Trading-Aware Agents in Sugarscape

A Deep Reinforcement Learning Approach to Adaptive Economic Behavior

Thanh Binh Lai

School of Accounting and Finance, Economics Research Group,
University of Vaasa, Wolffintie 32, Vaasa, 65200, Finland.

Contributing authors: laithanh@uwasa.fi;

Abstract

Trading is the lifeblood of modern economies, yet most agent-based models still represent it as an afterthought—something that happens *after* agents decide where to move or what to consume. This paper argues that such separation no longer reflects the reality of today's interconnected, data-driven markets, where trading awareness is embedded in every economic decision. To address this gap, we extend the classical *Sugarscape* model by introducing **trading-aware agents** that learn to jointly optimize movement and exchange using Deep Reinforcement Learning (DRL). Through this integration, the model updates one of the foundational agent-based frameworks in computational economics to reflect how modern economies operate: agents move not only toward resources but also toward opportunities to trade.

Using Proximal Policy Optimization (PPO), agents learn adaptive policies that internalize both spatial and market information. Simulation results show that trading-aware DRL agents achieve higher carrying capacity, faster price convergence, and more sustainable resource utilization than their rule-based counterparts. The approach transforms *Sugarscape* into a learning-based experimental laboratory for studying equilibrium formation, inequality dynamics, and policy interventions under realistic trading behavior. Beyond reproducing classical results, this new framework enables a generation of agent-based economic studies that mirror the complexity of real-world trading societies, where adaptation and exchange are inseparable.

Keywords: Agent-based modeling; Deep reinforcement learning; Sugarscape; Market equilibrium; Computational economics; Policy simulation

1 Introduction

In today's global economy, almost every economic action is intertwined with trade. From energy markets to digital platforms, individuals, firms, and algorithms continuously adjust their behavior in anticipation of future exchanges. Yet in many computational models of economic behavior—including the classical *Sugarscape* framework of Epstein and Axtell [1]—trading is treated as a secondary process, something that occurs only *after* agents decide where to move or what to consume. This simplification made sense in the 1990s, when computational limits constrained model complexity, but it now limits our ability to study economies in which trading itself shapes movement, learning, and survival.

Recent advances in DRL provide an opportunity to close this conceptual gap. Modern DRL algorithms such as PPO [2] and Soft Actor–Critic (SAC) [3] allow agents to learn stable, high-dimensional policies that jointly account for mobility, exchange, and adaptation within dynamic environments. These methods have already been applied to economic and multi-agent contexts—such as the *AI Economist* framework for adaptive taxation and market design [4], and multi-agent DRL (MARL) studies of resource sharing and social dilemmas [5, 6]—demonstrating that learning agents can endogenously develop cooperative and coordinated behaviors that were previously manually encoded in agent-based models. Embedding such adaptive learning into economic simulations thus represents both a theoretical update and a methodological necessity for studying modern economies driven by strategic, trading-aware behavior.

This paper introduces a **trading-aware *Sugarscape*** that unifies movement and trade through DRL. Agents learn with the PPO algorithm to maximize long-term welfare in a world of renewable resources, spatial heterogeneity, and market exchange. The resulting framework preserves the explanatory power of the original model while extending its scope to capture emergent equilibria, inequality dynamics, and sustainability under adaptive market behavior.

The contribution is twofold. First, it provides a **conceptual update** to the canonical *Sugarscape* model, replacing static heuristics with adaptive learning mechanisms inspired by bounded rationality [7]. Second, it establishes a **computational laboratory** for exploring policy experiments relevant to contemporary trading societies—such as taxation, resource pricing, and inequality mitigation—within a unified, data-generating environment. In doing so, this research responds to the growing call for agent-based economic models that reflect the adaptive, interconnected, and trading-oriented nature of the global economy [8].

The remainder of this paper proceeds as follows. Section 2 reviews related work in economic modeling, learning, and game-theoretic foundations. Section 3 details the simulation environment, agent design, and DRL training procedure. Section 4 reports comparative results on carrying capacity, market stability, efficiency, and inequality. Section 5 discusses theoretical and practical implications for computational economics, policy design, and game simulation.

2 Literature Review

2.1 Economic Models

Sugarscape represents one of the earliest and most influential attempts to operationalize microeconomic theory within an agent-based environment. In its variant Constant Growback with Traders, agents maximize a Cobb–Douglas utility function over two resources: *sugar* and *spice*—and engage in bilateral barter whenever their marginal substitution rates (MRS) differ [1]. Through decentralized exchanges, the system generates emergent prices that approximate the Pareto-efficient equilibria predicted by classical microeconomics. Agents trade until no further mutually beneficial exchanges are possible, equalizing their MRS and producing a stable equilibrium price [9]. Due to this elegant formulation, *Sugarscape* became a canonical platform for studying self-organized markets and wealth dynamics.

Despite its foundational role, the original design makes simplifying assumptions that limit behavioral realism. The spatial resource landscape is static, featuring fixed abundance peaks and predictable migration patterns. Agents move according to a hard-coded, myopic rule: within their vision range, they select the cell that maximizes immediate welfare;

$$W(sugar, spice) = sugar^{\alpha} spice^{1-\alpha},$$

where α reflects metabolic preferences.

$$\alpha_i = \frac{m_i^{sugar}}{m_i^{sugar} + m_i^{spice}},$$

Trading occurs only *after* movement, meaning that the decision to relocate does not consider potential future trades. Consequently, the model captures static equilibrium but omits the dynamic coupling between spatial movement, adaptive expectations, and market formation.

This omission raises an important question central to modern computational economics: how would outcomes change if agents could integrate trading expectations directly into their spatial strategy? Recent literature in multi-agent reinforcement learning suggests that agents can indeed learn to internalize such forward-looking motives. For example, Leibo et al. [5] shows conditions under which MARL agents become more or less cooperative in sequential social dilemmas, while Zheng et al. [4] demonstrate that reinforcement-learning agents can jointly discover optimal trading and taxation policies in simulated economies.

The present research addresses this behavioral gap by designing agents whose decision sets include not only spatial movement but also anticipatory trading motives. In doing so, it transforms *Sugarscape* from a static model of exchange into a **trading-aware agent-based economy**, thereby aligning micro-level decision-making more closely with the adaptive and expectation-driven behavior observed in real-world markets.

2.2 Learning in Agent-Based Economic Models

A growing body of work has incorporated learning algorithms into agent-based economic models to replace fixed heuristics with adaptive behavior. Brearcliffe and Crooks [10] extended the NetLogo implementation of *Sugarscape* to include evolutionary computation, Q-learning, and SARSA agents. Their experiments demonstrated that learning can alter emergent outcomes, but performance depends heavily on reward specification and environmental design—indeed, rule-based agents occasionally outperformed learning ones. Notably, their models emphasized foraging and conflict rather than market behavior; trading was not part of the learned action space.

