Adaptive Intelligence in Retail Space Optimization: Modeling the Coffee Shop Dilemma with Q-Learning Agents

Siranee Nuchitprasitchai[1], Kanchana Viriyapant[2], Kanjanee Satitrangseewong[3], May Myo Naing[4]
Faculty of Information Technology and Digital Innovation, KMUTNB, Bangkok, Thailand[1,2]
Kanokphon Digital, Bangkok, Thailand[3]
World Vision International, Bangkok, Thailand[4]

***Abstract*—This study models the "coffee shop dilemma", where customer attendance is discouraged by both overcrowding and emptiness. Using an agent-based model with Q-learning reinforcement learning, this study simulates the daily decisions of 100 agents over a one-year period. The results reveal a self-organizing attendance cycle around a 60% capacity threshold. This study demonstrates that customer satisfaction is not driven by visit frequency, but by adaptive decision-making strategies shaped by learned congestion values. Clustering analysis identifies distinct behavioral patron groups (e.g., Ultra-Frequent, Optimized) that emerge from these subtle value differences. The study provides a data-driven framework for optimizing shop space and customer flow, offering conceptual insights into balancing the needs of quick-service and long-stay customers by dynamically managing perceived occupancy.**

Keywords*—*El Farol Bar problem; agent-based modeling; Q-learning; reinforcement learning; customer behavior; congestion paradox; decision-making; coffee shop operations

I. Introduction

The modern coffee shop—exemplified by global chains that have evolved beyond a beverage outlet into a multifaceted "third place" that serves as both a social and professional hub distinct from home and office environments. This evolution introduces a complex operational challenge: managing two often conflicting customer flows. On one hand, the business must serve quick-service customers who prioritize speed and convenience during peak hours. On the other hand, it must accommodate remote workers, students, and long-stay patrons who value comfort and consistency, a key component of the brand's identity.

This tension gives rise to what may be termed the modern coffee shop dilemma, a real-world analogue of the El Farol Bar problem [1]. In the original formulation, agents independently decide whether to attend a bar, with the utility of attendance declining as crowding increases. Similarly, in coffee shops, customers decide whether to visit based on perceived crowding. Overcrowding deters visitors due to limited seating, excessive noise, and reduced comfort, while under-crowding can signal a lack of social energy or desirability. The resulting dynamic equilibrium emerges from individual decisions that collectively influence store congestion levels.

Although the El Farol problem has been widely studied in economics and complex systems theory, its application to high-volume retail environments remain limited. Prior research has predominantly emphasized theoretical agent behavior or macroeconomic implications, with less attention to data-driven micro-level business insights.

The main contributions of this study are summarized as follows:

- Operational reframing of El Farol: The El Farol Bar problem is translated into a retail coffee shop context, enabling the analysis of customer attendance decisions under perceived occupancy constraints rather than abstract capacity rules. Unlike classical El Farol formulations, capacity in this study represents a subjective comfort threshold rather than a hard physical limit.
- Satisfaction-based reward design: A reward-shaping mechanism is introduced that explicitly links customer satisfaction and dissatisfaction to congestion levels, allowing agents to learn from perceived crowding rather than visit frequency alone.
- Emergent behavioral segmentation: Post-learning clustering reveals distinct customer archetypes with stable attendance routines, demonstrating how heterogeneous strategies emerge from decentralized Q-learning dynamics.
- Retail space insights: The learning outcomes are connected to practical retail implications, showing how managing perceived occupancy can balance quick-service and long-stay customers without centralized control.

II. Related Work

This section reviews prior studies relevant to the present work across three complementary research streams. First, foundational models of the El Farol Bar problem and related congestion games are discussed to establish the theoretical basis for decentralized attendance decisions. Second, reinforcement learning approaches applied to congestion and coordination settings are reviewed, with emphasis on adaptive agent behavior. Finally, empirical and simulation-based studies on

perceived crowding and retail space optimization are examined to situate the model within operational retail contexts.

A. The El Farol Bar Problem

The El Farol Bar Problem, first introduced by Arthur [1], illustrates how boundedly rational agents make attendance decisions when utility depends on aggregate participation. This foundational model inspired a broad family of coordination and congestion problems, including the Minority Game [2], which formalizes inductive reasoning and self-organization under limited information.

Subsequent studies extended the El Farol paradigm through agent-based modeling (ABM) and reinforcement learning (RL) to analyze emergent equilibrium behavior and resource utilization [3]. Schosser [4] examined fairness considerations in the allocation of limited shared resources during the COVID-19 pandemic through an El Farol–type framework, analyzing how individual attendance decisions affect equitable outcomes under scarcity, without explicitly modeling infection dynamics.

The anti-coordination nature of the El Farol problem has been extensively studied through various methodological lenses. Guarnieri and Spadoni [5] investigated the role of social norms in anti-coordination decisions through experimental elicitation and priming methods. Their findings revealed that subjects tend to comply with perceived descriptive norms (empirical expectations about majority behavior) rather than injunctive norms (normative expectations about what is considered appropriate). As a result, overall attendance becomes less volatile and stays closer to the bar's capacity more consistently. This reduces the inefficiency of both overcrowding (too many go) and underutilization (too few go) [6].

