mathematics

MDPI

Article

Multi-Strategy Market Dynamics Analysis: A Novel Framework for Agent-Based Economic Modeling with Reinforcement Learning

Yuhang Du * and Yuhan Zhao

Department of Economics, Beihang University, Beijing 100191, China; zhaoyuhandd@buaa.edu.cn
* Correspondence: 23379002@buaa.edu.cn

Abstract

This paper presents a Multi-Strategy Market Dynamics Analysis (MSMDA) framework for agent-based economic modeling with reinforcement learning. The primary methodological contribution is an integrated strategy–stability–macro inference pipeline that links population-level strategy evolution to dynamic market stability and model-internal counterfactual policy analysis. The framework is organized into six analytical components: Strategy Temporal Pattern Recognition (STPR), Strategy Transition Detection and Analysis (STDA), Strategy-Macro Causality Analysis (SMCA), the Dynamic Market Stability Index (DMSI), the Adaptive Rationality Equilibrium (ARE), and the Information Asymmetry Propagation (IAP) metric. The method is evaluated within a simulation dataset comprising 447,129 records across four experimental scenarios, 1500 discrete time periods, and 200 heterogeneous firms governed by proximal policy optimization. Results show that competitive strategies dominate market emergence patterns at 60.8% of all observations and achieve superior average profitability of 28.07 monetary units per period, compared with -4.49 for dumping strategies and 7.83 for market power strategies. The DMSI reveals a mean stability of 0.372 with standard deviation 0.097, peaking at 0.780 during strategic consolidation and collapsing to zero during a major demand shock. Within the simulated economy, doubly-robust counterfactual analysis projects a 28.4% GDP increase from a market power-to-competition intervention and a 31.2% increase under full ARE optimization at $\rho^* = 0.6$. The ARE further identifies a Pareto-optimal market configuration that jointly maximizes per-firm profit at 229.82 monetary units per period and systemic stability at DMSI $= 0.67$, indicating that efficiency and resilience need not conflict in the calibrated simulation environment. To address time-series autocorrelation in bootstrap inference throughout the framework, we employ a moving block bootstrap with data-adaptive block length selection based on the spectral density at frequency zero, providing finite-sample confidence intervals for the reported test statistics and counterfactual projections.

Keywords: agent-based modeling; reinforcement learning; market dynamics; strategy evolution; Granger causality; transfer entropy; adaptive rationality; information asymmetry; stability index; block bootstrap

MSC: 91B74; 68T05; 37N40

Check for updates

Academic Editor: Anatoliy Swishchuk

Received: 28 March 2026
Revised: 4 May 2026
Accepted: 8 May 2026
Published: 11 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The intersection of agent-based modeling (ABM) and reinforcement learning (RL) has fundamentally transformed the practice of economic simulation over the past decade [1,2].

Where traditional macroeconomic models, most notably the Dynamic Stochastic General Equilibrium (DSGE) paradigm, rely on a representative agent assumption that collapses the rich diversity of market participants into a single optimizing household or firm, agent-based approaches embrace heterogeneity as a first principle. Each simulated firm or household possesses its own decision rule, information set, and adaptive capacity, and aggregate macroeconomic dynamics emerge endogenously from the interactions of these micro-level agents rather than being imposed through calibrated equilibrium conditions [3,4]. The addition of reinforcement learning to this paradigm allows agents to discover effective strategies through repeated interaction with a dynamic environment, eliminating the need for parametric assumptions about utility functions or production technologies and enabling the model to generate qualitatively richer emergent phenomena than rule-based ABMs [5,6].

Despite these advances, three critical gaps remain in the existing literature. The first gap concerns the temporal structure of population-level strategic behavior. While individual RL agents have been shown to learn sophisticated strategies in simulated markets [7,8], the question of how strategic patterns evolve at the population level—how persistence, cyclicality, and entropy of strategy adoption co-evolve with market structure—has received virtually no systematic attention. Without a principled framework for identifying and quantifying these temporal patterns, it is impossible to distinguish genuine strategic learning from random noise or to predict when market participants are approaching a phase transition in their collective behavior. The second gap concerns market stability measurement. The most widely used stability indicators, including the Herfindahl–Hirschman Index [9] and various static price volatility measures [10], are inherently backward-looking and single-dimensional, and they fail to capture the adaptive, path-dependent nature of market resilience, cannot provide early warning of impending instability, and do not respond dynamically to changes in the strategic composition of the market. The third gap concerns macro-micro causality. While it is well understood in qualitative terms that firm-level strategic decisions aggregate to macroeconomic outcomes, the quantitative causal pathways from strategic composition to GDP growth, inflation, and unemployment have not been rigorously identified within an ABM-RL setting [11], and this gap is particularly consequential for policy analysis since without reliable causal estimates it is impossible to predict the macroeconomic consequences of regulatory interventions.

