Q-LEARNING AGENTS IN THE EL FAROL BAR PROBLEM

Candidate Number: QRDG7
Department of Economics
University College London
London

ABSTRACT

The El Farol bar problem is a canonical game-theoretic problem for studying market dynamics under resource constraints, and it serves as a stylised representation of the challenges of coordination across domains. The rich emergent properties of the El Farol bar problem have previously been explored using evolutionary dynamics and heuristics-based strategies, and we hope to contribute to the literature by introducing reinforcement learning agents with varying levels of sophistication. In our simulations, the volatility of attendance of Q-learning agents decreased over time, resembling Arthur 1994's results, as agents adapted to the dynamic environment. Furthermore, through systematically varying the simulation parameters, we found that the 'memory' and the degree of exploration of agents are positively correlated with policy entropy: moderate exploration and memory introduced healthy diversity and stability, and excessively sophisticated learners often pushed the system into suboptimal regimes characterised by instability.

***K*eywords** Reinforcement Learning · Agent-Based Modelling · Game Theory · Computational Economics

1 Introduction

> "For a solitary animal egoism is a virtue that tends to preserve and improve the species: in any kind of community it becomes a destructive vice."
>
> —Erwin Schrodinger, *What is Life?*

Self-interest lies at the heart of economics, and it has characterised the modern paradigms of economic schools of thought. Adam Smith, the father of economics, claimed that 'It is not from the benevolence of the butcher, the baker, or the brewer that we expect our dinner, but from their regard to their own interest.' Though Smith claims that efficient outcomes can be achieved through the pursuit of self-interest, Hobbes was more cynical towards the 'state of nature'. He suggested that a society of egoists, without the governance of a central authority, will condemn life to be 'solitary, poor, nasty, brutish, and short'.

In line with Smith's beliefs, neoclassical models endowed agents with full rationality, complete information, giving rise to homo economics, the economic man. With the homo economicus at the heart of economic theory, it is commonly assumed that agents are completely rational, making decisions based on the maximisation of utility functions. However, the reductionist elegance of these models owes more to mathematical convenience than to realism. Heterodox schools therefore embrace bounded rationality, modelling agents who learn and adapt inductively in non-stationary environments. The concept of equilibrium is another point of contention among the neoclassicalists, Austrians, and complexity economists; in the place of a static, steady-state system, they envision an inherently dynamic, non-equilibrium market in constant flux. Though the general equilibrium framework has laid a formal foundation, the most compelling representations of the economy come from complexity science.

The El Farol bar problem was first introduced in Arthur 1994, and it has become a canonical model for studying congestion, adaptation, and coordination in multi-agent systems. The El Farol bar problem and the generalised Minority Game are stylised representations of market mechanics, and illustrates the inductive nature of decision-making observed empirically in agents. Through the seemingly simple decision of choosing whether or not to attend a bar on a given night, the model provides a rich illustration of the emergence of the system's self-organising equilibrium from micro-level interactions. This begs the following questions: Can boundedly rational and self-interested agents self-organise into efficient attendance patterns without central control? In addition to efficiency, how equitable is the distribution of rewards under resource constraints?

To capture agents' inductive adaptation in a principled way, we integrate a reinforcement learning mechanism into the El Farol bar problem. In particular, we employ Q-learning—a simple, model-free algorithm that estimates the optimal action-value function $Q(s, a)$ by trial-and-error updates—because it naturally aligns with bounded rationality and discrete choice dynamics. Q-learning requires no prior model of attendance transitions and can flexibly accommodate varying memory lengths, learning rates, and exploration strategies. By tuning the discount factor γ and exploration parameter (e.g. temperature or ϵ), agents balance the exploitation of high-reward attendance predictions against the exploration needed to track a non-stationary environment. This approach thus preserves the El Farol bar problem's hallmark of emergent, self-organising coordination while endogenously learning forecasting heuristics from experience rather than imposing them analytically.

This dissertation extends Arthur 1994 original El Farol bar problem by replacing the original heuristic-based forecasting framework with tabular Q-learning agents with tunable parameters. We show that the Q-learning agents preserves the model's ability to reproduce the characteristic damping of attendance fluctuations, followed by a convergence to an 'ϵ-neighbourhood' around the threshold, while enabling a systematic exploration of how memory length, learning rate and exploration strategy affect emergent dynamics. Additionally, we analysed and quantified the equity of the system by computing the Gini coefficient and Jain's index. Our key contribution is a reinforcement-learning framework that derives parameters endogenously through agent interaction, unlike Marsili and Challet 2001 who calibrated their agent-based model analytically. Our approach offers a more granular view of how decentralised reinforcement learning drives coordination and phase changes in complex adaptive systems.

This paper proceeds as follows. Section 2 provides a formalisation of the El Farol bar problem, alongside a game-theoretic solution to the El Farol bar problem. A literature review of learning-based algorithms applied to the El Farol bar problem and is provided in Section 3, followed by a discussion of the complexity of the El Farol bar problem, and the emergent phenomena which subsequently arises in Section 4. In Section 5, we introduce our methodology and the implementation of reinforcement learning in the El Farol bar problem. Details of the results are explained in Section 7, which addresses the stochastic nature of simulation-based analysis, ensuring that external validity is achieved through obtaining robust results. Finally, we conclude with a discussion of our results, and we hope to shed light on the significance and the implications of the El Farol bar problem.

2 The El Farol bar problem

The El Farol bar problem can be modelled as an infinitely-repeated two-action stochastic game in which each agent chooses whether to attend a bar on any given evening and receives a payoff based on whether attendance exceeds capacity. If attendance does not exceed the threshold, attendees gain utility; if it does, they suffer a poor outcome and staying home will be a dominant strategy, which in that case yields a higher payoff.

Definition 2.1 (Stochastic Game). *A stochastic game is a tuple*

$$\big(n,\, \mathcal{S},\, \mathcal{A},\, P,\, \mathcal{R},\, \gamma\big),$$

where

(a) n *is the number of agents;*

(b) $\mathcal{S}$ *is the state space;*

(c) $\mathcal{A} =_{i=1}^{n} \mathcal{A}^i$ *is the joint action space, where* $\mathcal{A}^i$ *is the set of actions available to agent* i*;*

(d) $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \;\to\; [0, 1]$ *is the state–transition probability function;*

(e) $\mathcal{R} = \big(R^1, \ldots, R^n\big)$ *is the reward function profile, with*

$$R^i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \;\to\; \mathbb{R}$$

specifying the reward to agent i*;*

(f) $\gamma \in [0, 1)$ *is the discount factor, which balances immediate and future rewards.*

The El Farol bar problem and the Minority Game have been extensively studied using analytical game theory models. These approaches provide a baseline for understanding what outcomes might be expected if agents acted in a deductive, rational manner, or if the system could be solved like a traditional game. However, the ill-defined nature of the El Farol bar problem necessitates inductive solutions, and methods in this camp highlight the contrast between deductive solutions and the adaptive outcomes observed with learning agents.

Definition 2.2 (El Farol Stage Game)**.** *The El Farol stage game is a one-stage, simultaneous-move stochastic game defined by the tuple*

$$\Gamma \;=\; \big(n,\; \{s_0\},\; \mathcal{A},\; P,\; R,\; \gamma\big),$$

where

(a) n *is the number of agents;*

(b) *the state space is the singleton* $\{s_0\}$*;*

(c) *the joint action space is*

$$\mathcal{A} \;=\; {}_{i=1}^{n}\mathcal{A}^i \quad \textit{with} \quad \mathcal{A}^i = \{0,1\};$$

(d) *the transition kernel is*

$$P(s_0 \mid s_0, a) = 1 \quad \forall\, a \in \mathcal{A};$$

(e) *the reward functions* $R = (R^1, \ldots, R^n)$ *satisfy, for each* i*,*

$$R^i(s_0, a, s_0) = \begin{cases} G, & a_i = 1 \textit{ and } \sum_{j \neq i} a_j < C, \\ B, & a_i = 1 \textit{ and } \sum_{j \neq i} a_j \geq C, \\ S, & a_i = 0, \end{cases} \qquad G > S > B,\; C \in \mathbb{Z}^+;$$

(f) $\gamma \in [0, 1)$ *is the discount factor (irrelevant for this single-stage game).*

We can show that there is no symmetric pure strategy Nash equilibrium: if all agents use the same pure strategy (i.e. all go or all stay), the system will yield inefficient outcomes in the form of full attendance, or vice versa. However, there exist many asymmetric pure-strategy equilibria where a specific subset of players always go and the others always stay, where the population is split into the proportion (number of agents - threshold: threshold, i.e. 40:60). This will not yield profitable deviations for agents on either side, as any unilateral deviations from each subset will result in a worse payoff. For the 60 that are attending, if they decide to not attend then they will miss out on an enjoyable night as the bar is not yet congested, and those that are not attending will be at the tipping point, so not attending will be more favourable than attending.