Jäger [11] proposed a neural-network framework for agent-based models, applying it to *Sugarscape* as a test case. His agents learned movement policies guided by a utility-based reward function, successfully reproducing canonical Sugarscape dynamics such as clustering and heavy-tailed wealth distributions. However, as in Brearcliffe and Crooks, trading remained an exogenous rule rather than a learned behavior. These studies thus confirmed that reinforcement learning can replicate classical emergent phenomena but did not extend it to market optimization.

The present work advances this literature by embedding trading directly into the learning process. Agents learn trading strategies by observing neighbours' MRS and local prices within their state space, enabling the DRL policy to optimize over both movement and exchange. By employing the PPO algorithm [2], this framework stabilizes training in high-dimensional environments and allows agents to discover equilibrium-seeking behaviors.

2.3 Reinforcement Learning Agents

Within the reinforcement-learning paradigm, each agent faces a Markov Decision Process (MDP) defined by the tuple (S, A, R, P, γ). Here S denotes the set of possible states (including local resources, neighbouring agents, and internal holdings), A the available actions, R the reward function, P the transition dynamics, and γ the discount factor. The agent seeks a policy $\pi(a|s)$ that maximizes expected discounted return

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right].$$

In the extended *Sugarscape*, the action space includes both directional moves and implicit trading decisions, while the reward function aligns with economic utility and survival. Through repeated interaction, the PPO algorithm updates policy parameters to improve long-term welfare, effectively approximating a solution to the MDP that governs the environment [2].

This learning framework transforms the Sugarscape from a static heuristic system into a dynamic economy of adaptive agents. Each agent acts as a decentralized optimizer, continuously updating its strategy to balance resource harvesting, trading opportunities, and survival—a closer analogue to real economic actors.

2.4 Game Theory and the Agent Movement Rule

The classical Sugarscape rules can be reinterpreted through a game-theoretic lens. The trading mechanism (*Rule T*) functions as a bilateral bargaining game: two agents exchange sugar and spice until both reach the Pareto frontier, where their MRS converge. The trading price p—the geometric mean of their MRS—ensures that each transaction lies within the core of the game, yielding mutual improvement and local market clearing [1].

Movement decisions (*Rule M*) can similarly be viewed as repeated spatial games against nature. Each agent chooses the cell that maximizes instantaneous utility, effectively performing a greedy search for a local Nash equilibrium in resource allocation. However, because agents cannot forecast future trading partners, these myopic rules miss potential gains from anticipated exchange.

By endowing agents with reinforcement-learning policies, the present study unifies these two layers of decision-making. Agents now evaluate spatial moves in light of expected trading payoffs, transforming movement into a strategic game that integrates logistics and market foresight. This synthesis bridges economic theory and modern AI: bounded-rational agents learn to approximate general-equilibrium outcomes through decentralized adaptation, demonstrating how micro-level learning can yield macro-level order [4–6].

2.5 Research Hypotheses and Questions

Building on the theoretical framework of reinforcement learning in the Sugarscape environment, this study is guided by the following hypotheses and research questions:

- **H1 (Carrying Capacity Advantage).** DRL agents that jointly optimize movement and trading will achieve a higher carrying capacity (survival rate) than rule-based agents under identical environmental conditions.
  Rationale: The integrated policy enables DRL agents to anticipate future barter opportunities and allocate resources more efficiently, reducing starvation and raising the sustainable population level.
- **H2 (Market Equilibrium and Price Stability).** DRL-driven societies will exhibit faster convergence of market prices and lower volatility compared with heuristic rule-based societies.
  Rationale: Adaptive trading strategies allow DRL agents to arbitrage local price discrepancies more rapidly, pushing the decentralized market closer to Pareto-efficient equilibrium.
- **H3 (Resource Efficiency and Welfare).** DRL agents will utilize environmental resources more efficiently—harvesting a greater fraction of regrown sugar and spice—and generate higher aggregate welfare (total wealth) than rule-based agents.
  Rationale: Learning-based policies internalize both consumption needs and opportunity costs, enabling near-optimal allocation of resources through proactive foraging and trade.
- **H4 (Inequality Dynamics).** The wealth distribution of DRL agents will stabilize at a lower Gini coefficient than that of heuristic populations.

Rationale: Extensive trading redistributes resources from surplus to deficit agents, mitigating inequality that would otherwise persist under static rules.

Research Questions

(i) How does embedding trading intent into movement decisions affect the emergent macroeconomic equilibrium?

(ii) Does learning-based adaptivity improve social welfare, equality and sustainability without central coordination?

2.6 Theoretical Framing: Bounded Rationality and Adaptive Learning

Classical economic agents are often modeled as perfectly rational optimizers. By contrast, the canonical *Sugarscape* already embodies bounded rationality: agents have limited vision, act on local information, and follow fixed heuristics for movement and trade [1]. Building on this foundation, the present study introduces a *learning-based form of bounded rationality*, in which agents adapt policies through experience rather than fixed rules. Each agent updates its behavior according to observed outcomes, linking reinforcement learning with adaptive expectations and evolutionary dynamics. This enables bounded-rational agents to approach equilibrium-like regularities—such as stable local prices and trading networks—without assuming global knowledge.

A key extension concerns spatial behavior. While the original model restricts each site to single occupancy,[1] the present framework relaxes this constraint, allowing controlled co-location. Agents learn congestion–access trade-offs, discovering when proximity fosters trade and when crowding imposes costs. This shift from hard-coded exclusion to learned spatial coordination extends bounded rationality from the cognitive to the institutional level, emphasizing adaptation and feedback as core drivers of emergent order.

3 Methodology

The methodological framework directly operationalizes the hypotheses outlined in Section 2. Each experiment tests one or more of the proposed hypotheses (H1–H4) by comparing rule-based and DRL-driven societies under controlled environmental conditions. The following subsections describe the simulation environment, agent architecture, training procedures, and evaluation metrics corresponding to these hypotheses.

3.1 Simulation Environment and Model Design

The environment builds upon the *Sugarscape Constant Growback Model with Traders* implemented in Mesa [12]. A two-dimensional 50×50 grid represents a landscape of two renewable resources—sugar and spice—each cell endowed with maximum capacities that regenerate at a rate of one unit per step. Spatial heterogeneity is introduced

[1] In *Sugarscape*, agents may move only to empty cells within their vision range; see Epstein and Axtell (1996, ch. 2).

through resource "peaks," forming areas of high abundance separated by low-yield valleys. In contrast to the original toroidal configuration, boundary wrapping is disabled so that agents must adapt to edges.

Two types of spatial environments are considered to evaluate agents' adaptability. In the **deterministic resource landscape** setting, the spatial distribution of sugar and spice is constant across all simulation episodes, featuring two stable resource peaks separated by low-yield valleys. This configuration enables comparison with the classical *Sugarscape* baseline. In contrast, the **stochastic resource landscape** redraws resource endowments randomly at the start of each episode while preserving the same statistical distribution. This regime tests whether learned policies generalize to unseen spatial layouts and resource heterogeneity.

A population of $N = 500$ heterogeneous agents inhabits the grid. Each agent is characterized by: (i) a vision range v (cells visible in each von Neumann direction), (ii) metabolic rates m for sugar and spice, (iii) initial endowments w, and (iv) *age* and survival status *alive*/*death*. Agents consume resources equal to their metabolism each step and die when any stock reaches zero.