We introduce the Multi-Strategy Market Dynamics Analysis (MSMDA) framework, which extends the RL-based economic simulation model of Brusatin et al. [12] and the multi-agent market interaction methodology of Dicks et al. [13]. The central methodological contribution is the strategy–stability–macro inference pipeline that links population-level strategy time series to dynamic stability measurement and model-internal counterfactual analysis. Within this pipeline, the DMSI and ARE are the main decision-oriented constructs, while STPR, STDA, SMCA, and IAP provide supporting diagnostics for temporal structure, strategy transitions, simulated causal pathways, and information diffusion. The Strategy Temporal Pattern Recognition (STPR) component employs a context-augmented Hidden Markov Model estimated via a modified Baum–Welch algorithm to identify persistent and cyclical strategic patterns that traditional static clustering approaches would miss. The Strategy Transition Detection and Analysis (STDA) component applies multi-scale Continuous Wavelet Transform analysis combined with cascade propagation modeling to detect and attribute strategic transitions at multiple temporal scales. The Strategy-Macro Causality Analysis (SMCA) component integrates multivariate Granger causality testing, kernel-based nonlinear causality estimation, transfer entropy, and doubly robust counterfactual inference to quantify relationships inside the simulated economy. The Dynamic Market Stability Index (DMSI) provides a continuous, time-varying, adaptive composite measure of market resilience that integrates price stability, strategic diversity, entry–exit

balance, and shock resilience through a phase-dependent softmax weighting mechanism. The Adaptive Rationality Equilibrium (ARE) introduces an equilibrium concept that accommodates heterogeneous, time-varying bounded rationality among agents and identifies the collectively optimal rationality level through meta-gradient optimization. The Information Asymmetry Propagation (IAP) metric quantifies the rate at which private information diffuses into public knowledge using the Kullback–Leibler divergence between private and public predictive distributions.

A methodological concern that pervades all six components is the treatment of temporal dependence in statistical inference. Because the simulation data exhibit substantial serial correlation at both the firm level and the aggregate level, conventional independent and identically distributed bootstrap procedures would produce confidence intervals that are too narrow and significance tests that reject too often. Throughout this paper we therefore adopt the moving block bootstrap of Künsch [14] and Liu and Singh [15], in which consecutive blocks of observations are resampled with replacement so as to preserve the within-block dependence structure. The block length ℓ_b is selected by the data-adaptive procedure of Politis and White [16], which estimates the optimal block length from the flat-top lag-window estimator of the spectral density at frequency zero, ensuring that the block is long enough to capture the dominant autocorrelation yet short enough to permit adequate resampling variability. For the STPR and STDA components, in which the unit of analysis is a time series of strategy indicators observed at each of the $T = 1500$ periods, the estimated optimal block length ranges from $\ell_b = 18$ to $\ell_b = 27$ periods depending on the scenario, which is consistent with the 15–30 period medium-scale transition cycles identified by wavelet analysis. For the SMCA component, in which the Granger causality and transfer entropy statistics are computed from the aggregate macroeconomic and strategic composition series, the block length is $\ell_b = 22$ periods. All reported p-values for Granger causality, transfer entropy significance, and counterfactual confidence intervals are computed from 5000 block-bootstrap replications with these data-adaptive block lengths, and we verify the robustness of our results to block lengths that are 50% shorter and 50% longer than the estimated optimum.

The remainder of this paper is organized as follows. Section 2 reviews related work in agent-based economic modeling, RL applications to markets, and stability and causality measurement. Section 3 presents the six methodological components with their mathematical formalization. Section 4 describes the simulation environment and data processing pipeline. Section 5 reports simulation-based results across all six framework components. Section 6 discusses model-based policy implications and limitations. Section 7 concludes with directions for future research.