Atilgan et al. [7] explored collective behavior in the El Farol Bar through the lens of memory horizon and selection criteria for prediction algorithms. Their work demonstrated that the distribution of algorithm clusters varies significantly with shorter agent memories, directly impacting long-term attendance dynamics. They identified a critical memory horizon where correlations in attendance deviations take longer to decay, suggesting a phase transition in collective behavior.

However, existing El Farol and Minority Game formulations primarily investigate coordination and equilibrium properties in abstract settings, without explicitly modeling recurring customer decisions driven by perceived congestion in operational retail environments, which is the focus of the present study.

B. Reinforcement Learning in Congestion Games

In applying Q-learning to a Minority Game, Zheng et al. [8] demonstrated that optimal coordination emerges when agents balance exploration and exploitation, governed by a specific temperature parameter. This balance prevents the system from being trapped in suboptimal, exploitative periodic states or degrading into random, exploratory behavior. Compared to conventional methods, the Q-Learning algorithm achieves improved financial performance and able to yield the highest financial returns through its dynamic adaptation to evolving market conditions and its effective management of price demand complexities [9], [10].

Kossack [11] extended reinforcement learning approaches by introducing an emotional machine framework that incorporates artificial emotions (satisfaction, fear, ambition) as states influencing decision-making. Applied to a 3-player El Farol scenario, this work demonstrated that emotional states evolve through differential equations weighted by personality parameters, producing different collective outcomes. When satisfaction dominates, agents form stable coalitions; when fear dominates, volatility increases; and when ambition dominates, agents pursue aggressive entry strategies leading to congestion.

While reinforcement learning has been extensively applied to congestion and coordination games, prior studies largely emphasize convergence behavior or algorithmic performance. In contrast, the present study examines how learned congestion valuations translate into persistent attendance routines and interpretable behavioral segments within a retail like environment, rather than examining how learned congestion valuations translate into stable attendance routines and interpretable customer segments in a retail context. Research on multi-agent coordination problems, such as traffic signal control, demonstrates the efficacy of reinforcement learning in managing distributed congestion [12].

In contrast, the present study focuses on how learned congestion valuations translate into persistent attendance routines and interpretable behavioral segments in a retail-like environment.

C. Retail Space Optimization

Prior research has established that perceived crowding is a critical factor influencing consumer behavior in retail environments. Zein et al. [13] demonstrate that high human and spatial density directly shape customer satisfaction through emotional responses, where perceived crowding negatively impacts pleasure and arousal, leading to approach behaviors. For instance, the optimal rush-hour staffing model (3 cashiers, 3 baristas) addressed the checkout bottleneck, cutting customer wait times by 40\% and increasing throughput [14] and also optimized retails space usage [15]. This underscores that it is not just objective occupancy but the customer's perception of that density that dictates their experience and loyalty.

This foundation is directly relevant to the "coffee shop dilemma" in our study. While Zein et al. establish a causal link between density and emotion in a retail context, this study aims to investigate how this perception drives adaptive decision-making in a recurrent visitation scenario and operationalize their core finding—that crowding perception alters behavior—by using an agent-based model to simulate how customers learn to avoid dissatisfaction.

A key strength of this approach, as evidenced in institutional policy work, is its capacity to model how macro-level patterns emerge from micro-level behaviors—a core focus of this simulation [16].

In contrast to empirical and optimization-based retail studies that treat crowding as an exogenous condition, this work models perceived occupancy as an endogenous signal learned through repeated interaction, enabling the analysis of adaptive customer behavior and emergent attendance patterns over time.

III. Experimental Setup

The study employs a modeling approach that balances analytical tractability with realistic behavior. The 60\% capacity threshold approximates the subjective transition from comfortable activity to perceived overcrowding in retail environments, rather than denoting a strict physical constraint. The 60\% threshold is not claimed to be universal, but represents a plausible comfort boundary used to induce congestion dynamics; sensitivity analysis is left for future work. This design frames dissatisfaction as a function of experienced congestion. Customer decisions are modeled as binary (attend or not), capturing the recurring nature of real world visit choices while maintaining simplicity. To preserve interpretability, a tabular Q-learning method is preferred over more complex deep reinforcement learning techniques. This enables direct examination of how persistent congestion influences long-term attendance strategies and the emergence of distinct behavioral groupings.

A. Environment Parameters

The simulation assumes a total of $N = 100$ agents, with a bar capacity set at 60% of N to induce congestion dynamics. Time is modeled in discrete rounds, where all agents simultaneously decide whether to attend the bar, allowing sufficient convergence of learning dynamics, as in Table I.

TABLE I. Simulation Parameters

Parameter	Value
Agents (N)	100
Simulation days (T)	365
Capacity threshold (C)	0.6
Learning rate (α)	0.1
Discount factor (γ)	0.9
Initial exploration rate (εinitial)	0.3
Minimum exploration rate (εmin)	0.01
Satisfaction penalty (ϕ)	-2

The model parameters were selected to balance realism, computational efficiency, and learning stability.

