. (2003). Coupled replicator equations for the dynamics of learning in multiagent systems. *Physical Review E, 67*(1), 015206. https://doi.org/10.1103/PhysRevE.67.015206

35. Hommes, C. H. (2013). *Behavioral rationality and heterogeneous expectations in complex economic systems*. Cambridge University Press. https://doi.org/10.1017/CBO9781139208727

36. Crutchfield, J. P., & Young, K. (1989). Inferring statistical complexity. *Physical Review Letters, 63*(2), 105–108. https://doi.org/10.1103/PhysRevLett.63.105

37. Radner, R. (1975). Satisficing. *Journal of Economic Theory, 10*(2), 336–342. https://doi.org/10.1016/0022-0531(75)90004-1

38. Shapley, L. S. (1964). Some topics in two-person games. In M. Dresher, L. S. Shapley, & A. W. Tucker (Eds.), *Advances in game theory* (pp. 1–28). Princeton University Press.

39. Rubinstein, A. (1998). *Modeling bounded rationality*. MIT Press.

40. Sato, Y., Akiyama, E., & Farmer, J. D. (2002). Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences, 99*(7), 4748–4751. https://doi.org/10.1073/pnas.032086299

41. Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. Oxford University Press.

42. Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

43. Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139173179

44. Kandori, M., Mailath, G. J., & Rob, R. (1993). Learning, mutation, and long run equilibria in games. *Econometrica, 61*(1), 29–56. https://doi.org/10.2307/2951773

45. Smith, J. M. (1982). *Evolution and the theory of games*. Cambridge University Press. https://doi.org/10.1017/CBO9780511806292

46. Taylor, P. D., & Jonker, L. B. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences, 40*(1–2), 145–156. https://doi.org/10.1016/0025-5564(78)90077-9

47. Shalizi, C. R., & Crutchfield, J. P. (2001). Computational mechanics: Pattern and prediction, structure and simplicity. *Journal of Statistical Physics, 104*(3–4), 817–879. https://doi.org/10.1023/A:1010388907793

48. Mertikopoulos, P., & Sandholm, W. H. (2016). Learning in games via reinforcement and regularization. *Mathematics of Operations Research, 41*(4), 1297–1324. https://doi.org/10.1287/moor.2016.0778

49. Milionis, J., Papadimitriou, C., Piliouras, G., & Spendlove, K. (2023). An impossibility theorem in game dynamics. *Proceedings of the National Academy of Sciences, 120*(41), e2305349120. https://doi.org/10.1073/pnas.2305349120

50. Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

51. Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review, 85*(5), 1313–1326.

52. Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88*(4), 848–881.

53. Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 2672–2680). Curran Associates, Inc.

54. Peshkin, L., Kim, K.-E., Meuleau, N., & Kaelbling, L. P. (2000). Learning to cooperate via policy search. In C. Boutilier & M. Goldszmidt (Eds.), *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI 2000)* (pp. 489–496). Morgan Kaufmann

55. Gabaix, X. (2019). Behavioral inattention. In B. D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of behavioral economics: Applications and foundations 1* (Vol. 2, pp. 261–343). North-Holland. https://doi.org/10.1016/bs.hesbeh.2018.12.001

56. Kandori, M. (1992). Social norms and community enforcement. *The Review of Economic Studies, 59*(1), 63–80. https://doi.org/10.2307/2297925

57. Kantz, H., & Schreiber, T. (2003). *Nonlinear time series analysis* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511755798

58. Farmer, J. D., & Packard, N. H. (1986). The geometry of chaos. *Physica D: Nonlinear Phenomena, 22*(1–3), 187–201.

59. Lo, A. W. (2004). The adaptive markets hypothesis. *The Journal of Portfolio Management, 30*(5), 15–29. https://doi.org/10.3905/jpm.2004.442611

60. Crawford, V. P. (1995). Adaptive dynamics in coordination games. *Econometrica, 63*(1), 103–143. https://doi.org/10.2307/2951697

61. Ellison, G. (1993). Learning, local interaction, and coordination. *Econometrica, 61*(5), 1047–1071. https://doi.org/10.2307/2951493

62. Young, H. P. (1998). *Individual strategy and social structure: An evolutionary theory of institutions*. Princeton University Press.

Appendix A: Complete Simulation Figures

The following pages present the diagnostic analysis for the Battle of the Sexes game (section *4.1.5.1* in the main text). The companion Python code (structured_instability_simulations.py) reproduces all figures and can be extended to additional games and learning rules.