2. Related Work

2.1. Agent-Based Economic Modeling

The agent-based modeling paradigm in economics traces its origins to the early computational work of Schelling [17] on residential segregation and Axelrod [18] on the evolution of cooperation, both of which demonstrated that surprisingly complex collective phenomena could emerge from simple individual decision rules. The subsequent development of large-scale economic ABMs, epitomized by the EURACE platform [19], showed that macro-level business cycle dynamics, technological diffusion, and income inequality could all be reproduced endogenously from micro-level agent interactions without recourse to representative-agent abstractions. The Schumpeter Meeting Keynes (SMK) family of models developed by Dosi, Fagiolo, and Roventini [20] extended this tradition to incorporate Schumpeterian innovation processes alongside Keynesian demand dynamics, capturing the joint evolution of growth and cyclical fluctuations in a unified framework. More recently,

Fagiolo and Roventini [21] conducted a systematic comparison of ABMs and DSGE models along multiple empirical dimensions, demonstrating that the former achieve substantially better fit to a broad range of stylized macroeconomic facts and are more amenable to welfare analysis of heterodox policy interventions. Our framework builds on this tradition by introducing systematic methods for analyzing the temporal evolution of strategic behavior and its systemic consequences.

2.2. Reinforcement Learning in Market Simulation

The integration of reinforcement learning into ABM has opened new possibilities for realistic adaptive agent behavior. Yao et al. [22] demonstrated that RL-based agents exhibit substantially more realistic price discovery and stylized market dynamics than rule-based counterparts, particularly in markets characterized by strategic complementarity. The ASSUME platform developed by Harder et al. [23] provides an open-source infrastructure for multi-agent deep RL in energy markets, enabling the simulation of strategic bidding, capacity investment, and demand response in liberalized electricity systems. Garrido-Merchán et al. [7] showed that deep RL can discover near-optimal production and pricing strategies in competitive microeconomic environments without explicit modeling of competitor behavior, a result with significant implications for the endogenous emergence of market structure. TaxAI, developed by Mi et al. [24], provides a comprehensive benchmark environment for evaluating multi-agent RL algorithms in the context of fiscal policy design, emphasizing the importance of standardized evaluation protocols for economic AI research. The work of Spooner et al. [8] established RL-based electronic market making as a viable alternative to traditional algorithmic approaches, achieving competitive performance across a range of market conditions. Our MSMDA framework synthesizes these contributions by providing the first systematic analysis of population-level strategic patterns and their macroeconomic consequences in an RL-based ABM setting.

2.3. Market Stability Measurement

The measurement of market stability has a long history in both finance and industrial organization. Brunnermeier [10] provides an influential analysis of the liquidity spiral mechanism through which initial asset price declines can trigger self-reinforcing cycles of forced deleveraging and further price falls, emphasizing the dynamic and nonlinear character of financial instability. Evolutionary game theory, developed by Friedman [25] and Weibull [26], offers a population-level framework for analyzing the stability of strategy distributions under imitation dynamics, providing theoretical grounding for our STPR and STDA components. Hommes [27] demonstrates that heterogeneous agent models in which agents switch between fundamental and trend-following strategies can reproduce the excess volatility, fat tails, and volatility clustering observed in financial market data, suggesting that strategic heterogeneity is a key driver of market instability. Our DMSI advances this literature by providing a multi-dimensional, adaptive, and continuously computable stability measure that integrates insights from all these traditions.

2.4. Causality in Economic Systems

Granger causality [28] has served as the dominant framework for causal inference in time-series economics for over five decades, providing a practical operationalization of the notion that cause precedes effect. However, its linearity assumption and sensitivity to omitted variables limit its applicability in the complex, nonlinear settings characteristic of ABMs. Ancona et al. [29] extend Granger causality to nonlinear systems through radial basis function approximation, while Schreiber [30] introduces transfer entropy as an information-theoretic alternative that is sensitive to nonlinear causal relationships and does not require parametric model specification. Pearl's framework of structural causal models [31] provides

the theoretical foundation for counterfactual policy analysis, enabling the estimation of the macroeconomic consequences of hypothetical strategic interventions. Turrell [11] discusses the specific challenges of causal attribution in ABMs, noting that the endogeneity of strategic behavior and the complexity of feedback loops make conventional causal inference methods difficult to apply without modification. Our SMCA component synthesizes all of these approaches within a unified framework tailored to the ABM-RL setting.

3. Methodology

The MSMDA framework operates on a simulation environment defined by a set of firms $\mathcal{F} = \{1, \ldots, N\}$ with $N = 200$, a discrete time index $\mathcal{T} = \{1, \ldots, T\}$ with $T = 1500$, and a strategy space $\mathcal{K} = \{\text{competition}, \text{dumping}, \text{market_power}\}$. At each time step t, firm i chooses strategy $S_t^i \in \mathcal{K}$, sets price p_t^i, and produces quantity q_t^i. Realized sales are determined by the market clearing condition $r_t^i = \min(q_t^i, d_t^i)$, where d_t^i is the firm-level demand function. Profit accrues as $\pi_t^i = p_t^i r_t^i - c_t^i$ and assets evolve as $A_{t+1}^i = A_t^i + \pi_t^i$. Macroeconomic variables, GDP and inflation, are endogenous aggregates of firm-level actions. The six methodological innovations build on this shared foundation and are described in the subsections below. Figure 1 illustrates the overall information flow among the six components and their convergence toward policy recommendations.