Unlike the continuous-replacement regime of Epstein and Axtell [1], agents in the baseline simulations are *not respawned* after death. This non-replacement setting enables a clear assessment of population decline, resource depletion, and the environment's effective carrying capacity under different learning policies. For analyses of market dynamics, however, the model is extended to include fertility and finite lifespans following the demographic formulation of Epstein and Axtell [1]. This demographic extension produces continuous population turnover, allowing long-run equilibrium processes to emerge as generational learning and market renewal interact over extended horizons. Together, these complementary regimes capture both short-term ecological limits and long-term economic equilibria within a unified framework.

In the rule-based baseline, agents move according to *Rule M*: within their vision, they select the nearest cell maximizing welfare $W(sugar, spice)$. Adjacent agents with differing MRS then trade following *Rule T*, negotiating an exchange rate $p = \sqrt{\text{MRS}_i \text{MRS}_j}$ and trading until utilities equalize. In the DRL extension, agents still follow Rule T when trading, but Rule M is no longer limited to maximizing welfare through harvesting alone. Instead, movement decisions are learned through PPO, allowing agents to seek locations that offer both resource gains and improved trading opportunities. Thus, spatial exploration and market interaction become jointly optimized, as agents learn to anticipate and exploit profitable exchange possibilities.

3.2 Observation, Action, and Reward Design

3.2.1 Observation Space

Each agent receives a Markovian observation vector containing both external and internal states. Externally, the agent perceives a four-channel tensor representing its local neighborhood: (1) sugar amounts, (2) spice amounts, (3) welfare surplus from harvesting each visible cell, and (4) neighboring agents' MRS values if present. Internally, a six-element vector encodes the agent's own sugar and spice stocks, metabolic

rates, vision range, and current MRS. Formally, the combined observation is

$$o_t = \{\text{tensor}_{(4v+1)\times 4},\ \text{vector}_{\mathbb{R}^6}\}.$$

This structure provides spatial, economic, and personal context, ensuring that the learning process satisfies the Markov property.

3.2.2 Action Space

Agents can move one cell per step in the von Neumann neighborhood:

$$A = \{\text{North}, \text{South}, \text{East}, \text{West}, \text{Stay}\}.$$

Trading is not an explicit action; it occurs automatically when adjacent agents can improve joint utility. Therefore, the DRL policy influences market outcomes indirectly by positioning agents advantageously for exchange, letting PPO discover when to approach or avoid trading partners.

3.2.3 Reward Function

The reward function operationalizes the link between individual incentives and collective outcomes, allowing agents to learn behavior that balances consumption, movement, and survival. The Cobb–Douglas Utility Reward is adopted to quantify per-step welfare improvements derived from resource acquisition and metabolic maintenance.

Each agent's instantaneous reward equals the marginal gain in welfare from consumption relative to its previous state:

$$r_t = W(sugar'_t, spice'_t) - W(sugar_t, spice_t) - 1 \times death,$$

where $W(sugar, spice) = sugar^{\alpha}\, spice^{1-\alpha}$ represents the Cobb–Douglas utility function over the two resources. The exponent $\alpha \in [0, 1]$ denotes the agent's metabolic preference for sugar relative to spice.

When evaluating potential actions, the destination state (s'_t, e'_t) denotes the expected quantities of sugar and spice remaining after movement and metabolic deduction. Accordingly, $W(sugar'_t, spice'_t)$ reflects the agent's *expected welfare* at the candidate cell, while $W(sugar_t, spice_t)$ represents its *current welfare*. The terminal penalty term (-1) introduces a negative reward upon death, stabilizing temporal-difference learning and preventing myopic strategies [13, 14]. This specification aligns the reinforcement signal with economic intuition: agents are rewarded for transitions that yield higher future utility and penalized for depleting resources or reaching fatal states.

By framing reinforcement feedback in terms of Cobb–Douglas welfare, this scheme embeds classical microeconomic behavior—diminishing marginal utility and balanced consumption—within the learning process. It enables agents to discover spatial and trading policies that jointly maximize lifetime utility under resource constraints,

thereby bridging traditional welfare economics and adaptive reinforcement learning dynamics.

3.3 DRL Policy Training

3.3.1 MDP Formulation

Each agent solves a Markov Decision Process with state $s_t = o_t$, action $a_t \in A$, and reward r_t. The objective is to learn a policy $\pi(a|s)$ maximizing expected discounted return

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right],$$

with $\gamma \in [0, 1]$ is the discount factor. The Bellman expectation equation

$$V_{\pi}(s) = \mathbb{E}_{a\sim\pi, s'\sim P}[r(s, a) + \gamma V_{\pi}(s')]$$

defines the critic's value target.

3.3.2 Proximal Policy Optimization (PPO)

Training uses the on-policy actor–critic PPO algorithm [2]. The clipped objective

$$L_{\text{clip}}(\theta) = \mathbb{E}_t\Big[\min\big(\rho_t(\theta)\hat{A}_t,\ \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\big)\Big],$$

where $\rho_t(\theta) = \pi_\theta(a_t|s_t)/\pi_{\text{old}}(a_t|s_t)$, stabilizes updates by constraining policy changes. The overall loss combines policy, value, and entropy terms, optimized with the Adam algorithm. A shared neural network with three 256-unit ReLU layers outputs both action probabilities and value estimates. Hyperparameters include learning rate 5×10^{-4}, batch size 2048, and 3 epochs per PPO update.

3.3.3 Training Process

Training was implemented via Unity ML-Agents Toolkit [15], using centralized training with decentralized execution. All agents share parameters, producing efficient sample usage and convergence within roughly 5×10^5 agent-steps. Episodes terminate when either all agents die or a time limit $T = 150$ is reached. After convergence, the trained policy is frozen for evaluation.

3.4 Model Modifications

Several modifications ensure comparability and stable reinforcement learning performance.

- **Finite horizon:** each episode terminates after $T = 150$ simulation steps in the baseline regime (and $T = 500$ in demographic extension), rather than running indefinitely as in the original Sugarscape. This finite horizon prevents long-run population depletion and spatial dispersion that reduces trading opportunities, while enabling episodic reinforcement-learning training.

- **No respawn:** dead agents are not replaced within an episode. This non-replacement setting allows clear measurement of population decline, resource depletion, and the environment's effective carrying capacity under different policies.
- **Demographic extension (fertility regime):** for market-price and long-run equilibrium analysis, the model incorporates fertility and finite lifespans following Epstein and Axtell [1]. Continuous population turnover enables generational learning and market renewal to coevolve over extended horizons.
- **Randomized maps:** spatial resource distributions vary across runs to test generalization and the robustness of learned strategies.
- **Population density:** $N = 500$ agents ensures frequent trading encounters and stable aggregate dynamics.
- **Equal vision:** all agents share a fixed vision range of $v = 10$ cells for consistent comparison across models.
- **Deterministic evaluation:** stochastic exploration is disabled when testing trained policies to ensure reproducible welfare and price outcomes.