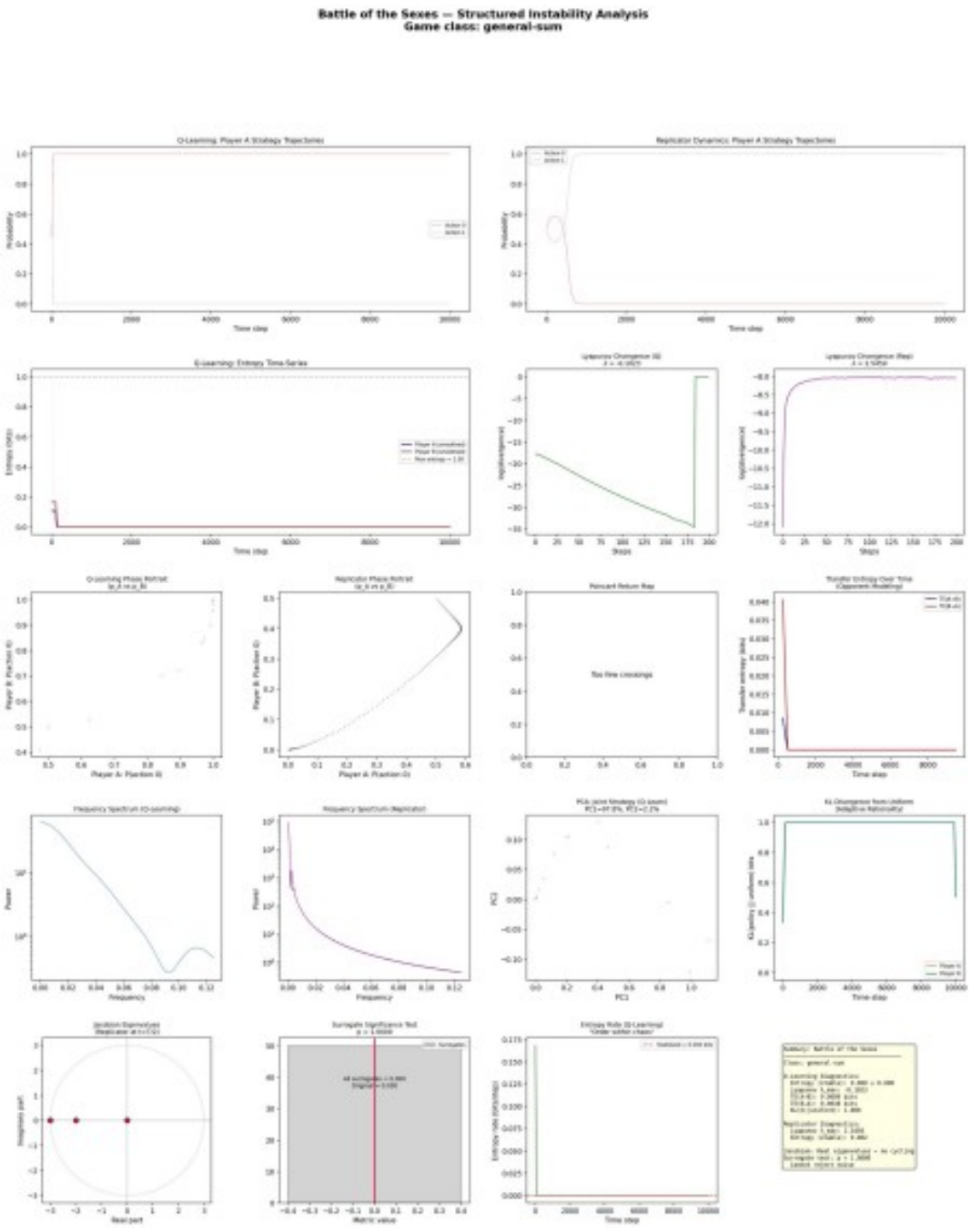

Figure A1. Battle of the Sexes: Full analysis showing convergence under both Q-learning and replicator dynamics.

Appendix B: Companion Code

The complete simulation code is provided as a self-contained Python script (structured_instability_simulations.py) requiring only numpy, scipy, and matplotlib, deposited at the GitHub repository. The code implements all learning algorithms, diagnostic tools, and visualization routines described in this paper. It can be run with: python structured_instability_simulations.py

The code is organized into five sections: (1) Game definitions for all eight games; (2) Learning algorithms (Q-learning with Boltzmann exploration and replicator dynamics with stochastic perturbation); (3) Analytical tools (Lyapunov exponent estimation, entropy rate, transfer entropy, KL divergence, PCA, FFT, Jacobian eigenvalue analysis, and surrogate testing); (4) Comprehensive simulation and analysis pipeline; and (5) Cross-game comparison and summary generation.

The link to the code is: https://github.com/pmcoder6105/structuredinstabilitysimulations