- Agents (N = 100): A population size that generates complex emergent behavior.
- Simulation days (T = 365): A one-year horizon that allows agents to experience multiple seasonal cycles and fully converge their learning strategies, ensuring observed patterns are stable and not transient.
- Capacity threshold (C = 0.6): The 60\% occupancy threshold where the venue shifts from "vibrant" to "crowded".
- Learning rate (α = 0.1): A value that controls how quickly agents update their Q-values based on new experiences. A rate of 0.1 ensures stable learning by preventing Q-values from fluctuating too drastically from a single day's outcome.
- Discount factor (γ = 0.9): Determines the importance of future rewards. A high value of 0.9 encourages agents to be farsighted, considering the long-term consequences of their attendance patterns rather than just immediate gratification.
- Initial exploration rate (εinitial = 0.3): The starting probability that an agent will choose is a random action. A 30\% rate promotes sufficient exploration of the action space in the early stages of the simulation to prevent premature convergence to suboptimal strategies.
- Minimum exploration rate (εmin = 0.01): The lower bound for exploration. A 1% rate ensures that agents never completely stop exploring, allowing them to adapt to slow changes in the collective attendance pattern over time.
- Satisfaction penalty (ϕ = -2): A scalar that quantifies the dissatisfaction of encountering a crowded venue. The value of −2 creates a strong negative reward for a "Punished Visit" (R = −1), making it a distinctly undesirable outcome compared to the high reward of a "Rewarded Visit" (R = 3).

The selected parameter values do not simulate a specific retail instance; rather, they are calibrated to facilitate the observation and examination of stable learning patterns that arise from constraints imposed by subjective congestion.

B. Reward Function

The reward function $R(a_t,d_t)$ is the core mechanism that encodes the "congestion paradox," guiding agent learning by quantifying the desirability of each outcome. The function takes the daily attendance at and an agent's decision d_t as inputs, where $d_t = 1$ signifies "Go" and $d_t = 0$ signifies "Stay." The function is formally defined as:

$$R(at, dt) = \begin{cases} 1 + \phi, & if\ d_t = 1 \ \wedge\ a_t > \mathrm{N} \cdot \mathrm{C}\ \text{(Punished Visit)} \\ 1 - \phi, & if\ d_t = 1 \ \wedge\ a_t \leq \mathrm{N} \cdot \mathrm{C}\ \text{(Rewarded Visit)} \\ 0, & if\ d_t = 0 \ \wedge\ a_t \leq \mathrm{N} \cdot \mathrm{C}\ \text{(Justified Absence)} \\ 1, & if\ d_t = 0 \ \wedge\ a_t > \mathrm{N} \cdot \mathrm{C}\ \text{(Strategic Avoidance)} \end{cases}$$

This reward structure is designed to reflect experiential outcomes rather than transactional utility, allowing agents to learn attendance strategies based on perceived satisfaction and avoidance of negative congestion experiences.

The rationale for each case, with the satisfaction penalty ϕ=−2, is as follows:

- Punished Visit (R=−1): The agent goes but finds the shop overcrowded (at > N·C). They receive a base reward of 1 for making a decision but a strong penalty ϕ, resulting in a net negative reward. This discourages visiting during peak times.
- Rewarded Visit (R=3): The agent goes and finds the shop pleasantly occupied (at ≤ N·C). The base reward is augmented by the negative of the penalty (−ϕ), creating a high positive reward. This reinforces visiting during optimal capacity.

- Justified Absence (R=0): The agent stays away and the shop is not crowded. This neutral reward reflects no gain or loss for correctly avoiding an unnecessary trip.
- Strategic Avoidance (R=1): The agent stays away and correctly avoids a crowded shop (at > N·C). The positive reward reflects the benefit of a smart, strategic decision to avoid a negative experience.

C. Simulation Procedure

The daily simulation procedure, outlined in Table II, executes the core agent-based learning cycle. Each step is elaborated below:

- Initialize Agents: Each agent i is initialized with a Q-table, Qi, with state-action values set to zero for the actions 'Stay' (0) and 'Go' (1), forcing learning from experience.
- Decay Exploration Rate: The exploration rate ε_t decays linearly from $\varepsilon_{initial} = 0.3$ to $\varepsilon_{min} = 0.01$, promoting early experimentation and later exploitation.
- Action Selection: Each agent uses an ε-greedy policy. With probability ε_t , it explores (random action); otherwise, it exploits by choosing arg max Qi.
- Calculate Attendance: The daily attendance at is computed as the sum of all agents' 'Go' decisions ($d_t^i = 1$), forming the environmental state.
- Assign Rewards: Each agent receives an immediate reward rt based on the function $R(a_t, d_t^i)$ from Eq. (1), implementing the congestion paradox.
- Update Q-values: Agents update their Q-value for the chosen action using the Q-learning rule, incorporating the immediate reward rt and the discounted future reward estimate (γ·maxQi).
- Log Data: Comprehensive data (global attendance a_t, individual actions d_t^i, rewards rt, and Q-values) is recorded for post-simulation analysis.