	State	
	Uncrowded	Crowded
Go to the bar	G	B
Stay at Home	S	S

where $G > S > B$

Figure 1: State-Dependent Payoff for Player i in the El Farol stage game.

However, these asymmetric equilibria require players to somehow agree on who goes and who stays, making them unlikely without coordination. This formal game-theoretic result underscores why El Farol bar problem is challenging: the mixed equilibrium achieves the efficient outcome on average but assumes rational randomising agents, whereas in reality, agents must learn or negotiate their roles. Similarly, the Minority Game has a unique symmetric Nash equilibrium in mixed strategies and a multitude of Nash equilibria in pure strategies where the minority condition is met by different groupings of players. These static equilibria set a reference point – for instance, the 50/50 randomising strategy in Minority Game or 60% attendance in El Farol bar problem – but they do not explain how agents might arrive at those outcomes.

Therefore, we have shown that there does not exist a pure strategy Nash equilibrium, and the asymmetric pure strategy Nash equilibrium will be invalid due to the structure of the game. The difficulty in achieving efficiency in the El Farol bar problem hinges on the task of coordination without communication. Given the ability to communicate, agents may explicitly collude through the use of contracts to decide the pattern of attendance.

Definition 2.3 (Nash equilibrium in a stochastic game)**.** *Consider a stochastic game given by the tuple*

$$\big(\mathcal{S},\, \mathcal{A} = (\mathcal{A}^1, \ldots, \mathcal{A}^n),\, P,\, \mathcal{R} = (\mathcal{R}^1, \ldots, \mathcal{R}^n),\, \gamma\big),$$

where $\mathcal{S}$ is the state space, each $\mathcal{A}^i$ is the action set of player i, P the state-transition kernel, each $\mathcal{R}^i$ the reward function of player i, and $\gamma \in [0, 1)$ the discount factor. A joint policy

$$\pi^* = (\pi^{1,*}, \pi^{2,*}, \ldots, \pi^{n,*})$$

is called a Nash equilibrium (L. Zhang and Liu 2021) if for every agent i and every state $s \in \mathcal{S}$,

$$V_i^{\pi^*}(s) \;\geq\; V_i^{\left(\pi^i, \pi^{-i,*}\right)}(s) \quad \textit{for all alternative policies } \pi^i,$$

i.e. no single player can improve its expected discounted return by unilaterally deviating from π^*.

The El Farol stage game admits a unique symmetric mixed-strategy equilibrium in which each player randomises their decision with a certain probability that is consistent with the indifference condition, namely, when their expected utility from attending is equalised to their expected utility from not attending.

$$\sum_{k=0}^{C-1} \binom{N-1}{k} p^k\,(1-p)^{N-1-k} = \sum_{k=C}^{N-1} \binom{N-1}{k} p^k\,(1-p)^{N-1-k},$$
$$p \approx \frac{C}{N}.$$

so that on average C agents attend each round. At this equilibrium, on average, the bar is at capacity, and no individual can improve their payoff by unilaterally deviating. Under this mixed-strategy Nash equilibrium (MSNE) each player attends with probability $p = 0.6$, so total attendance $A \sim \mathrm{Binomial}(n, p)$ satisfies

$$\mathbb{E}[A] \;=\; n\,p \;=\; 100 \times 0.6 \;=\; 60, \qquad \mathrm{Var}[A] \;=\; n\,p\,(1-p) \;=\; 100 \times 0.6 \times 0.4 \;=\; 24,$$

where the attendance is expected to converge asymptotically to 60, but the variance will manifest in random fluctuations around the target threshold, which is observed in the simulations. This is inefficient in that the bar-goers will regularly over-attend or under-attend, and rewards will be consistently lower than the pure strategy Nash equilibrium.

We extend the analysis of the El Farol Stage Game to the El Farol bar problem. In particular, consider the infinitely repeated El Farol bar stage game with $n = 100$ agents.

Definition 2.4 (El Farol bar problem)**.** *The El Farol bar problem is the infinitely repeated version of the El Farol stage game Γ. Formally, it is the stochastic game*

$$\overline{\Gamma} = \big(n,\; \mathcal{S},\; \mathcal{A},\; P,\; R,\; \gamma\big)$$

with the same components as Γ, played at each period $t = 0, 1, 2, \ldots$. At each t, every agent i selects action $a_t^i \in \{0, 1\}$ and receives reward

$$R^i(s_0, a_t, s_0),$$

with their long-run payoff given by the discounted sum $\sum_{t=0}^{\infty} \gamma^t\, R^i(s_0, a_t, s_0)$.

In the setting of infinitely repeated games, a fundamental theorem from game theory, the Folk Theorem, guarantees that any feasible payoff profile exceeding the one-shot Nash equilibrium payoffs can be sustained as a subgame-perfect equilibrium, yielding an unbounded family of possible outcomes.

Definition 2.5 (Folk Theorem in the El Farol bar problem)**.** *In an infinitely repeated stage game with discount factor γ sufficiently close to 1 and with credible punishment strategies available, any feasible payoff profile that strictly exceeds the one-shot Nash equilibrium payoffs can be achieved as a subgame-perfect equilibrium outcome. Consequently, players can coordinate on strategy profiles delivering higher long-run returns than the stage-game Nash equilibrium.*

The ill-defined nature of the El Farol bar problem was capitalised by Arthur to illustrate the shortcomings of deductive reasoning, which is the norm for general equilibrium models, when the fundamental assumptions of full rationality, complete information, among others, are hard to fulfill. Instead, agents are designed to be boundedly rational, and Arthur emphasised that agents in this problem cannot rely on rational expectations as there is no unique deductive solution; any successful or profitable strategy will soon be oversubscribed, and it is fruitful for agents to follow successful strategies to an extent, which can be interpreted as herd behaviour.

In a complex adaptive system, there is no set of stable equilibrium strategies, and any successful strategy will fail to remain profitable as the saturation of a successful strategy arbitrages unrealised profit. This is analogous to how the market price converges to the marginal cost in Bertand and Cournot competition as the number of firms approaches infinity. Since a market is a complex adaptive system, there exists uncertainty that is intrinsic to the system. For example, when a strategy becomes profitable, the convergence of adoption by agents will reduce the profitability of the strategy, and this will affect the distribution of strategies employed by all agents. Agents' strategies form an evolving ecology where profits are arbitraged away. Entropy comes into play as it measures the amount of disorder in the system, and it can be employed to measure the variation in the dispersion of the possible strategies.

Challet, Marsili, and Ottino 2004 developed a mathematical reformulation of the El Farol bar problem and demonstrated a deep connection to the Minority Game. They showed that under general conditions, even zero-intelligence agents can induce the average attendance to converge to the bar's capacity, implying that complex reasoning is not strictly required for achieving the 60/40 split in the long run. By specialising their model, they mapped the El Farol bar problem onto the well-known Minority Game and leveraged tools from statistical physics. Through this analysis, they derived a phase diagram that completely characterises the model's behavior. A crucial insight from their work and related Minority Game studies is the presence of phase transitions: for example, as the diversity of agents' strategies or the information available to agents changes, the system can shift from a symmetric phase (uncoordinated, high fluctuations) to a symmetry-broken phase where outcomes become predictable and coordination improves. In Minority Game terms, below a critical point agents essentially randomise, while beyond that point, agents start to correlate their actions. This formal treatment was one of the first to show that complex adaptive dynamics can be understood in terms of equilibrium phase transitions, giving a theoretical backbone to the patterns observed in simulations. It connects El Farol bar problem/Minority Game to the broader field of complex systems by showing how macroscopic regularities emerge from simple rules, and how these regularities can change dramatically with system parameters.

3 A Different Paradigm of Learning

Many theories of learning from psychology and neuroscience have inspired learning algorithms in artificial intelligence, particularly reinforcement learning. Reinforcement learning agents learn optimal policies through interacting with the environment, and receive rewards associated with the policy based on its success. In short, reinforcement learning agents engage in trial and error. This approach is less structured than paradigms such as supervised and unsupervised learning. For instance, Pavlovian conditioning and the experimental work of Kahneman and Tversky 1979. illustrate how boundedly rational decision-making processes operate. Advocating for a new paradigm, Arthur proposed modeling agents' decision-making with inductive reasoning, using an evolutionary-inspired methodology where successful strategies received more weight relative to other strategies, and thus became more prominent in an agent's repertoire. This heuristic approach contrasts with the strict deductive reasoning assumed in traditional models. Early works (e.g. Franke 2003; Whitehead 2008) showed that simple reinforcement learning dynamics allow El Farol bar problem agents to approach the optimal attendance on average, albeit with Nash equilibria that are deemed inequitable wth regards to the reward distribution.