Table 1 summarizes key differences between the original Sugarscape and the DRL extension.

Table 1: Comparison of baseline and DRL model parameters.

Parameter	Base-line	DRL extension
Resource distribution	Fixed	Fixed / Random
Population size	200 (respawn)	500 (no respawn)
Growth rate of resources	1	1
Initial resources	25–50	5–25
Metabolism	1–4	1–5
Vision radius	1–6	10
Time limit	1000	150/500

3.5 Evaluation Metrics and Experimental Design

Model performance is assessed on the same metrics used by Epstein and Axtell [1]:

1. **Carrying capacity:** number of agents alive at episode end; mean and variance computed over 50 runs.
2. **Market prices:** time series of sugar-to-spice exchange rates; convergence speed and volatility measure market stability.
3. **Efficiency:** total resource consumption, average wealth, and aggregate societal wealth quantify output and welfare.
4. **Inequality:** Gini coefficients computed on $W(sugar_i, spice_i)$ measure the general level of inequality in the economy, while their evolution over time provides information on how redistribution and trading dynamics shape wealth distribution during simulation.

5. **Dynamic patterns:** spatial dispersion and the consume–regrow ratio

$$CR_t = \frac{\text{resources grown in } t}{\text{resources consumed in } t},$$

evaluate sustainability; $CR \approx 1$ indicates optimal harvesting.

Each condition—rule-based versus DRL, static versus stochastic maps—is replicated 50 times with identical seeds. Statistical significance is verified via paired t-tests at $p < 0.01$.

3.6 Learning and Training Setup

Training hyperparameters are held constant across experiments: discount factor $\gamma = 0.99$, generalized advantage estimation $\lambda = 0.95$, entropy coefficient $\beta = 0.005$ (linearly annealed to zero), clipping ratio $\epsilon = 0.1$ (linearly annealed), batch size 2,048, rollout buffer size 20,480, and three epochs per PPO update. The environment hosts $N = 500$ agents with vision $v = 10$, resource regrowth rate 1, and random initial endowments. Each run proceeds for 5×10^6 environment steps, with performance summaries logged every 5×10^4 steps. All experiments are conducted locally on an Apple Mac mini (M4, 16 GB unified memory) using the Unity ML–Agents Toolkit with fixed random seeds to ensure reproducibility.

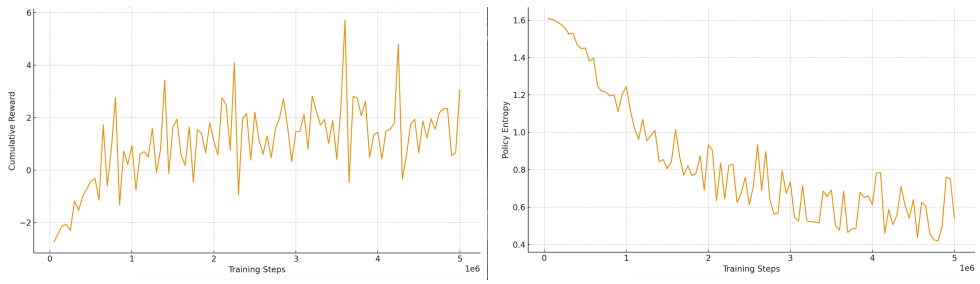
For learning diagnostics, we monitor two key indicators: (i) cumulative environment reward and (ii) policy entropy. The first reflects long-term performance improvements, while the second tracks the transition from exploration to exploitation. Convergence is assessed using a rolling-mean slope threshold ($< 5 \times 10^{-6}$) on the cumulative reward curve. Macro-level performance metrics—carrying capacity, mean wealth, equilibrium price dispersion, Gini coefficient, and episode length—are subsequently evaluated using the converged policies.

3.7 Learning Dynamics and Robustness

Before analyzing macroeconomic outcomes, we first verify that the DRL policy achieved stable convergence under the **Cobb–Douglas Utility** reward function. This reward structure links immediate resource consumption to utility increments, allowing agents to learn balanced foraging and trading behaviors consistent with microeconomic theory.

As shown in Figure 1, cumulative reward improved from -2.73 at 50,000 steps to 3.04 by the end of training, with a mean cumulative reward of 1.05 (s.d. 1.48). Entropy declined monotonically from 1.61 to 0.54, stabilizing near 2.3×10^6 steps, indicating a smooth transition from exploration to exploitation and convergence toward a deterministic, high-utility policy. The evolution of both metrics confirms that learning dynamics were well-behaved: the policy steadily increased expected returns while avoiding premature convergence or oscillatory behavior.

To assess robustness, the entropy and clipping coefficients (β, ϵ) of the PPO objective were perturbed by $\pm 20\%$. Final cumulative rewards varied by less than 10% from the baseline, and convergence timing shifted by fewer than 0.2×10^6 steps. These results



(a) Utility scheme: cumulative reward. (b) Utility scheme: policy entropy.

Fig. 1: Learning diagnostics for the Cobb–Douglas Utility scheme, showing convergence of cumulative reward and entropy over training steps.

demonstrate that the training process is stable and that the reported macroeconomic outcomes reflect genuine behavioral learning rather than sensitivity to hyperparameter tuning.

3.8 Computational Complexity and Scalability

The computational cost of the DRL-augmented *Sugarscape* arises mainly from the PPO training loop, which scales approximately linearly with the number of agents N, the episode length T, and the number of gradient updates per iteration E. Formally, the expected training complexity is:

$$O(N \times T \times E \times P),$$

where P denotes the number of policy parameters.

For this research, with $N = 500$ agents, $T = 150$, and $E = 3$ PPO epochs, the total training horizon of roughly 5×10^5 agent-time steps was sufficient for convergence. Each PPO iteration requires forward and backward passes through a three-layer neural network of 256 neurons per layer; on a modern GPU, this equates to approximately 0.02 s per 1,000 environment steps, or roughly three hours of wall-clock training per policy.

Scalability can therefore be achieved by:

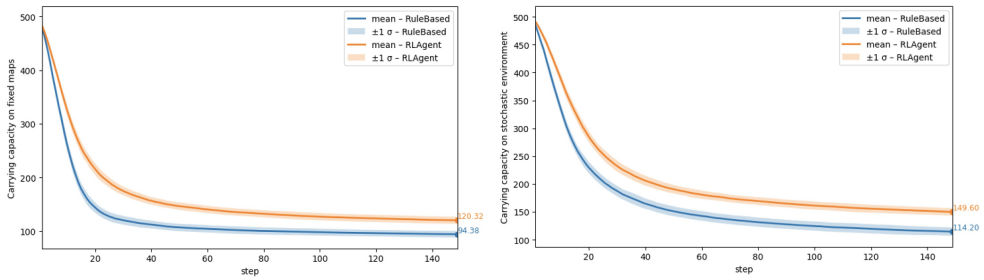
1. Parallelizing environment rollouts across multiple grid instances;
2. Parameter sharing among homogeneous agents (already implemented); and
3. Decentralized or asynchronous training where subsets of agents update independently.

These features make the model computationally feasible for large-scale experiments involving thousands of agents or multi-commodity extensions.