TABLE II. DAILY SIMULATION PROCEDURE

Step	Description
1	Initialize agents with Q-tables $Q_i = [0,0] for\ action\ \{0 = Stay, 1 = Go\}$
2	Decay exploration rate: $\varepsilon_t = \max\ (\varepsilon_{min}, \varepsilon_{t-1} - \frac{\varepsilon_{initial} - \varepsilon_{min}}{T})$
3	Action selection: For each agent, with probability ε t choose random action; otherwise, select arg max Qi
4	Calculate attendance: $a_t = \sum_{i=1}^{N} d_t^i$ Where d_t^i is agent i decision
5	Assign rewards: Compute $R(a_t, d_t^i)$ for each agent using Equation (1)
6	Update Q-values: $Q_i(d_t) \leftarrow Q_i(d_t) + \alpha[r_t + \gamma \cdot max Q_i - Q_i(d_t)]$
7	Log data: Record attendance, decisions, rewards, and Q-values for analysis

D. Data Collection

The following agent-level data was collected for post-simulation analysis:

- Agent_ID: Unique identifier (1–100)
- Total_Reward: Sum of all rewards over 365 days
- Historical_decision: Sequence of daily actions (Go/Stay)
- Cumulative_Reward: Time-series of cumulative rewards
- Q_Values: Final learned Q-values for Stay/Go actions

This comprehensive data collection enables clustering analysis and behavioral pattern identification as discussed in the results section.

E. Implementation Details

All experiments were executed using Python 3.10 with NumPy and Matplotlib libraries for computation and visualization. Data analytics, including clustering of attendance patterns and convergence plots, were performed using the pandas and scikit-learn libraries. Each experiment was repeated for 30 independent runs with different random seeds to ensure statistical robustness.

IV. RESULTS AND DISCUSSION

This is a summary of analytical steps, as in Table III.

TABLE III. SUMMARY OF ANALYTICAL STEPS

No.	Analytical Phase	Description
1	Overall Agent Performance	Analyzed the distribution of total cumulative rewards and the attendance over time across all agents to establish baseline performance and identify variance in strategy success.
2	Agent Clustering and Behavioral Analysis	Applied K-means clustering to agent behavioral features (frequency, Q-values, reward) to segment the population into distinct strategic archetypes (e.g., Ultra-Frequent, Optimized, Frequent).
3	Cluster Characteristics	Quantified and compared the properties of the identified clusters (size, Q-Go and Go Rate by Cluster, mean Q-values, mean reward) to define their strategic profiles.
4	Cluster Behavior and Decision Pattern Analysis	Visualized top weekly decision sequences for each cluster to reveal the temporal patterns and adaptive (or non-adaptive) nature of their strategies.
5	Discussion of Weekly Decision Patterns	Discuss on revealing behaviorally realistic routines which reflect strategic diversity and real-world influences.

A. Overall Agent Performance

The plot in Fig. 1 illustrates the cumulative reward trajectories of all Q-learning agents over 365 simulation days. The red line represents the population-average cumulative reward across n=100n=100 agents, while the shaded blue area denotes ±1±1 standard deviation. The dashed green lines mark the minimum and maximum cumulative rewards observed among agents.

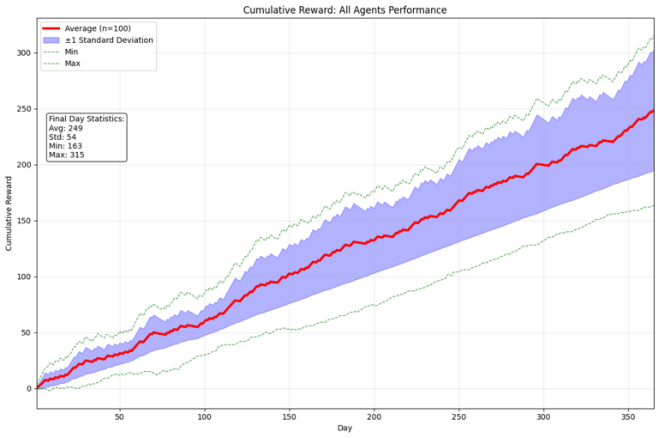

Fig. 1. Cumulative reward: all agents performance.

The learning trend indicates that agents progressively improve their cumulative rewards over time, suggesting that the Q-learning mechanism successfully guides decision adaptation. The mean cumulative reward at the final day reached 248.65, with a standard deviation of 54.24, a minimum of 163.00, and a maximum of 315.00. These values imply moderate heterogeneity in agent performance, reflecting differences in learned strategies and convergence rates.

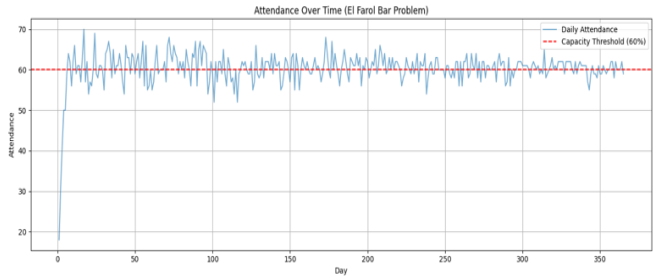

Fig. 2. Attendance overtime with 60 per cent threshold.