An interesting theoretical result by Zambrano 2004 addresses the relationship between boundedly rational learning and game-theoretic equilibrium in El Farol bar problem. Zambrano modeled the El Farol bar problem as a game and computed its mixed-strategy Nash equilibria, then proved that Arthur's inductive learning scheme converges to the same distribution of outcomes as those equilibria. In particular, he showed that the empirical frequency of bar attendance in Arthur's simulation matches the distribution implied by the mixed Nash equilibrium. This explains why Arthur's simulation yields seemingly random attendance from week to week, even though each agent uses deterministic prediction strategies – essentially, the ecology of strategies in the simulation acts as if each agent were randomising with a certain probability. Thus, inductive reasoning led the system to an equilibrium-like state. This work provides a bridge between the game-theoretic view and the learning view: it suggests that the outcome of decentralised adaptation can coincide with a correlated equilibrium of the underlying game. However, one should note that while the aggregate statistics match, the process by which it's reached is dynamic and path-dependent.

Zheng et al. 2023 introduced tabular Q-learning agents into the Minority Game. They demonstrated that learning agents can self-organise into an optimal resource allocation regime, effectively "solving" the Minority Game. Notably, when agents appropriately balance exploration with exploitation, the population converges to the optimal attendance where half the agents choose each option, maximising resource use. If agents veer toward purely greedy or purely random behaviors, only partial coordination or even anti-coordination emerges. A key finding is the role of moderate exploration: it helps agents escape cyclically suboptimal patterns and achieve global optimum coordination. The end state is characterised by a symmetry-breaking in action preferences – roughly half the agents consistently choose one action and half the other, yielding a stable equilibrium with efficient resource use. This symmetry-broken optimal allocation is robust to population size and other parameters, highlighting that basic Q-learning can solve the Minority

Game without pre-programmed strategies. Beyond the Minority Game, this work argues that such a reinforcement learning paradigm could help explain how decentralised agents solve resource allocation puzzles in economics.

S.-P. Zhang et al. 2023 combined the Minority Game with reinforcement learning agents to examine collective outcomes. They found that as tabular Qlearning agents iteratively adapt to maximise their individual payoffs, the global system state moves toward optimum resource allocation under certain conditions. Interestingly, the learning dynamics produce a self-organising oscillatory behavior that suppresses herding and overcrowding: instead of all agents swinging in unison, the population settles into an alternating attendance pattern-akin to a period-2 cycle-that keeps attendance near the desired capacity. The authors report evidence of a first-order phase transition in the system's collective behavior as key parameters such as learning rate or exploration rate vary. In one regime, the system self-organises into an optimal, stable oscillation; beyond a critical threshold, the coordination pattern breaks down abruptly – a hallmark of first-order transitions. They also introduce analytical tools to detect the emergence of these period-two oscillations. This study underscores that multi-agent reinforcement learning can exhibit rich emergent dynamics while driving the system toward high efficiency.

Chung, S. Chow, and Tumer 2018 tackled a multi-agent learning variant of the El Farol bar problem by extending it to a Multi-Night bar problem. Instead of one crowded bar night, there are multiple nights where each had its own capacity, and 100 agents had to choose which night to attend, learning via Q learning. A novel contribution is a probabilistic scheduling of learning: not all agents update their Q-tables every round. Instead, each agent has a probability of learning at a given time, which increases if its impact on reward is small. By staggering the learning updates this way, they reduced "noise" from simultaneous policy changes. The result was a faster and more stable convergence to optimal attendance distribution across nights. They also employed the difference reward technique to better align individual learning with the global objective. This work is relevant for multi-agent coordination in that it shows how modifying the learning process itself can dramatically improve collective outcomes in congestible resource scenarios. It demonstrates a practical approach to scaling El Farol bar problem-like problems to larger, more complex settings while maintaining coordination efficiency.

Whitehead 2008 re-framed the El Farol bar problem explicitly as a repeated potential game with reinforcement learning dynamics, proving that in the long run the population segregates into two groups: those who almost always go vs. those who almost always stay home– effectively a learned equilibrium that "sorts" the population into consistent behaviors. This sorting result aligns with the idea of symmetry-breaking observed in later reinforcement learning studies. Additionally, Franke 2003 applied basic reinforcement learning rules in El Farol bar problem and found that agents' attendance decisions do fluctuate around the bar's capacity, but the steady-state distribution of attendance frequencies under learning is quite different from the game's multiple Nash equilibria. In other words, reinforcement learning leads to a stable stochastic outcome that standard equilibrium analysis wouldn't predict. These earlier findings paved the way for recent research by highlighting that simple learning rules can solve the El Farol bar problem "coordination puzzle," albeit with dynamics and equilibria that are non-trivial. Current works build on this by using more formal analysis, as seen in S.-P. Zhang et al. 2023, or by harnessing more sophisticated reinforcement learning methhods to achieve coordination more efficiently.

We distinguish our approach by endowing each Q-learning agent with a tunable memory length M that directly governs how many past "crowded/not crowded" outcomes they condition on. Whereas prior reinforcement learning or propensity-based methods typically react only to the last round or the attendance figure, or use a fixed small predictor set. Our agents build Q-tables indexed by the last M bits of attendance history. By increasing M, agents can detect and exploit longer-range patterns or cycles in bar attendance—something simple one-step learners cannot—yet remain adaptive and tractable.

Moreover, although the Minority Game and the El Farol bar problem share the core idea of competing for a scarce resource, their emergent dynamics differ. In the El Farol bar problem, attendance tends to oscillate around the capacity threshold in quasi-periodic cycles whose typical length grows with M. In contrast, the Minority Game exhibits a statistical "phase transition" in volatility as the ratio

$$\alpha = \frac{2^M}{N}$$

crosses a critical value: too little or too much information yields high fluctuations and predictability, while at α_c agents collectively reach an efficient, unpredictable regime. By layering memory-parameterised Q-learning onto the El Farol bar problem, we combine the oscillatory richness of El Farol with the adaptive coordination of Minority Game, showing that moderate M optimises both stability and efficiency before overfitting noise reverses those gains.

4 Complexity and Emergent Phenomena in the El Farol bar problem

Emergence characterises complex systems, where an "entity has properties or behaviors that its parts do not have on their own, and emerge only when they interact in a wider whole." The concept of emergence is inherently incompatible with the philosophy of neoclassical models, where aggregate behaviour is taken to be the sum of individual behaviour, and where analytical solutions are derived at the expense of realism. We therefore turn to agent-based models to capture these interactive complexities. This bottom-up approach differs from the top-down approach of neoclassical models: agents are heterogeneous and free to interact with other agents under simple instructions, crudely simulating how a real market operates with boundedly rational individuals (Tesfatsion and Judd 2006; Farmer and Foley 2009). Furthermore, agent-based models treat shocks as endogenous, emergent behaviors arising from microscopic interactions, offering a perspective to study how local decisions collectively and endogenously produce bubbles, crashes, and other macroscopic phenomena (LeBaron 2006).

The El Farol bar problem is widely studied for its rich emergent behaviour that arises from the interactions between the many learning and adapting agents. Rather than settling into a static equilibrium, these complex systems exhibit ongoing fluctuations, collective oscillations, or abrupt transitions between regimes of behavior, i.e. markets may oscillate between periods of clustered volatility. Of all complex systems, financial markets perhaps exhibit the most complex dynamics. Contrary to physical complex systems, where the individual interacting units are inanimate particles, the subunits of interest in financial markets are conscious beings, influenced by social and psychological factors.

As S.-P. Zhang et al. 2023 demonstrate, even the simplest reinforcement-learning rules can generate rich, non-trivial temporal patterns at the collective level. In a reinforcement learning-driven Minority Game, for example, attendance does not settle on a fixed value each round but instead self-organises into a persistent oscillation—alternating between slight under- and over-capacity in the following periods. This period-two cycle emerges without any external coordinator or forced scheduling and it effectively forestalls the herding that would otherwise lead most agents into the majority and guarantee failure. Such oscillatory dynamics echo classic predator–prey or evolutionary cycles, and in the context of the El Farol bar problem they reflect a tacit agreement among agents: they take turns occupying the minority or majority, thereby sustaining an efficient use of the shared resource. However, in more realistic simulations where parameters which dictates the dynamics of the system, such as the number of agents or the threshold, it is more difficult for agents to tacitly collude and self-organise into a cyclical pattern.

A related body of work highlights that multi-agent adaptive systems exhibit genuine phase transitions as key parameters shift. In the original Minority Game, there is a critical ratio of strategy diversity to player count at which attendance variance is minimised—below this threshold the system is efficient and symmetric, while above it a biased, less efficient phase takes over. Zheng et al. 2023 show that an intermediate exploration rate unlocks near-optimal coordination, whereas too little or too much exploration drives the population into partial coordination or outright anti-coordination. S.-P. Zhang et al. 2023 go further, identifying a first-order transition in learning parameters where the system abruptly jumps from a stable, coordinated regime to a volatile, uncoordinated one. This notion of criticality will be of interest, as we will explore the sensitivity and implications of varying model parameters. The interpretation of such critical parameters extends beyond markets, and it also applies to domains such as urban planning and the internet, where the bottleneck often arises due to an inefficient allocation of resources.