4 Results

4.1 Carrying Capacity and Survival Dynamics

To assess Hypothesis H1, the analysis first compares population survival across DRL and rule-based societies, examining whether learning-based agents sustain higher carrying capacity under equivalent environmental constraints. The carrying capacity of the system—defined as the number of agents surviving to the end of an episode—serves as a fundamental indicator of resource-use efficiency. Figure 2 compares population survival under rule-based and DRL-driven societies across both fixed and stochastic resource maps.



(a) Deterministic resource landscape. (b) Stochastic resource landscape.

Fig. 2: Comparison of carrying capacity between DRL and rule-based agents under static and stochastic environments.

Both populations experience an initial die-off as resources become scarce, followed by stabilization once equilibrium between consumption and regrowth is reached. However, DRL agents consistently sustain a larger population, approximately 28% higher under fixed maps and 32% under stochastic conditions. Paired t-tests confirm the statistical significance of this difference ($p < 0.001$).

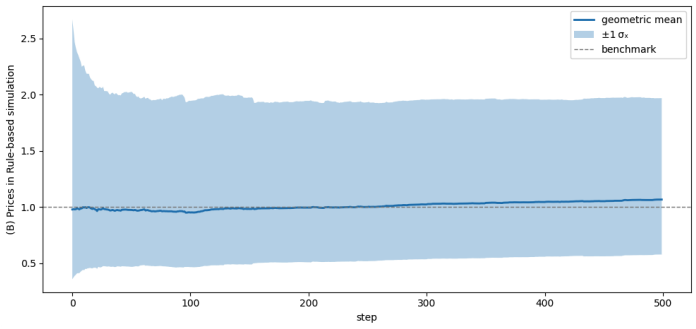
The advantage arises from adaptive foraging and trading: learning agents anticipate scarcity, alter trajectories accordingly, and engage in exchanges that prevent starvation. By contrast, rule-based agents follow rigid heuristics and often remain trapped in depleted areas. These adaptive dynamics yield higher aggregate survival without overexploiting the environment—an emergent increase in the system's effective carrying capacity.

4.2 Emergent Market Prices and Equilibrium

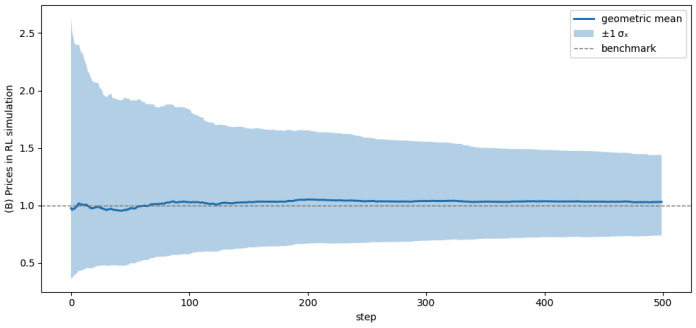
The next analysis focuses on Hypothesis H2, exploring how DRL-driven agents influence market equilibrium—specifically, whether adaptive learning accelerates price convergence and dampens volatility relative to heuristic traders. A defining feature of *Sugarscape* economies is the endogenous emergence of market prices from decentralized barter. To observe long-run equilibrium dynamics more realistically, we extend the baseline model by incorporating **fertility and finite agent lifespans** following

the demographic formulation of Epstein and Axtell [1]. This addition creates continuous population turnover, allowing equilibrium formation to be analyzed over extended horizons where generational learning and market renewal interact.

Figures 3 display the temporal evolution of the geometric mean and standard deviation of prices across 50 simulation runs for rule-based and DRL-driven societies, respectively. Shaded areas represent ± 1 standard deviation around the mean, indicating the degree of price dispersion within each system.



(a) Rule-based agents.



(b) RL agents.

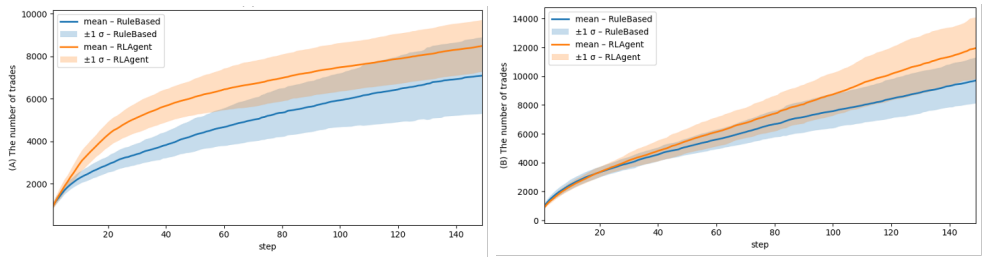
Fig. 3: Standard deviation of geometric mean prices over time in Rule-based and DRL economies under fertility and finite-lifespan conditions.

In both systems, prices initially fluctuate widely before gradually stabilizing near the equilibrium ratio ($p \approx 1$), consistent with Pareto-efficient trade. However, the **rule-based economy** exhibits persistent volatility throughout the 500-step horizon, reflecting slower correction of local price imbalances and delayed information diffusion. By contrast, the **DRL economy** demonstrates a sharper and sustained decline in variance: price dispersion falls by roughly half within the first 200 steps and continues to contract thereafter. This indicates that adaptive agents internalize equilibrium-seeking behavior through repeated trading experience, effectively learning to dampen arbitrage cycles.

The results suggest that DRL agents collectively function as distributed market stabilizers. Their learned policies balance localized resource demand and supply more efficiently, enabling a faster approach to the statistical equilibrium benchmark of Foley [16]. In contrast, the rule-based society remains trapped in cyclical over- and under-shooting typical of heuristic adjustment. By learning to anticipate future trade opportunities, DRL agents reduce market noise, accelerate price convergence, and produce a more resilient equilibrium trajectory across generations.

4.3 Trade Frequency and Volume

Further evidence pertaining to Hypothesis H2 is provided by investigating trading activity itself, evaluating how learned strategies alter the tempo and volume of exchanges within DRL economies. Figure 4 illustrates cumulative trade volume and tempo. DRL agents execute roughly 1.2 times as many trades as rule-based agents and maintain smoother trading intensity across time.



(a) Deterministic resource landscape. (b) Stochastic resource landscape.

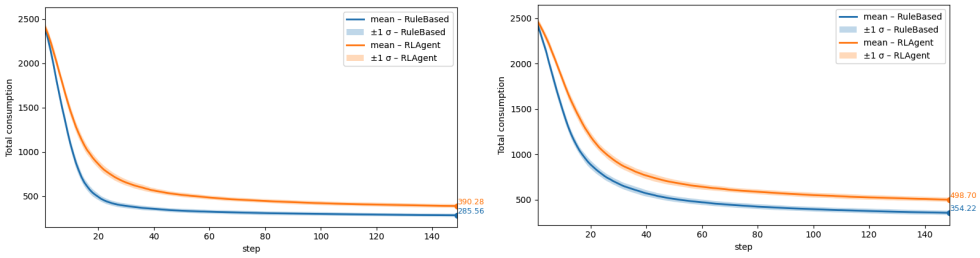
Fig. 4: Cumulative number of trades and trading tempo for DRL vs. rule-based societies.