Fig. 2 illustrates the average attendance over time compared with the bar capacity threshold. The system exhibits convergence toward the capacity level after an initial transient phase, indicating that agents collectively learn to balance the exploration–exploitation trade-off. The variance of attendance decreases as episodes progress, confirming the stabilization of the learning dynamics.

Overall, the steady upward trajectory of the mean reward curve demonstrates that the agents collectively adapt to the attendance-constrained environment inherent in the El Farol Bar problem. The increasing spread over time reflects diversity in learning outcomes—some agents adopt efficient attendance policies, yielding higher rewards, whereas others stabilize at suboptimal attendance frequencies.

B. Agent Clustering and Behavioral Analysis

A k-means clustering procedure with k=3 was applied to characterize heterogeneity among agents using two behavioral features: the learned Q-value for attending (Q_Go) and the attendance probability (GoRate). The clustering separated the population into three distinct groups. The scatter plot of agents with centroids is presented in Fig. 3, and the cluster summary statistics are reported in Table IV.

Fig. 3. Agent strategy cluster.

TABLE IV. CLUSTER SUMMARY STATISTICS

Cluster	Count	Q_Go (mean)	GoRate	Total Reward (mean)
0	60	8.71	0.92	291.30
1	37	3.54	0.09	181.03
2	3	5.00	0.66	229.67

The three clusters are interpreted as follows:

Cluster 0 — Regular Attenders: Cluster 0 (60 agents) exhibits the highest Q_Go values (mean = 8.71) and the highest attendance probability (mean GoRate = 0.92). Agents in this group consistently choose to attend and, on average, obtain the largest cumulative reward (mean total reward = 291.30), representing the dominant "frequent attender" or "optimistic learner" archetype.

Cluster 1 — Non-Attenders / Cautious Agents: Cluster 1 (37 agents) is characterized by low Q_Go (mean = 3.54) and very low attendance probability (mean GoRate = 0.09). These agents adopt cautious strategies, rarely attending; as a consequence they obtain lower cumulative rewards (mean total reward = 181.03), reflecting missed opportunities when the bar is below capacity as well as avoidance of congestion penalties.

Cluster 2 — Opportunistic / Small Sample: Cluster 2 contains only 3 agents with intermediate Q_Go (mean = 5.00) and moderate attendance (GoRate = 0.66). Descriptively, these agents are opportunists, balancing attendance with restraint for intermediate rewards (mean = 229.67). However, because this cluster comprises only three samples, it is too small for reliable inferential comparisons (e.g., hypothesis testing or robust summary statistics). Therefore, Cluster 2 is reported here for completeness and qualitative interpretation only; subsequent comparative analyses and statistical tests focus on the two large clusters (Cluster 0 and Cluster 1).

C. Cluster Characteristics

The clustering reveals three distinct behavioral archetypes: persistent regulars (Cluster 0), cautious non-attenders (Cluster 1), and a very small opportunistic group (Cluster 2). The dominance of Cluster 0 in cumulative reward suggests that proactive attendance strategies can be advantageous when the collective system converges toward the bar capacity. The small size of Cluster 2 warrants caution in its interpretation: it may reflect a transient or niche strategy that emerged under the

specific random seed and parameterization of this simulation, rather than a robust population mode.

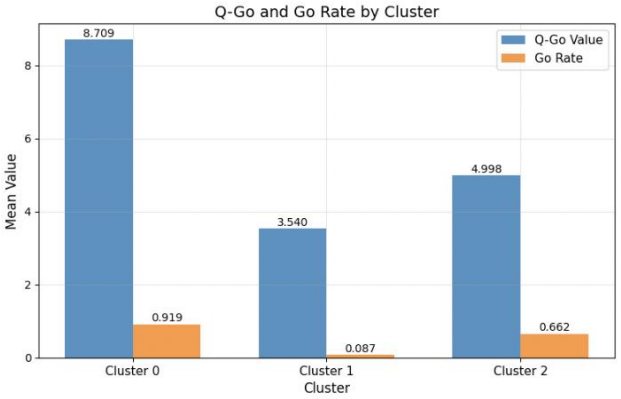

Fig. 4. Q-go and go rate by cluster.

Overall, these clustering results support the view that Q-learning agents self-organize into clearly separated behavioral groups with markedly different attendance policies and long-term payoffs. The explicit exclusion of the small Cluster 2 from inferential comparisons preserves the statistical validity of subsequent analyses, while all observed patterns are reported for transparency.

The comparative visualization in Fig. 4 highlights the mean Q-values, Go rates, and total rewards across clusters. Cluster 0 clearly dominates in terms of performance, suggesting that proactive strategies yield the greatest long-term benefits when collective learning drives the population toward an optimal equilibrium.

D. Cluster Behavior and Decision Pattern Analysis

To further investigate the heterogeneity of agent behaviors, a post-hoc clustering analysis was performed based on agents' weekly decision sequences and cumulative performance. Three distinct clusters were identified; however, Cluster 2 was excluded from interpretation due to its extremely small sample size (n = 3), which is unlikely to represent a statistically meaningful behavioral pattern. The analysis therefore focuses on Cluster 0 and Cluster 1.