The bar-attendance dynamics are often contrasted to models of financial markets or traffic flow, where collective outcomes emerge from agents trying to predict each other, and the inherent asymmetry which arises from their mutual predictions that produces outcomes of interest. The El Farol bar problem's oscillatory solution is analogous to a self-organised division of resources seen in animal foraging or network congestion problems. Some recent work in multi-agent reinforcement learning more broadly has explored social welfare and emergent cooperation in Matrix games and Markov games.

Emergent dynamics are also influenced by the diversity in agent behaviors and limitations. Agents with varying levels of rationality or intelligence, along with other parameters, have been explored previously, and this heterogeneity in agent parameters can introduce varying levels of efficiency in adapting to the equilibrium, where sophisticated agents display more promising signs of discovery. Other works show a vested interest in the long-run dynamics of the system, employing methods similar to replicator-mutator dynamics or Fokker-Planck equations to describe the time evolution of the probability distribution of the macro-variable.

5 Methodology

In this section, we present the reinforcement learning mechanism through which agents inductively learn to decide whether to attend the bar or not. This essentially boils down to a sophisticated forecasting exercise-sophisticated as the system is non-stationary. This is analogous to a central banker trying to predict the rate of unemploment: they are engaged in a feedback loop where their policies affect the expectations of the general public, which in turn affect the efficacy of their policies. As the system adapts to new information and actions, the convergence guarantees of the Q-learning algorithm originally established by Watkins and Dayan 1992 are invalidated, as it assumes that agents operate within a stationary Markovian environment. Despite the lack of formal convergence proofs, Q-learning remains a powerful, simple heuristic, often yielding approximate equilibria.

Our model applies Q-learning to an agent-based simulation of the El Farol bar problem. The learning of optimal action–state pairs, namely the "quality" of each state–action pair, is facilitated by the Q-function, which maintains a table of Q-values corresponding to all feasible state–action combinations. Each Q-value is updated via trial-and-error learning, a fundamental process in both intelligent and non-intelligent systems.

State $s \in \{0,1\}^m$	**Stay Home** $(a = 0)$	**Attend** $(a = 1)$
$0\,0 \ldots 0$		
$0\,0 \ldots 1$		
$0\,0 \ldots 0, 1$		
$\vdots$	$\vdots$	$\vdots$
$1\,1 \ldots 1$		

Table 1: Q-Table for memory length m

The independent Q-learning agents operate under the partially observed markov decision process (POMDP) framework, which is the multi-agent extension of the standard markov decision process framework, accounting for the partial observability of states and actions in general-sum stochastic games.

Definition 5.1 (Partially Observable Markov Decision Process). *A* Partially Observable Markov Decision Process *is a tuple*

$$\mathcal{P} \;=\; \big(\mathcal{S},\; \mathcal{A},\; \mathcal{O},\; T,\; Z,\; R,\; \gamma\big),$$

where

- $\mathcal{S}$ *is a (finite) set of* states*;*
- $\mathcal{A}$ *is a (finite) set of* actions*;*
- $\mathcal{O}$ *is a (finite) set of* observations*;*
- $T \colon \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ *is the* state-transition kernel*, with*
$$T(s' \mid s, a) \;=\; \Pr\big(S_{t+1} = s' \mid S_t = s,\; A_t = a\big);$$
- $Z \colon \mathcal{S} \times \mathcal{A} \times \mathcal{O} \to [0,1]$ *is the* observation kernel*, with*
$$Z(o \mid s', a) \;=\; \Pr\big(O_{t+1} = o \mid S_{t+1} = s',\; A_t = a\big);$$
- $R \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ *is the* reward function*, so that taking action a in state s yields expected reward $R(s,a)$;*
- $\gamma \in [0,1)$ *is the* discount factor *weighting future rewards.*

At each time step t, the agent chooses $A_t \in \mathcal{A}$, the environment transitions from S_t to S_{t+1} according to T, emits an observation O_{t+1} via Z, and the agent receives reward $R(S_t, A_t)$.

Learning via Q-values is a departure from the evolutionary-based approach of Arthur. In Arthur (1994), agents are all initialised with a set of strategies, and the propensity of playing that strategy is updated through employing various strategies, assessing their profitability, then the Q-value is updated accordingly. In our approach, agents are initalised with a Q-table, with uniformly random values, which provides a source of heterogeneity amongst agents, and the Q-values are updated with the temporal difference update rule:

$$Q_{t+1}(s_t, a_t) \;=\; Q_t(s_t, a_t) \;+\; \alpha\big[r_t + \gamma \max_{a'} Q_t(s_{t+1}, a') \;-\; Q_t(s_t, a_t)\big].$$

Granted the Q-learning agents have learned the optimal solution, their optimised Q-function, Q*, must then satisfy the Bellman optimality equation:

$$Q^*(s,a) \;=\; R(s,a) \;+\; \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) \max_{a' \in \mathcal{A}} Q^*(s', a').$$

The Q-values of the executed strategies—e.g. attending when the recent attendance history was $\mathbf{h} = (1, 1, 0)$ (comfortable two weeks ago, crowded three weeks ago)—are updated each iteration. Strategies that yield positive reward r_t are reinforced with larger Q-values and hence have higher selection probability upon revisiting the same state s.

Our main contribution is to introduce an extended memory parameter M and study its effect on predictive performance. We conjecture the existence of an optimal memory length M^* at which agents best capture the temporal dynamics without overfitting: too small M misses long-run cycles, whereas too large M locks agents into outdated patterns. Reinforcement learning relies heavily on the Markov property, and this is embedded in the states of each agent's Q-table, where we assume that the attendance decision for the current state is solely conditional on the previous state. Since the optimal forecasting rule goes beyond a one-period forecast, our inclusion of memory encapsulates and condenses the relevant information into one state, preserving the validity of the Markov property.

Definition 5.2 (Markov Property)**.** *A stochastic process $\{X_t\}_{t\geq 0}$ (or, in a controlled setting, $\{(S_t, A_t)\}$) satisfies the* Markov property *if the conditional law of the next state and reward depends only on the current state–action pair, and not on the earlier history. In the context of a Markov decision process with state S_t, action A_t, next state S_{t+1}, and reward R_{t+1}, this reads*

$$\Pr\bigl(S_{t+1} = s',\, R_{t+1} = r \mid S_t = s,\, A_t = a,\, H_t\bigr) = \Pr\bigl(S_{t+1} = s',\, R_{t+1} = r \mid S_t = s,\, A_t = a\bigr),$$

where $H_t = (S_{t-1}, A_{t-1}, \dots, S_0, A_0)$ denotes the full history up to time $t - 1$.

Furthermore, we enhance the exploration–exploitation mechanism by complementing the traditional ϵ-greedy rule with a softmax policy. At time t, the agent selects action a in state s with probability

$$\pi_t(a \mid s) \;=\; \frac{\exp\bigl(Q_t(s,a)/\tau_t\bigr)}{\sum_{a'} \exp\bigl(Q_t(s,a')/\tau_t\bigr)},$$

where the inverse-temperature τ_t may be annealed or adapted to the sample variance of recent Q-values. This yields a smooth transition from exploration (τ_t large) to exploitation ($\tau_t \to 0$), avoiding the abrupt schedule $\epsilon \leftarrow \epsilon \cdot \kappa$ often used in ϵ-greedy, and typically enjoying stronger convergence in the pure-strategy limit.

Contrary to Zheng et al. 2023, where the states are encoded as the number of attendees on any given night, information about prior attendance and the prior states is encoded as binary strings of length M, with each binary digit representing whether or not a state was congested or overcrowded. By extending the availability of information, this allows for further sophistication while retaining fidelity to the inductive cognitive reasoning process proposed by Arthur. Our approach can still be interpreted as heuristic, as agents have a notion or intuition of the 'quality' of certain states, for example, commuters often have an intuitive notion of how congested roads may be depending on the time of day, representng the different states in this context, and often decide to take alternative paths during peak hours or vice versa.

Each agent maintains a Q-table of action–value estimates $Q(s, a)$ over the 2^m possible histories when its memory length is m. Hence the table size grows as

$$\mathcal{O}\bigl(2^m\bigr),$$

independent of the number of agents. This approach differs from Zheng et al. 2023, where the states consists of the attendance A on any given night, and the table grows at

$$\mathcal{O}\bigl(n\bigr),$$

where millions of states must be maintained for hyper-realistic, large-scale simulations (see Kaszowska-Mojsa and Pipień 2023). Additionally, many states are redundant, and are likely to remain sparse since many state–action pairs are never visited. In our implementation, m is unlikely to exceed 2^4 states as attendance efficiency deteoriates past $m = 4$ (see 5, 6). Studies of the Minority Game show that the resulting attendance distribution often converges to either a unimodal or a bimodal shape.