Smoother trading tempo enhances market liquidity and contributes to faster convergence of prices toward equilibrium. In behavioral terms, the DRL population functions as a collective of "aggressive market-makers," continuously correcting mispricing, while heuristic agents act as slower, reactive traders.

4.4 Efficiency and Resource Utilization

In testing Hypothesis H3, the study compares the efficiency of resource exploitation and aggregate welfare generation between DRL and rule-based populations, determining whether adaptive agents achieve superior environmental utilization. Economic efficiency is assessed through total resource consumption, average wealth, and aggregate social wealth. Figure 5 plots total sugar and spice consumption over time in both environments.

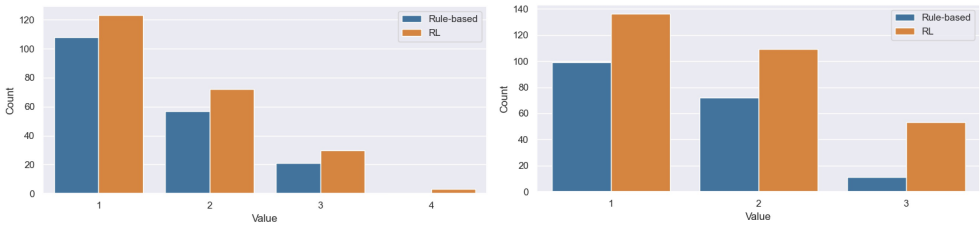
DRL societies consume 37% and 41% more resources than rule-based counterparts in deterministic and stochastic maps respectively, despite identical inflows, reflecting improved allocation rather than overharvesting. Average individual wealth rises



(a) Deterministic resource landscape. (b) Stochastic resource landscape.

Fig. 5: Total resource consumption over time under static conditions.

by about 7%, and total societal wealth increases proportionally with population size. The metabolic distribution (Figure 6) further reveals that DRL economies sustain a larger share of high-metabolism agents, indicating that trading redistributes resources toward those with greater needs—a phenomenon consistent with Cohen [17]'s ecological argument that trade expands effective carrying capacity.



(a) Deterministic resource landscape. (b) Stochastic resource landscape.

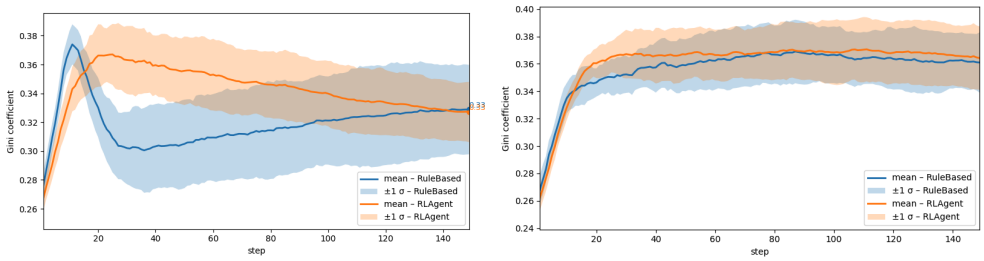
Fig. 6: Distribution of agents' metabolism levels in the final episodes.

The DRL agents' proactive harvesting and strategic exchanges minimize resource idleness. Unlike rule-based agents that cluster excessively on resource peaks, DRL populations distribute themselves efficiently across the landscape, ensuring near-complete utilization of the environment's regenerative capacity.

4.5 Inequality and Wealth Distribution

Hypothesis H4 addresses the dynamic evolution of wealth dispersion in DRL economies, testing the proposition that adaptive trading and redistribution mechanisms transform the initial surge in inequality into long-run equilibria exhibiting lower Gini indices and reduced distributional skewness.

Figure 7 tracks Gini dynamics for individual wealth. Early in simulation, DRL societies exhibit a transient surge in inequality as agents exploit resource peaks aggressively. Over time, however, extensive trading redistributes wealth, lowering inequality relative to the rule-based baseline. By episode end, both resources stabilize around

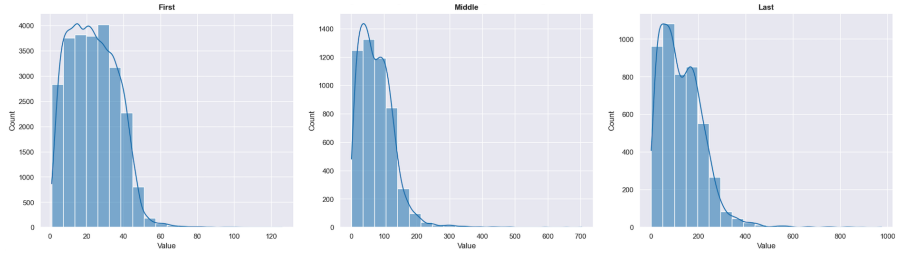


(a) Deterministic resource landscape. (b) Stochastic resource landscape.

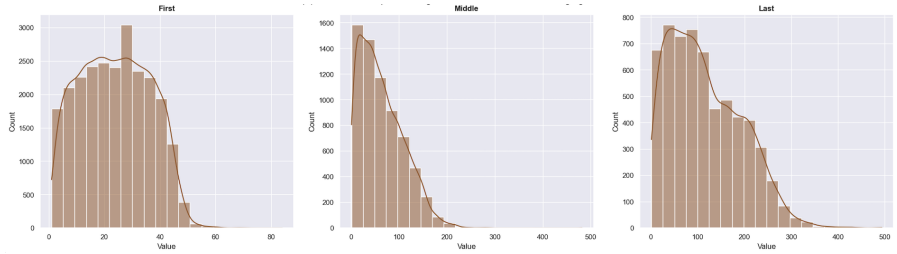
Fig. 7: Temporal evolution of Gini coefficients for individual wealth.

$G \approx 0.33$ in deterministic case, but DRL runs display narrower variance, indicating more consistent equilibrium outcomes.

The wealth histograms (Figure 8) confirm right-skewed, heavy-tailed distributions characteristic of kinetic exchange economies [18]. However, the DRL distribution shows a thinner upper tail, implying fewer extreme fortunes and fewer destitute agents—a collectively more equitable state.



(a) Rule-based agents.

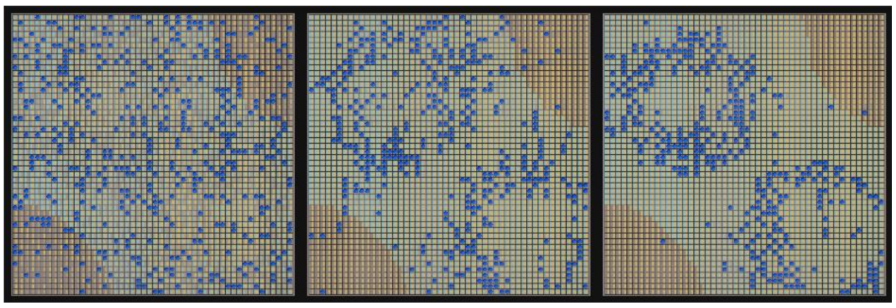


(b) RL agents.