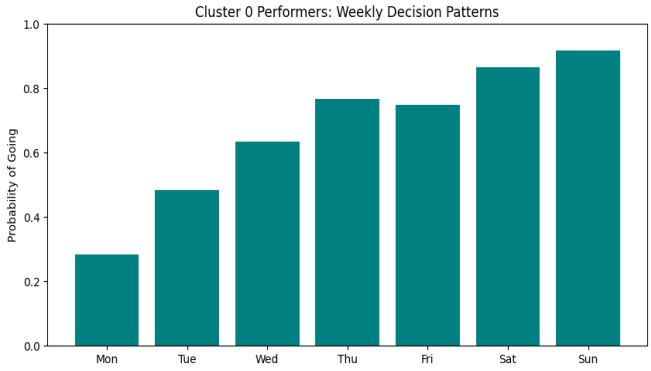

Fig. 5. Cluster 0 weekly decision patterns.

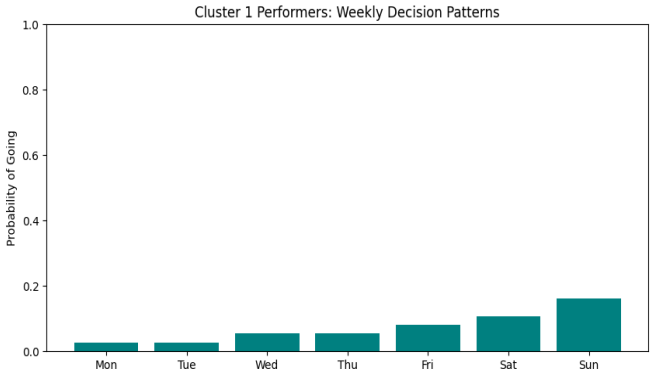

Fig. 6. Cluster 1 weekly decision patterns.

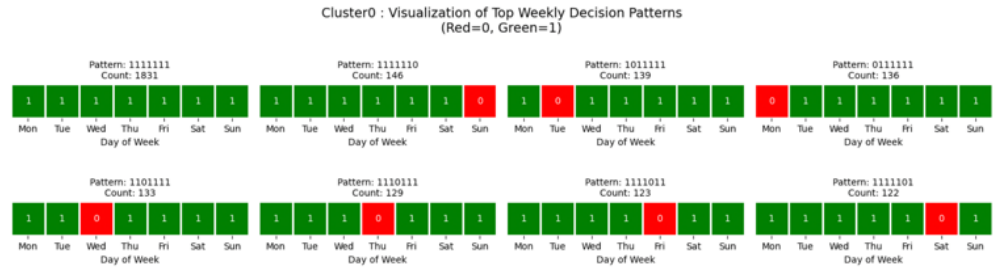

Fig. 7. Cluster 0 : visualization of top weekly decision patterns.

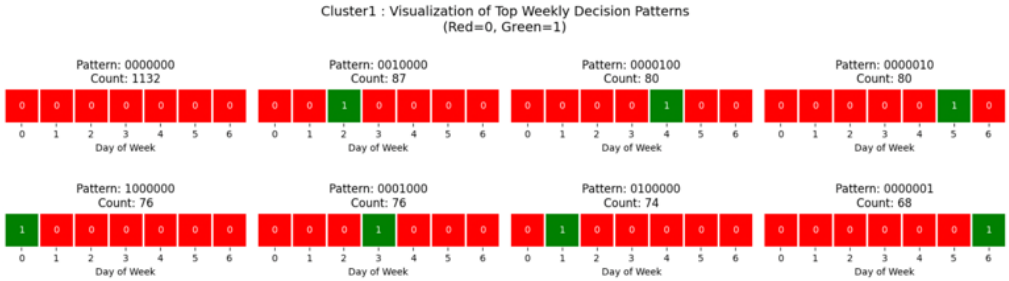

Fig. 8. Cluster 1 : visualization of top weekly decision patterns.

1) Cluster 0: Consistent high-attendance strategy: Fig. 5 shows that Cluster 0 agents exhibit a steadily increasing probability of attendance throughout the week, starting from approximately 0.28 on Mondays and peaking near 0.92 on Sundays. This clear upward trend indicates that agents in this cluster are not only frequent attendees but also increasingly confident in their decision to attend as the week progresses.

When mapped to their behavioral sequences, these agents adopt highly consistent attendance patterns. Fig. 7 illustrates the dominant weekly decision patterns observed in Cluster 0. Most agents in this group exhibit highly consistent attendance behaviors, frequently following the pattern 1111111 or slight variants such as 1111110 and 1111101, indicating that they tend to attend the bar almost every day of the week. This persistent attendance pattern suggests a highly exploitative strategy where agents have learned that the long-term expected reward of attending is greater than skipping, even under conditions of possible congestion.

Cluster 0 achieved the highest rewards, indicating that, under the modelled assumptions, a stable high-attendance strategy yields higher cumulative rewards. Despite incurring regular overcrowding penalties, their consistent exploitation of the venue yielded superior cumulative gains, demonstrating that persistent participation dominates more adaptive strategies.