We quantify the dispersion of attendance decisions at time t via the Shannon entropy

$$H_t \;=\; -\sum_{s\in\{0,1\}^n}\sum_{a\in\{0,1\}} \pi_t(s,a)\,\log \pi_t(s,a),$$

where $\pi_t(s, a)$ is the empirical frequency with which action $a \in \{0, 1\}$ is chosen in state s. A drop in H_t—herding—can increase average payoffs up to a critical threshold, beyond which individual rewards erode. Tracking H_t over time therefore not only reveals emerging herding dynamics but may also serve as an early warning of rising volatility or impending phase transitions. As an extension, one could use deep-learning approximations of the underlying Fokker–Planck equation to obtain an analytical solution to the time-evolution of the strategy distribution.

6 Results

In this section, a description of the results from our simulations is provided. Firstly, we will assess the "big picture" results to verify that our simulation was able to reproduce the dynamics observed in Arthur 1994, which will then allow us to examine the effects of the key parameters on the phenomena of interest.

6.1 Learned State–Action Values

Table 6.1 reports the converged Q–values for two sample agents (0 and 1) with memory length $M = 3$. Each row corresponds to one of the $2^3 = 8$ possible binary histories $s \in \{0,1\}^3$, and the columns $Q_0(s)$ and $Q_1(s)$ give the learned action-values for "stay home" ($a = 0$) and "attend" ($a = 1$), respectively. Notice that agent 0's table differs markedly from agent 1's—for instance, in state $(1,0,0)$ agent 0 has $Q_1 = 0.00$ while agent 1 has $Q_1 = 1.23$. However, there are Q values which are common across agents, despite their heterogeneity otherwise. Given that the profitability of states is dependent on the joint action of other agents, we cannot definitively discriminate optimal actions from suboptimal ones. This heterogeneity in Q–tables reflects that agents develop distinct heuristics: some states become strongly associated with attendance for one agent but not for another. Such diversity of learned policies prevents trivial lock-step behavior and gives rise to the rich coordination and de-synchronisation phenomena that characterise El Farol dynamics.

Agent	State	$\mathbf{Q_0}$	$\mathbf{Q_1}$
0	(0,0,0)	0.41	1.63
	(0,0,1)	0.68	0.00
	(0,1,0)	0.27	0.75
	(1,0,0)	1.09	0.00
	(0,1,1)	0.77	0.00
	(1,1,0)	0.04	1.23
	(1,0,1)	0.85	0.14
	(1,1,1)	0.48	0.00
1	(0,0,0)	1.16	0.06
	(0,0,1)	1.17	0.00
	(0,1,0)	2.18	0.28
	(1,0,0)	0.00	1.23
	(0,1,1)	0.94	0.00
	(1,1,0)	0.00	1.41
	(1,0,1)	0.00	1.99
	(1,1,1)	0.03	0.66

Table 2: State-action Q-values for Agent 0 and Agent 1

6.2 Attendance Dynamics

Figure 2 plots the total attendance over 200 rounds (with capacity $C = 60$). As in Arthur 1994, we observe an initial "exploration" phase of large, erratic swings around the red dashed line $A_t = C$, followed by rapid damping of fluctuations. After roughly 50–75 iterations, the process settles into a quasi-stationary regime in which A_t remains within an ε–neighborhood of C, yet continues to exhibit small, chaotic oscillations due to residual non-stationarity. These regularities—high early volatility, exponential-like decay of swings, and sustained near-capacity attendance—have been documented in both the original El Farol bar problem. Our results confirm that Q-learning agents reproduce these hallmark patterns while also generating heterogeneous, history-dependent behavior at the micro level.

6.3 Attendance Efficiency

In the El Farol bar problem, efficiency is maximised when the attendance A_t satisfies

$$A_t \approx C,$$

where C is the bar's capacity. Since volatility is inversely related to efficiency, a low attendance variance $\mathrm{Var}(A_t)$ implies more stable, efficient use of the bar.

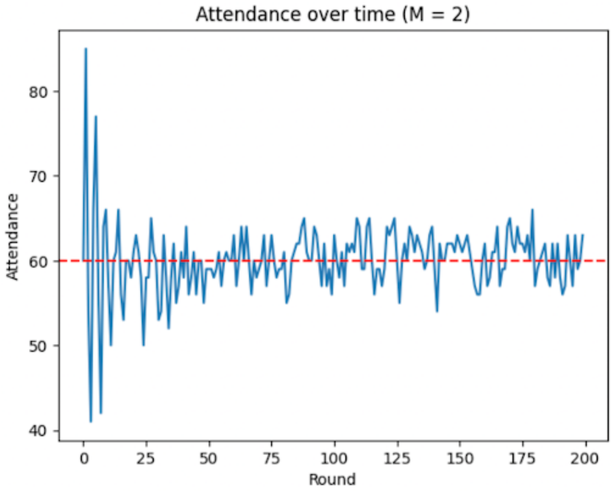

Figure 2: Attendance over 200 rounds for agents with memory length $M = 2$. The red dashed line marks the capacity threshold $C = 60$.

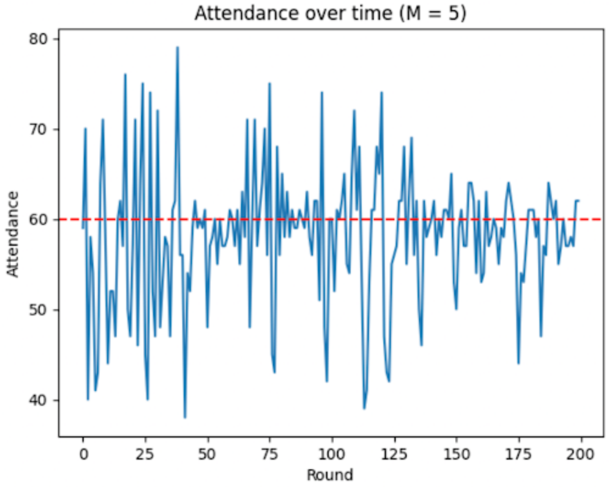

Figure 3: Attendance over 200 rounds for agents with memory length $M = 5$. The red dashed line marks the capacity threshold $C = 60$.

Attendance efficiency refers to how optimally the limited resource is utilised without waste or congestion. In the El Farol bar problem context, it is highest when the bar is neither under-attended nor overcrowded. In the Minority Game context, maximal efficiency corresponds to as many agents as possible being on the winning side each round. In other words, efficiency is maximised when attendance fluctuates around the optimal capacity threshold. A perfectly efficient outcome means the attendance each round is at the capacity for the El Farol bar problem, so that no resource is wasted and no excess crowding occurs. For example, Chen and Gostoli 2015 note that the "optimal use of the public facility" in El Farol means achieving neither idle capacity nor congestion.

Figure 2 (for $M = 2$) shows that, after an initial exploration phase with large swings, attendance rapidly settles into a narrow band around C. The variance of A_t in the steady state is small, indicating that nearly every round uses the bar's capacity efficiently without chronic under- or over-crowding. This behaviour mirrors Arthur's original findings and demonstrates that short memory suffices for agents to coordinate tightly. In contrast, Figure 3 (for $M = 5$) exhibits more pronounced and persistent swings away from C. Although the mean attendance remains near the capacity, the larger variance implies frequent under- and over-shooting. In effect, agents with longer memory over-react to outdated attendance patterns, reducing overall efficiency.

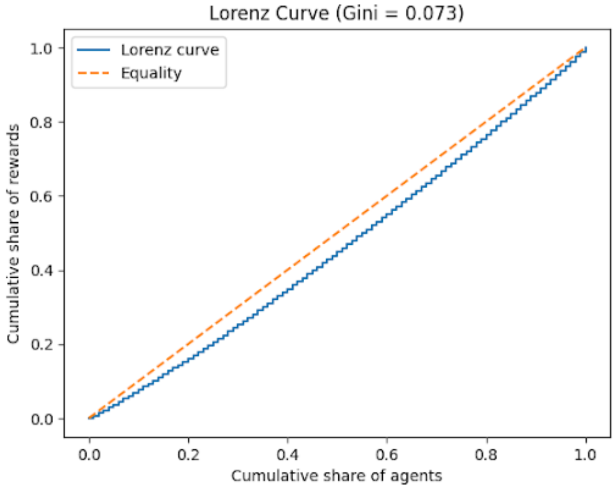

Figure 4: Lorenz curve of cumulative rewards for 100 agents (Gini coefficient = 0.073). The dashed orange line represents perfect equality; the slight bowing of the blue curve indicates a small degree of inequality, confirming a nearly uniform distribution of payoffs.

6.4 Equity of Outcomes

This is an important evaluation metric as one of the main questions we hope to answer is whether the pursuit of self-interest can lead to an efficient allocation of resources. However, it is not sufficient for a system to be solely efficient; in real-world systems efficiency is often achieved by a subset of agents at the expense of other subsets. The efficiency-equity trade-off thus arises as a product of this tension, and it continues to be ubiquitous across many social systems.