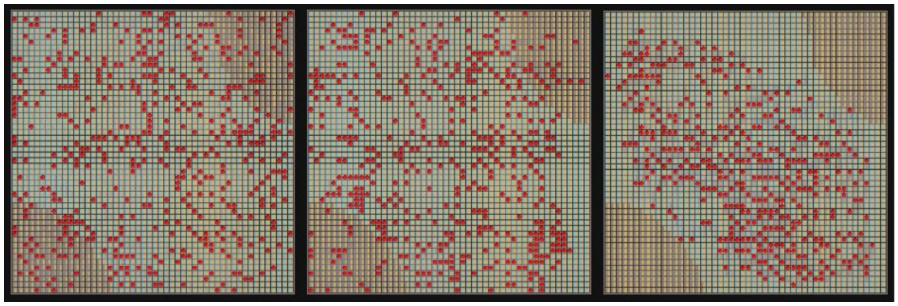
Fig. 8: Wealth distribution of DRL and rule-based agents at episode termination.

4.6 Dynamic Spatial and Sustainability Patterns

Finally, to synthesize the preceding hypotheses (H1–H3), this subsection analyzes the spatial and temporal patterns of resource use, demonstrating how adaptive learning fosters sustainable equilibrium across the landscape. To better isolate migration and trading behavior, the spatial map was slightly modified: the sugar resource was retained into localized peaks to encourage directional foraging, while the spice distribution was retained uniform across the grid. This adjustment simplifies the analysis by emphasizing agents' movement dynamics and interregional trade responses without altering aggregate resource inflows.



(a) Rule-based agents.



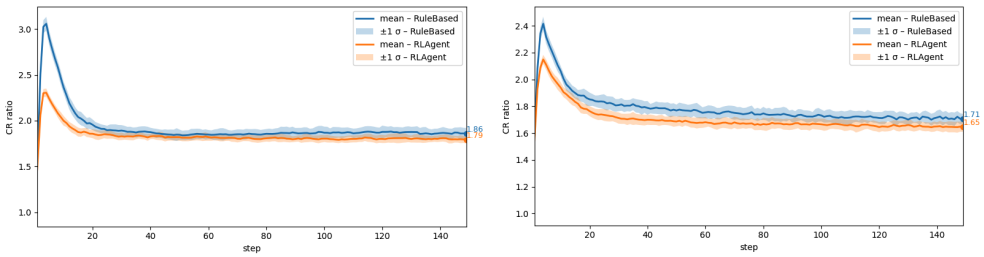
(b) RL agents.

Fig. 9: Spatial distribution of agents and emergent trade corridors in DRL society.

Beyond numerical metrics, spatial organization offers insight into emergent economic structure. Rule-based agents cluster densely around static resource peaks, leading to cyclical depletion and migration bursts. DRL agents, in contrast, maintain broader dispersion and form interlinked trade corridors between high-resource zones (Figure 9). These channels enable interregional exchange and stabilize the system against local shortages—an emergent analogue of trade networks in real economies.

Resource sustainability is measured by the consume-regrow ratio as shown in Figure 10. The DRL society maintains CR roughly 6–7% lower than the rule-based model (paired t-test $p < 10^{-8}$), signifying closer adherence to sustainable equilibrium.

Lower CR despite higher metabolism implies that DRL agents internalize intertemporal trade-offs—harvesting aggressively when surplus exists but relocating before local depletion. This adaptive restraint yields a self-organized balance between exploitation and renewal.



(a) Deterministic resource landscape. (b) Stochastic resource landscape.

Fig. 10: Consume-regrow ratio under static and stochastic environments.

Overall, the DRL-enhanced economy achieves superior outcomes in survival, efficiency, market stability, and sustainability. Adaptive policies amplify individual welfare while preserving environmental resilience, illustrating how micro-level learning can collectively move an economy toward near-optimal equilibrium.

5 Implications and Applications

5.1 Public Policy Modeling and Sustainability Planning

Integrating deep reinforcement learning into the *Sugarscape* framework transforms it into a flexible experimental laboratory for economic policy design. By allowing agents to autonomously learn resource-use and exchange strategies, the model enables systematic exploration of how alternative institutional environments—such as taxation, redistribution rules, or conservation incentives—shape long-run outcomes. Because learning dynamics unfold in a controlled, risk-free setting, policymakers can investigate equilibrium responses ex ante and assess trade-offs among efficiency, equity, and sustainability. This mirrors advances in multi-agent computational economics, such as the two-level *AI Economist* architecture Zheng et al. [4], where adaptive policy interacts with learning agents to achieve desired welfare goals. In a similar spirit, the DRL-enhanced Sugarscape can be extended to evaluate fiscal or regulatory instruments that respond dynamically to observed behavior, allowing policy to co-evolve with agents' strategies.

Beyond sustainability applications, the learning-based trading environment offers broader implications for understanding complex economic systems. Because agents internalize scarcity and exchange signals endogenously, the model naturally reveals feedback mechanisms and coordination failures that static frameworks may overlook. This creates fertile ground for examining how frictions in trade and resource access propagate through decentralized economies. For instance, selective constraints

on exchange or strategic withholding of resources—conceptual analogues to real-world trade barriers or export restrictions—could be introduced as policy experiments within the environment. Such scenarios would illuminate how disruptions to trading networks influence long-run welfare, distributional outcomes, and system resilience. Recent progress in multi-agent reinforcement learning for economic governance [19] suggests that these computational laboratories can help clarify when adaptive strategies mitigate or amplify shocks in interconnected socio-economic systems, offering insights relevant to contemporary global trade dynamics and coordinated policy responses.

5.2 Economic Simulation and Game Design

The DRL agents developed here also have practical applications in interactive simulations and games. Because they learn to balance production, consumption, and trade autonomously, such agents can populate virtual economies that evolve without manual scripting. In serious-game or educational contexts, they can illustrate complex economic phenomena—market formation, inequality, and sustainability—in real time. In entertainment games, DRL-driven non-player characters could act as rational traders or competitors, responding intelligently to scarcity and price signals. Prior work by McCarlie and Hunter [20] showed that even simple AI controllers can stabilize simulated economies; DRL agents extend this capacity to realistic, multi-commodity markets.

Embedding these learning economies into digital environments creates a two-way benefit. Developers gain richer, more believable worlds, while researchers obtain continuous behavioral data from large-scale simulations. Such cross-fertilization between economics and game AI points toward a new generation of hybrid research-and-entertainment platforms where adaptive economies evolve alongside human decision-makers.

6 Conclusion and Future Work

This study demonstrates that embedding trading-aware DRL agents within the classical *Sugarscape* framework allows micro-level adaptive behavior to produce macro-level equilibrium stability. By jointly optimizing movement and market exchange, learning agents internalize both spatial and trading incentives, resulting in higher survival rates, more equitable wealth distributions, and sustainable use of renewable resources. The findings highlight DRL not merely as a computational method but as a conceptual bridge that links behavioral adaptation, economic theory, and policy-oriented simulation.