2) Cluster 1: Consistent high-attendance strategy: By contrast, the weekly attendance patterns in Fig. 6 for Cluster 1 agents reveal a markedly different behavioural dynamic. Attendance probabilities remain below 0.20 across all weekdays, with only a modest rise on weekends (peaking at 0.16 on Sundays). This indicates a strong inclination to avoid attendance most of the time, suggesting either: 1) heightened sensitivity to congestion penalties or 2) a persistent underestimation of the long-term benefits of frequent attendance.

In addition, agents in Cluster 1 demonstrate more diverse and selective attendance patterns, as shown in Fig. 8. Their top decision sequences often include multiple "0" entries (non-attendance days), suggesting an explorative or cautious strategy. Such agents appear to attend intermittently to avoid congestion penalties, but this moderation also limits their cumulative reward growth compared to Cluster 0.

Their decision sequences often contain multiple "0" days, resulting in sparse attendance policies. With a mean *QGo* value of just 3.54 and a very low *GoRate* (0.09), Cluster 1 agents embody a conservative or exploratory policy, where the avoidance of potential overcrowding outweighs the pursuit of maximum rewards. This strategy, while adaptive in balancing risk and reward, is ultimately suboptimal under the given simulation settings. Indeed, their average cumulative reward (181.03) lags significantly behind Cluster 0, highlighting that overly cautious attendance reduces the opportunity for long-term gain.

E. Discussion of Weekly Decision Patterns

A granular analysis of the top weekly decision sequences within each cluster reveals profound insights into the learned strategies and their potential real-world correlates.

1) Cluster 0: The persistence of high-frequency patterns: Within Cluster 0 (Ultra-Frequent), a single dominant pattern emerged: an "Always go" strategy. Designated Pattern 1, it was used 1,831 times, far more than any other. Its persistence indicates that agents learned that the high reward of a successful visit outweighs occasional penalties. This reflects a necessity-driven and risk-tolerant strategy for a venue that is an essential daily routine.

2) The emergence of rest-day patterns: Other patterns in Cluster 0 and patterns in Cluster 1 (Optimized) frequently featured an absence, or "Stay" decision, on a specific day—most mapped to a Sunday. This emerging "rest day" is a significant finding. This interpretation is speculative and intended as an analogy rather than a claim about real-world behaviour.

From a behavioral standpoint, this can be interpreted in two ways:

- Learned Depletion: Agents may have implicitly learned that after a sustained period of attendance, the marginal utility of visiting diminishes or the probability of fatigue (both their own and systemic overcrowding) increases. A rest day serves as a strategic reset.
- Real-World Rhythms: This pattern strongly mirrors human social behavior, where Sunday is culturally designated as a day of rest and preparation for the upcoming week. The agents have effectively discovered that avoiding the venue on this day is a robust strategy, possibly because it aligns with a period of lower overall utility or higher opportunity cost for going out.

3) Strategic heterogeneity in cluster 1: The lower frequency of pattern usage for non-dominant strategies, particularly in the Optimized cluster (Cluster 1), highlights the role of strategic flexibility. The "sudden" decision to go out on a day that is typically a rest day, or to deviate from a routine, can be interpreted as the model's representation of stochastic real-world influences.

F. Theoretical and Managerial Implications

The results of this study offer several implications that extend beyond the specific simulation setting, contributing to theory, methodology, and retail practice. At the theoretical level, the findings reinforce the El Farol Bar problem's central insight that bounded rational agents can achieve stable collective outcomes through repeated learning rather than centralized coordination. Importantly, the observed equilibrium emerges in response to perceived congestion rather than objective capacity constraints, highlighting the role of subjective evaluation in shaping attendance decisions.

This suggests that equilibrium behavior in congestion games may be driven as much by learned perceptions as by physical limits, particularly in recurrent decision environments such as retail visitation.

From a methodological perspective, the analysis demonstrates the value of combining reinforcement learning with post-learning behavioral clustering. While Q-learning governs individual adaptation during the simulation, clustering enables ex post identification of stable behavioral archetypes that are not explicitly encoded in the model. The results further illustrate how reward shaping functions as a critical design lever, influencing not only convergence properties but also the diversity of emergent strategies within the agent population. This layered analytical approach supports more interpretable insights than aggregate performance measures alone.

Taken together, these findings imply that retail space management should not be viewed solely as a problem of maximizing customer throughput or maintaining high occupancy levels. Instead, the way customers experience and interpret congestion appears to play a central role in shaping attendance behavior. Design choices such as seating configuration may influence perceived occupancy, spatial layout, and time-based incentives may therefore serve as practical tools for influencing perceived occupancy. In settings where customers make repeated visit decisions, such perception-sensitive approaches are likely to offer greater flexibility than strategies based exclusively on fixed capacity targets.

V. Conclusion

This work explored how adaptive learning influences attendance behavior in a congested retail environment by framing the coffee shop dilemma as a multi-agent reinforcement learning problem. Rather than treating customer behavior as a function of visit frequency alone, the analysis indicates that collective regularities emerge from how individuals gradually interpret and respond to congestion experiences. More specifically, differences in how agents experience satisfaction or dissatisfaction under varying occupancy conditions influence how their decisions evolve over time. These differences are reflected in the emergence of stable attendance routines, as well as in persistent variation in behavior among agents exposed to the same environment.