Fairness in the context of the El Farol bar problem refers to how evenly the payoffs are distributed among the agents. Even if the group as a whole achieves high efficiency, some agents might consistently win while others consistently lose – a fairness issue. A perfectly fair outcome would be one in which all agents have equal long-run average utilities where no one monopolises the wins, whereas unfair outcomes exhibit a wide gap between “winners” and “losers” over time.

Fairness has become a concern in later studies of the El Farol bar problem because pure efficiency does not guarantee equity. In fact, near the optimal coordination point, it’s possible for a small set of agents to capture disproportionate rewards by consistently being on the winning minority side, while others lag behind. Ho, F. K. Chow, and Chau 2004 found that “wealth inequality in the Minority Game is very severe near the point of maximum cooperation among players,” i.e. when volatility is minimal. This indicates a trade-off: the state with minimal volatility can lead to win distributions that are highly uneven, as a result of some agents consistently timing the minority side correctly. On the other hand, in a less coordinated, or more random, state, each agent’s chance of winning might be more equal but the resource is underutilised. Kiniwa, Koide, and Sandoh 2012 showed that if agents use very little memory or very simplistic strategies, both volatility is high and inequality is high; many agents get stuck losing while a lucky few might win often.

To evaluate equity in the El Farol Bar Problem, we employ three complementary measures: the Gini coefficient, Jain’s fairness index, and the Shannon entropy of the reward distribution. These metrics respectively capture inequality, overall fairness, and the dispersion of cumulative payoffs among agents, providing a multifaceted assessment of how rewards are shared in the long-run equilibrium.

Definition 6.1. *Gini Coefficient The Gini coefficient measures inequality in a distribution. It ranges from 0 to 1, which corresponds to perfect equality and complete inequality, respectively. The Lorenz curve plots the cumulative share of total rewards against the cumulative share of agents. If all agents share equally, the curve is the 45° line; the more it bows below that line, the greater the inequality.*

Definition 6.2. *Jain’s Fairness Index*

$$J = \frac{\left(\sum_{i=1}^{N} R_i\right)^2}{N \sum_{i=1}^{N} R_i^2}$$

where R_i is agent i’s total reward. A value of $J = 1$ corresponds to perfect fairness (all R_i equal), while $J < 1$ indicates growing disparity among agents’ rewards.

Definition 6.3. *Entropy of the Reward Distribution*

$$H = -\sum_{i=1}^{N} p_i \log p_i, \quad p_i = \frac{R_i}{\sum_{j=1}^{N} R_j}.$$

Entropy is also used to measure and quantify the degree of homogeneity of actions the distribution of attendance decisions among agents, and its employed again to measure how uniformly rewards are spread. Higher entropy implies a more uniform distribution, where entropy is maximised when rewards are uniform amongst agents, while lower entropy indicates concentration of rewards in fewer hands.

With a Gini coefficient of just 0.0731 and a Jain's fairness index of 0.9836, accompanied by a reward entropy of 4.5967 out of a maximum 4.6052 for 100 agents, the Q-learning agents achieved an equitable spread of cumulative rewards. In 4, the Lorenz curve almost coincides with the line of equality, confirming that no small minority of agents consistently outperforms the rest. This near-uniformity, sustained by a judicious balance of exploration, memory length, and the learning rate, indicates that the learning dynamics reliably converge to a symmetric equilibrium rather than a polarised or "winner-takes-all" outcome. This is supported by Figure 2, where we observe relatively uniform fluctuations around the threshold, which implies that agents are uniformly distributed above and below the threshold.

The concept of equity varies signifcantly across economic and philosophical schools of thought. The contrast is encapsulated in the tension between the aggregate reward and the distribution of individual rewards. A utilitarian judges equity by the sum total of welfare, indifferent to how it is distributed so long as overall benefit is maximised. In our simulation, utilitarians would note that Q-learning agents achieve high average rewards while keeping volatility low—so the collective reward is maximised. The near-equal distribution of payoffs is a bonus but not strictly required: what matters is that the market-analogue maximises aggregate payoff, and achieves an efficient outcome.

John Rawls argued that social and economic inequalities are permissible only if they benefit the least-advantaged members of society (Rawls 1971). Here, the very low Gini coefficient (0.073) and high Jain index (0.984) show that even the worst-off agents earn payoffs close to the average. From a Rawlsian standpoint, the Q-learning agents satisfy the difference principle: any residual inequality still leaves every agent nearly as well-off as their peers, so the system can be considered just. In accordance with both utilitarianism and rawlsian, the Q-learning agents deliver an outcome that is at once efficient and equitable, characterised by high total reward and low volatility, and an equal distribution of rewards, respectively.

6.5 Sweeps over Parameter Values

The nature of simulations subjects the results to both aleatoric and epistemic uncertainty. Aleatoric uncertainty is irreducible, irrespective of the permutation of parameter values, while epistemic uncertainty arises from model misspecification and can be reduced by refining our design. Such sweeps also assess whether parameters are sensitive to initial conditions or if there exist critical values at which phase transitions occur. In our parameter-sweep procedure, we vary one parameter at a time—holding all others fixed—to isolate its individual effect. This one-factor-at-a-time approach not only clarifies each parameter's influence but also enables us to pinpoint phase-transition thresholds, for example, between regimes of low and high volatility. In the appendix, we present plots illustrating how the long-run mean attendance, its variance, and the attendance entropy evolve as each model parameter is systematically varied in one-factor-at-a-time sweeps.

6.5.1 Memory Length

Varying each agent's memory length M reveals a critical balance: with very short memory, agents react too noisily to the most recent outcome, yielding low average attendance but also low volatility, while very long memory leads them to over-weight outdated information, synchronising their actions and driving persistent overcrowding with high variance. Empirically, as M rises from 1 to 8, mean attendance climbs beyond the capacity threshold, volatility first surges then plateaus, and policy entropy steadily declines, indicating more deterministic, herd-like behaviour (10). In practice, no single "sweet spot" entirely eliminates swings—too little memory induces erratic oscillations, and too much locks in suboptimal patterns—mirroring the critical-memory phenomenon observed in the Minority Game. 5 confirms that agents are most efficient at forecasting when endowed with a memory of two periods, and we see that attendance becomes underutilised as agents form more sophisticated forecasts. However, rewards are maximised when $M \approx 3.5$, as seen in 6. With $M \approx 3.5$, the average attendance is marginally below the threshold, and this may yield higher average rewards as it is approximately efficient, but it evades overattendance, and the corresponding penalty (7, 8).

For a certain range of parameters, the attendance figures of our simulations replicated Arthur's classic pattern: an initial "exploration" phase marked by large, erratic swings as agents tested their forecasting rules. As soon as they learned

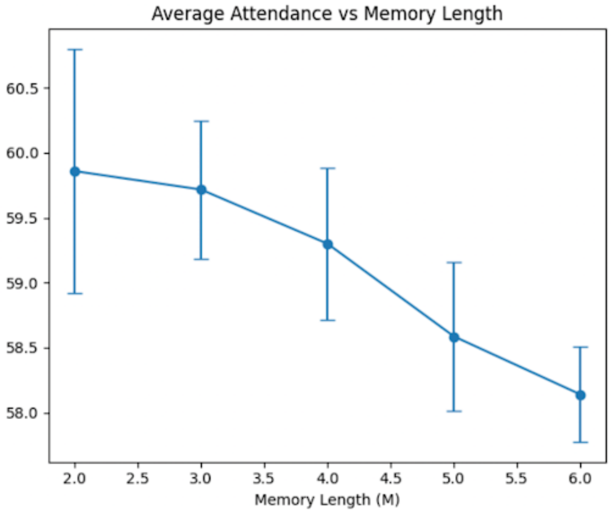

Figure 5: Average attendance as a function of memory length M, with vertical bars indicating one standard deviation across runs.

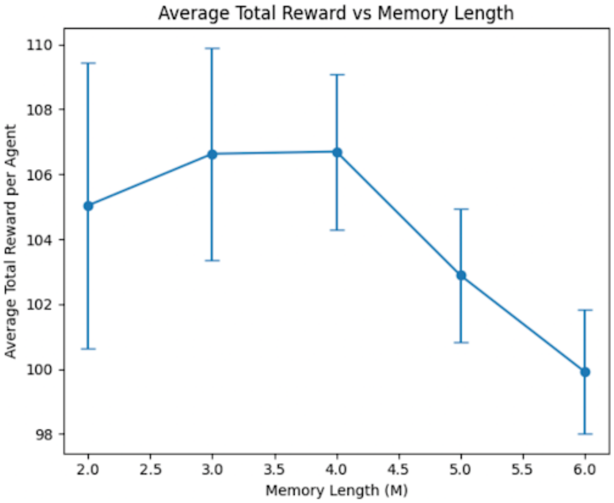

Figure 6: Average total reward per agent as a function of memory length M, with vertical bars representing one standard deviation across simulation runs.

both the bar's capacity and each other's behaviors, these fluctuations rapidly dampened. Eventually, the system settled into a sustained equilibrium in which volatility remained low and roughly constant over time. However, the system never only achieves approximate convergence as non-stationarity is inherent to the El Farol bar problem. Approximate (ϵ-neighbourhood) convergence manifests as follows: for a parameter $\epsilon \in \mathbb{R}$, the long run attendance, A, satisfies

$$\left|A - C\right| < \varepsilon$$

6.5.2 Attendance Threshold

Sweeping the bar's capacity C between extreme values shows that entropy is minimised when $C \approx N/2$; at low thresholds agents overwhelmingly abstain or herd, producing large attendance swings and moderate entropy, whereas at very high thresholds "attend" becomes almost risk-free, collapsing both exploration and variance. Around the balanced point, the minority mechanism stabilises attendance, yielding the lowest volatility and peak policy entropy. As C increases from 30 to 70, mean attendance rises linearly, volatility plunges, and entropy peaks near the midpoint before falling—consistent with the phase-transition behaviour of adaptive coordination models.