The results also suggest that adaptive coordination and decentralized learning can replicate features of real economies—market stabilization, inequality mitigation, and sustainable resource management—emerging organically from boundedly rational behavior. This supports the growing view that agent-based models enriched with learning mechanisms can serve as powerful laboratories for studying complex economic dynamics and institutional design.

Looking ahead, future research will extend this framework toward a unified platform that integrates trading-aware DRL with **structural estimation** methods. Such integration would enable the recovery of fundamental behavioral primitives—preferences, expectations, and adjustment costs—from empirical data and embed them directly into the simulation. By calibrating DRL agents to real-world microeconomic evidence, the model could evolve from a theoretical laboratory to an empirically grounded decision environment. This unified framework would allow researchers to reproduce and test dynamic policy interventions, explore multi-commodity economies, and analyze co-evolving institutions within adaptive, data-driven simulations.

In summary, the combination of reinforcement learning, agent-based modeling, and structural estimation offers a promising pathway toward realistic, interpretable, and policy-relevant computational economics. Through this synthesis, future research can bring simulation-based methods closer to the empirical richness and predictive capacity required for understanding and sustaining complex economic systems.

Supplementary information. The supplementary materials include: (i) replication code (Python/Unity) and configuration files; (ii) pre-generated datasets used in the figures/tables; and (iii) a README with reproduction instructions. Files are provided on Zenodo (DOI: 10.5281/zenodo.17466369).

Acknowledgements. The author gratefully acknowledges the support of the University of Vaasa and the School of Accounting and Finance. Special thanks are extended to Professor Panu Kalmi for his supervision and insightful guidance, and to colleagues in the Economics research group for their valuable discussions and feedback on earlier drafts of this study.

Declarations

- Funding: This research received no external funding and was conducted as part of the author's doctoral research at the University of Vaasa, Finland.
- Competing interests: The author declares that there are no competing interests relevant to the content of this article.
- Ethics approval and consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Data availability: The datasets and code supporting the findings of this study are publicly available on Zenodo (DOI: 10.5281/zenodo.17466369). The pre-generated simulation datasets are located in the `TradingAwareSugarscapeReplication/` directory.
- Materials availability: All materials required to train the model are provided in the `TradingAwareSugarscapeReplicationForTraining/` directory within the supplementary Zenodo repository.
- Code availability: The replication package reproduces all figures and tables using Jupyter notebooks and Python scripts contained in the `TradingAwareSugarscapeReplication/` directory at the same Zenodo repository.

- Author contribution: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Visualization, Writing—Original Draft Preparation, and Writing—Review and Editing: Thanh Binh Lai.

References

[1] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. MIT Press, Cambridge, MA (1996)

[2] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017) https://doi.org/10.48550/arXiv.1707.06347

[3] Haarnoja, T., Zhou, A., Abbeel, P., Levine, S.: Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv preprint arXiv:1801.01290 (2018) https://doi.org/10.48550/arXiv.1801.01290

[4] Zheng, S., Trott, A., Srinivasa, S., Parkes, D.C., Socher, R.: The ai economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science Advances **8**(18), 2607 (2022) https://doi.org/10.1126/sciadv.abk2607

[5] Leibo, J.Z., Zambaldi, V., Lanctot, M., Marecki, J., Graepel, T.: Multi-agent reinforcement learning in sequential social dilemmas. Proceedings of the Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2017)

[6] Hughes, E., Leibo, J.Z., Phillips, M.G., Tuyls, K., Duéñez-Guzmán, E.A., García Castañeda, A., Dunning, I., Zhu, T., McKee, K.R., Koster, R., Roff, H., Graepel, T.: Inequity aversion improves cooperation in intertemporal social dilemmas. In: Advances in Neural Information Processing Systems 31 (NeurIPS 2018), Montréal, Canada, pp. 3330–3340 (2018). https://papers.neurips.cc/paper/7593-inequity-aversion-improves-cooperation-in-intertemporal-social-dilemmas.pdf

[7] Simon, H.A.: Models of Man: Social and Rational — Mathematical Essays on Rational Human Behavior in a Social Setting. Wiley, New York (1957). Reprinted by MIT Press, 1982

[8] Tesfatsion, L.: Modeling economic systems as locally-constructive sequential games. Journal of Economic Methodology **24**(4), 384–409 (2017) https://doi.org/10.1080/1350178X.2017.1382068

[9] Axtell, R.L.: The complexity of exchange. The Economic Journal **115**(504), 193–210 (2005)

[10] Brearcliffe, D., Crooks, A.: Creating intelligent agents: Combining agent-based modeling with machine learning. In: Proceedings of the 2020 Conference of the Computational Social Science Society of the Americas, pp. 31–58. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-83418-0_3

[11] Jäger, G.: Using neural networks for a universal framework for agent-based models. Mathematical and Computer Modelling of Dynamical Systems **27**(1), 162–178 (2021) https://doi.org/10.1080/13873954.2021.1889609

[12] Mesa Project Team: Sugarscape Constant Growback Model with Traders [Computer software example]. In: Mesa: Agent-Based Modeling in Python. Available at https://mesa.readthedocs.io/stable/examples/advanced/sugarscape_g1mt.html (2025)

[13] Ng, A.Y., Harada, D., Russell, S.J.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Bratko, I., Džeroski, S. (eds.) Proceedings of the Sixteenth International Conference on Machine Learning (ICML '99), pp. 278–287. Morgan Kaufmann, San Francisco, CA (1999). https://dl.acm.org/doi/10.5555/645528.657613

[14] Sert, E., Bar-Yam, Y., Morales, A.J.: Segregation dynamics with reinforcement learning and agent-based modeling. Scientific Reports **10**(1), 13535 (2020) https://doi.org/10.1038/s41598-020-68447-8

[15] Unity Technologies: Unity Machine Learning Agents Toolkit (ML-Agents). Available at https://github.com/Unity-Technologies/ml-agents (2025)

[16] Foley, D.K.: A statistical equilibrium theory of markets. Journal of Economic Theory **62**(2), 321–345 (1994) https://doi.org/10.1006/jeth.1994.1018

[17] Cohen, J.E.: Population growth and earth's human carrying capacity. Science **269**(5222), 341–346 (1995) https://doi.org/10.1126/science.7618100

[18] Boghosian, B.M.: Kinetics of wealth and the pareto law. Physical Review E **89**(4) (2014) https://doi.org/10.1103/PhysRevE.89.042804

[19] Rudd-Jones, J., Musolesi, M., Pérez-Ortiz, M.: Multi-agent reinforcement learning simulation for environmental policy synthesis. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), pp. 2890–2895 (2025). https://dl.acm.org/doi/10.5555/3709347.3744041

[20] McCarlie, P., Hunter, A.: Using game ai to control a simulated economy. In: Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021) – Volume 2, pp. 629–634. SCITEPRESS – Science and Technology Publications, Vienna, Austria (2021). https://doi.org/10.5220/0010212306290634 . https://www.scitepress.org/Papers/2021/102123/102123.pdf