Viewed more broadly, these results relate to existing work on congestion games and bounded rationality. In line with the El Farol Bar problem, the findings suggest that coordinated outcomes can arise without centralized control or complete information. Notably, the equilibrium observed in this study is guided by learned perceptions of congestion rather than by explicit awareness of physical capacity limits. This observation underscores the role of subjective evaluation in recurrent decision environments and supports the inclusion of perception-driven learning mechanisms in models of collective behavior involving shared resources.

The study also carries implications for retail space optimization and customer flow management. The presence of distinct behavioral patterns indicates that customers may respond differently to congestion even when exposed to identical operational conditions. As a result, managing retail spaces solely through capacity targets or throughput maximization may overlook important behavioral dynamics. Paying closer attention to how occupancy is experienced—through layout decisions, seating arrangements, or time-based incentives—may therefore provide additional flexibility in influencing customer behavior. Further research may extend this modeling approach by considering richer state descriptions, incorporating social interaction effects, or grounding the model in observational retail data to better examine adaptive behavior in complex service settings.

Moreover, this study is subject to limitations. The model abstracts away social interaction, heterogeneous preferences, and empirical calibration. Future work may integrate observational retail data, richer state representations, and adaptive capacity thresholds to further validate the findings.

References

[1] W. B. Arthur, "Inductive reasoning and bounded rationality," The American Economic Review, vol. 84, no. 2, pp. 406–411, 1994.

[2] D. Challet and Y.-C. Zhang, "Emergence of cooperation and organization in an evolutionary game," Physica A: Statistical Mechanics and its Applications, vol. 246, no. 3-4, pp. 407–418, 1997.

[3] S.-H. Chen and U. Gostoli, "Agent-based modeling of the el farol bar problem," in Simulation in Computational Finance and Economics: Tools and Emerging Applications, IGI Global Scientific Publishing, 2013, pp. 359–377.

[4] J. Schosser, "Fairness in the use of limited resources during a pandemic," PLoS ONE, vol. 17, e0270022, 2022. doi: 10.1371/journal.pone.0270022.

[5] P. Guarnieri and L. Spadoni, "Norms and anti-coordination: Elicitation and priming in an El Farol Bar Game experiment," Dept. Econ. Manag., Univ. Pisa, Pisa, Italy, Discussion Paper no. 303, 2024.

[6] S.-H. Chen and U. Gostoli, "Coordination in the El Farol Bar problem: The role of social preferences and social networks," Journal of Economic Interaction and Coordination, vol. 12, no. 1, pp. 59–93, 2017.

[7] C. Atilgan, A. R. Atilgan, and G. Demirel, "Collective behavior of El Farol attendees," Advances in Complex Systems, vol. 11, no. 4, pp. 629–639, 2008.

[8] G. Zheng, W. Cai, G. Qi, J. Zhang, and L. Chen, "Optimal coordination in Minority Game: A solution from reinforcement learning," arXiv preprint arXiv:2312.14970, 2023.

[9] M. Apte, P. Datar, K. Kale, and P. R. Deshmukh, "Dynamic retail pricing via Q-learning—A reinforcement learning framework for enhanced revenue management," in 2025 1st International Conference on AIML-Applications for Engineering & Technology (ICAET), 2025, pp. 1–5.

[10] S. Huang and Y. Yang, "Reinforcement learning optimization strategies for dynamic pricing and inventory control in e-commerce retail," IEEE Transactions on Artificial Intelligence, vol. 5, no. 2, pp. 123–135, 2024.

[11] P. Kossack, "An Emotional Machine goes into a Bar," 2024. [Online]. Available: https://www.researchgate.net/profile/Philip-Kossack/publication/385490885_An_Emotional_Machine_goes_into_a_Bar/links/6726095b5852dd723ca452c9/An-Emotional-Machine-goes-into-a-Bar.pdf

[12] P. Michailidis, I. Michailidis, C. R. Lazaridis, and E. Kosmatopoulos, "Traffic Signal Control via Reinforcement Learning: A Review and Innovations," arXiv preprint arXiv:2406.12345, 2024.

[13] T. C. Mehta and B. Shah, "Study on effects of perceived crowding in retail spaces and futuristic design solutions for positive consumers' experience and retention," Onomázein, vol. 61, pp. 405–415, 2023.

[14] D. Buzali, S. Elizondo, S. Muñiz, and O. Sánchez, "Simulation-optimization of a coffee shop in business district: A case study of Starbucks in Mexico City," International Journal of Operations Research, vol. 15, no. 3, pp. 45–58, 2023.

[15] L. Grando, J. R. E. Leite, and E. L. Ursini, "Agent-based simulation for drone charging in an internet of things environment system," arXiv preprint arXiv:2509.10867, 2025.

[16] A. Borsos, A. Carro, A. Gitelmo, M. Hinterschweiger, J. Kaszowska-Mojsa and A. Uluc, "Agent-based modeling at central banks: recent developments and new challenges," Bank of England Staff Working Paper, no. 1122, Feb. 2025.