6.5.3 Exploration Parameters: ϵ-Greedy and Softmax

Comparing ϵ-greedy and softmax reveals similar exploration–exploitation trade-offs but different levers of control. In ϵ-greedy, a small $\epsilon \approx 0.1$ injects enough randomness to break herd-lock and minimise volatility, while $\epsilon = 0$ yields pure exploitation with chronic overcrowding and high variance, and $\epsilon \to 1$ yields near-random play with high entropy and low volatility. Softmax exploration, governed by temperature τ, mirrors these regimes—low τ acts greedily, high τ approaches uniform randomness—and an intermediate τ often achieves the best coordination. Critically, computing a two-action softmax (two exponentials and a normalising sum) imposes negligible overhead, even in large simulations.

The exploration parameter may also be interpreted as a rough proxy for the "risk appetite" of agents. A high ϵ or high τ means the agent willingly tries actions whose payoffs are uncertain, much as a risk-seeking investor samples volatile assets in hope of discovering hidden gains. Conversely, a low exploration setting corresponds to a more risk-averse stance: the agent sticks closely to its best-known action and shuns unknown alternatives. That said, exploration is really a mechanism for information-gathering rather than direct preference over risk; it trades short-term payoff stability for the chance of long-term learning. Still, in the El Farol bar context, tuning ϵ or τ upward does indeed make agents behave more "boldly," accepting higher variance in attendance outcomes in pursuit of potentially better strategies—an interpretation perfectly in line with the notion of risk appetite.

6.6 Learning Rate

The learning rate α governs how the rate at which agents update their Q-values and therefore balances speed of adaptation against stability. Low α (~ 0.01) yields conservative updates and moderate volatility, while very high α (~ 0.9) produces rapid but noisy learning with large, synchronised attendance swings and reduced entropy. An intermediate α (~ 0.1–0.5) typically offers the best trade-off, allowing agents to approach capacity efficiently without locking into erratic, herd-driven cycles—reflecting the classic convergence trade-offs of reinforcement learning.

7 Conclusion

Our implementation of tabular Q-learning agents in the El Farol bar problem successfully reproduces Arthur's classic dynamics—initial high volatility followed by convergence to an ε-neighbourhood around the capacity threshold. We demonstrate that attendance efficiency, measured by low $\mathrm{Var}(A_t)$ and $|A_t - C| < \varepsilon$, is achieved with a range of memory lengths, exploration strategies, and learning rates.

Equity metrics further confirm that self-interested reinforcement learning agents need not sacrifice fairness for efficiency: the observed Gini coefficient of 0.0731, Jain's index of 0.9836, and reward entropy close to its maximum indicate an almost uniform distribution of long-run payoffs. From both utilitarian and Rawlsian perspectives, the Q-learning framework attains high total reward while ensuring that even the least-advantaged agents fare nearly as well as the average.

Parameter sweeps reveal clear phase-transition thresholds: memory length M exhibits a critical regime where volatility is minimised without overcrowding, capacity C peaks in efficiency at $C \approx N/2$, and exploration parameters (ϵ or temperature τ) balance exploration–exploitation to avoid herd-lock or randomness. These one-factor-at-a-time analyses validate that our findings are not artifacts of specific parameter choices but reflect genuine emergent phenomena.

Embedding Q-learning into the El Farol bar problem provides a more sophisticated mechanism for modelling boundedly rational agents, and their self-organisation into efficient and equitable equilibria. Our results open avenues for richer extensions—such as deep-reinforcement learning approximations of the underlying Fokker–Planck dynamics—and suggest that decentralised reinforcement learning can serve as a unifying mechanism in the study of complex adaptive systems.

Furthermore, we plan to extend our framework to more realistic settings by incorporating greater agent heterogeneity, introducing a dynamic attendance threshold, and gradually adding new agents. These enhancements will challenge the stability of learned coordination, since implicit collusion based on fixed state–action values may break down under non-stationary population and environmental changes, thereby testing the robustness and generality of our results.

References

Arthur, W. Brian (May 1994). "Inductive Reasoning and Bounded Rationality". In: *The American Economic Review* 84.2. Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association, pp. 406–411.

Challet, Damien, Matteo Marsili, and Gabriele Ottino (Dec. 2004). *Shedding Light on El Farol*. Tech. rep. Posted: 6 Dec 2004. CentraleSupélec; Abdus Salam International Centre for Theoretical Physics; University of Fribourg.

Chen, Shu-Heng and Umberto Gostoli (2015). "Coordination in the El Farol Bar Problem: The Role of Social Preferences and Social Networks". In: *Journal of Economic Interaction and Coordination* 10.1, pp. 59–93. DOI: 10.1007/s11403-015-0150-z.

Chung, Jen Jen, Scott Chow, and Kagan Tumer (2018). *When Less Is More: Reducing Agent Noise with Probabilistically Learning Agents*. Extended Abstract, Oregon State University. Jen Jen Chung (jenjen.chung@oregonstate.edu), Scott Chow (chows@oregonstate.edu), Kagan Tumer (kagan.tumer@oregonstate.edu).

Farmer, J. Doyne and Duncan Foley (2009). "The economy needs agent-based modelling". In: *Nature* 460, pp. 685–686. DOI: 10.1038/460685a.

Franke, Reiner (July 2003). "Reinforcement learning in the El Farol model". In: *Journal of Economic Behavior & Organization* 51.3, pp. 367–388. DOI: 10.1016/S0167-2681(02)00152-X.

Ho, K. H., F. K. Chow, and H. F. Chau (2004). "Wealth Inequality in the Minority Game". In: *Physical Review E* 70.6, p. 066110. DOI: 10.1103/PhysRevE.70.066110.

Kahneman, Daniel and Amos Tversky (Mar. 1979). "Prospect Theory: An Analysis of Decision under Risk". In: *Econometrica* 47.2, pp. 263–292.

Kaszowska-Mojsa, Jagoda and Mateusz Pipień (2023). "Macroprudential Policy in a Heterogeneous Environment—An Application of Agent-Based Approach in Systemic Risk Modelling". In: *Journal of Economic Modelling* XX.YY. Jagoda Kaszowska-Mojsa (Polish Academy of Sciences); Mateusz Pipień (Cracow University of Economics). These authors contributed equally., pp–pp. DOI: 10.XXXX/XXXXXXXXX.

Kiniwa, Jun, Takeshi Koide, and Hiroaki Sandoh (2012). "A New Variant of the Minority Game: Asset Value Game and Its Extension". In: *Proceedings of the 4th International Conference on Agents and Artificial Intelligence (ICAART)*, pp. 15–22. DOI: 10.5220/0003715800150022.

LeBaron, Blake (2006). "Agent-based Computational Finance". In: *Handbook of Computational Economics*. Ed. by Kenneth L. Judd and Leigh Tesfatsion. Vol. 2. blebaron@brandeis.edu. Elsevier, pp. 1187–1233.

Marsili, Matteo and Damien Challet (2001). "Trading Behavior and Excess Volatility in Toy Markets". In: *Advances in Complex Systems* 4.1, pp. 3–17. DOI: 10.1142/S0219525901000024.

Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Tesfatsion, Leigh and Kenneth L. Judd, eds. (2006). *Handbook of Computational Economics, Volume 2: Agent-Based Computational Economics*. Amsterdam, The Netherlands: Elsevier.

Watkins, Christopher J. C. H. and Peter Dayan (1992). "Q-learning". In: *Machine Learning* 8.3-4, pp. 279–292. DOI: 10.1007/BF00992698.

Whitehead, Duncan (Sept. 2008). *The El Farol Bar Problem Revisited: Reinforcement Learning in a Potential Game*. Tech. rep. 186. Working paper. Edinburgh School of Economics Discussion Paper Series.

Zambrano, Eduardo (Feb. 2004). *The Interplay Between Analytics and Computation in the Study of Congestion Externalities: The Case of the El Farol Problem*. Tech. rep. 97-06-060. Received November 2003; Accepted February 2004. University of Notre Dame; Banco Central de Venezuela.

Zhang, Lin and Xinquan Liu (2021). "Traffic Game Model with the Contract Model". In: *Scientific Programming* 2021. Academic Editor: Jiwei Huang, p. 6189075. DOI: 10.1155/2021/6189075.

Zhang, Si-Ping et al. (2023). "Self-organizing Optimization and Phase Transition in Reinforcement Learning Minority Game System". In: *To appear*. Details forthcoming.

Zheng, Guozhong et al. (Dec. 2023). *Optimal Coordination in Minority Game: A Solution from Reinforcement Learning*. arXiv preprint arXiv:2312.14970. arXiv: 2312.14970 [physics.soc-ph].

A Appendix

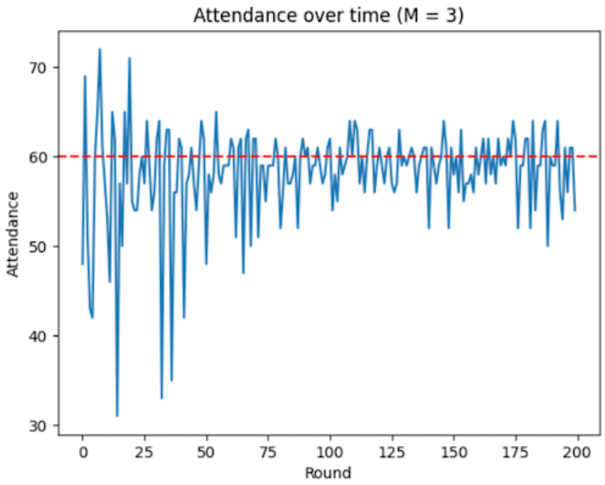

Figure 7: Attendance over 200 rounds for agents with memory length $M = 4$.

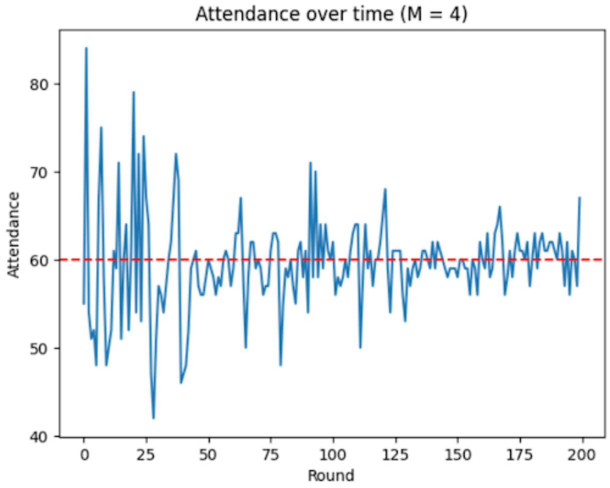

Figure 8: Attendance over 200 rounds for agents with memory length $M = 4$.

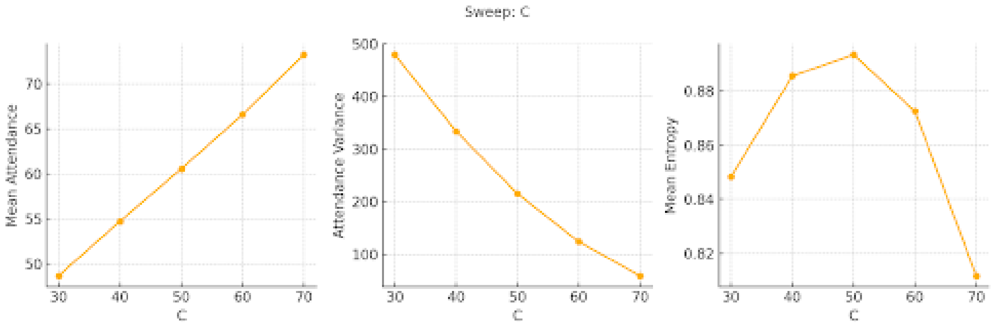

Figure 9: Effects of varying the capacity threshold C on (left) mean attendance, (center) attendance variance, and (right) mean policy entropy.

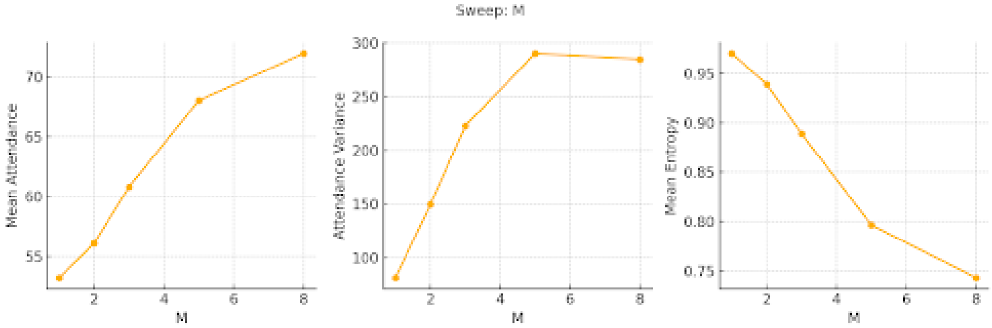

Figure 10: Effects of varying memory length M on (left) mean attendance, (center) attendance variance, and (right) mean policy entropy.

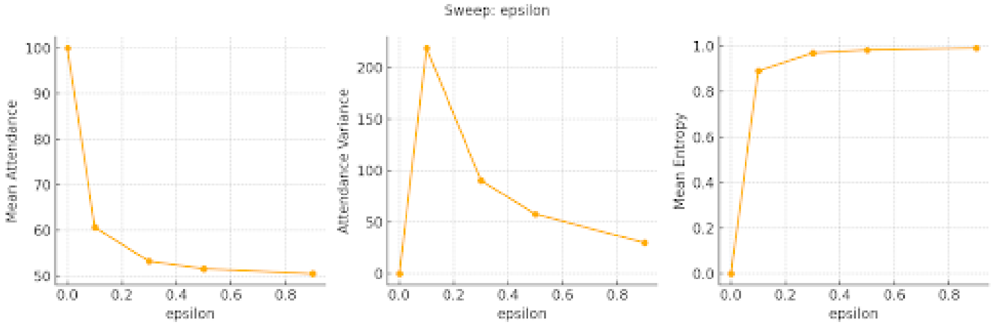

Figure 11: Effects of varying the exploration parameter ϵ on (left) mean attendance, (center) attendance variance, and (right) mean policy entropy.

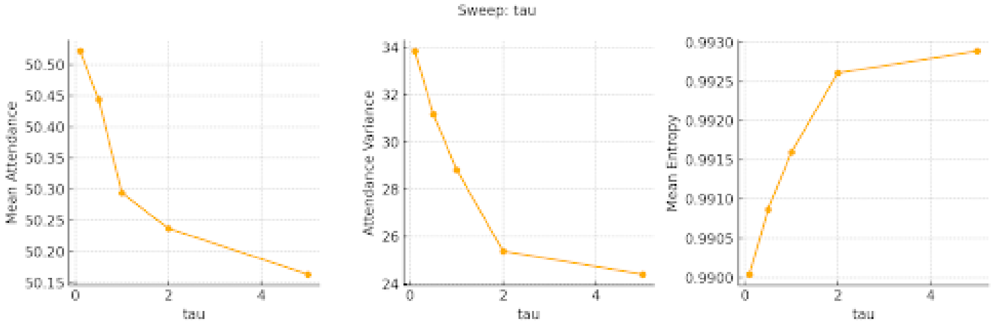

Figure 12: Effects of varying the exploration parameter τ on (left) mean attendance, (center) attendance variance, and (right) mean policy entropy.

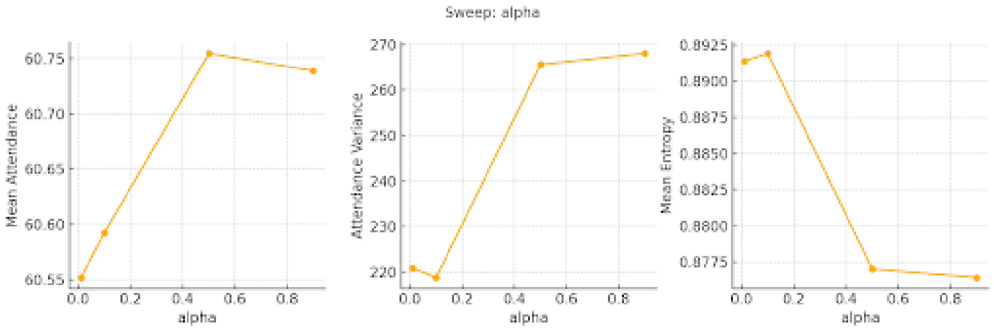

Figure 13: Effects of varying the learning rate α on (left) mean attendance, (center) attendance variance, and (right) mean policy entropy.