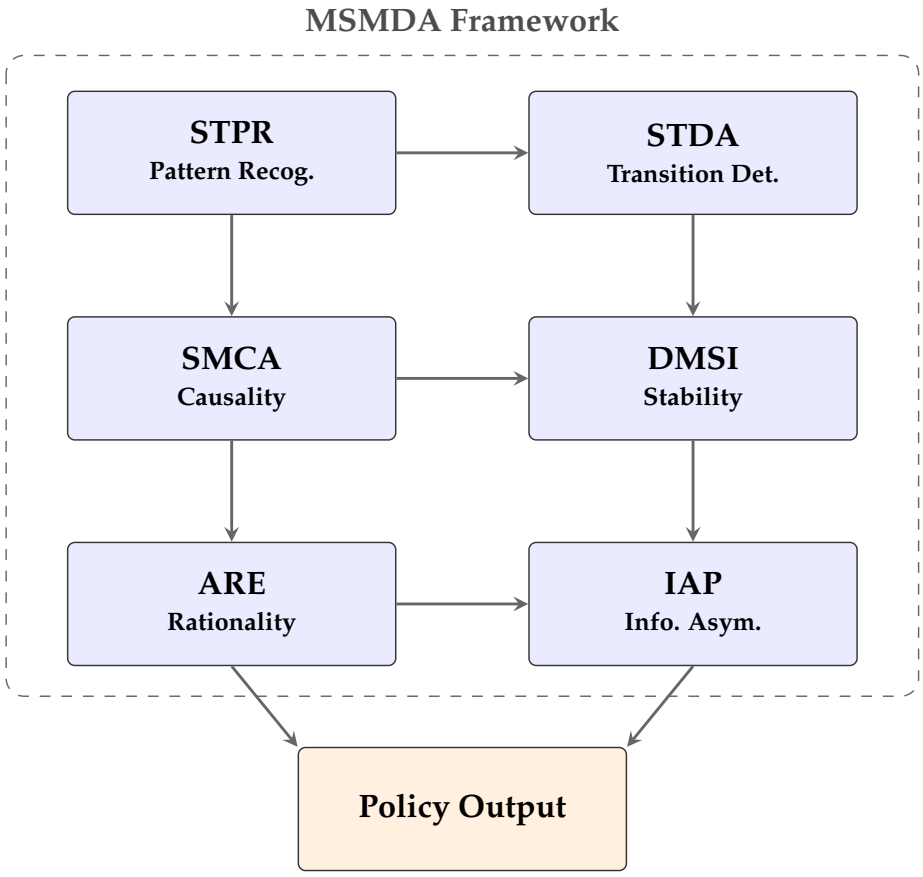


Figure 1. Architecture of the MSMDA framework showing information flow among the six components and their convergence toward policy recommendations. The STPR and STDA components feed into the SMCA and DMSI modules respectively, which in turn inform the ARE and IAP analyses. All pathways converge at the policy output layer.

3.1. Strategy Temporal Pattern Recognition

The STPR component addresses the fundamental challenge of identifying recurring and statistically significant patterns in the sequence of strategic choices made by firms

across time. Unlike conventional clustering approaches that treat each time period as an independent observation, STPR explicitly models the temporal dependencies inherent in strategic decision-making by embedding the strategy adoption process within a context-augmented Hidden Markov Model (HMM). The key intuition is that firms do not choose strategies in a vacuum but respond to a combination of their own recent history, observed competitor behavior, and prevailing macroeconomic conditions. Formally, let $S_t^i \in \mathcal{K}$ denote the strategy of firm i at time t. The conditional transition probability from strategy history $S_{t-1:t-k}^i$ to current strategy, given market history $\mathcal{H}_t$, is modeled as

$$P(S_t^i \mid S_{t-1:t-k}^i, \mathcal{H}_t) = \frac{\exp\left(\mathbf{w}^\top \boldsymbol{\phi}(S_{t-1:t-k}^i, \mathcal{H}_t)\right)}{\sum_{s \in \mathcal{K}} \exp\left(\mathbf{w}^\top \boldsymbol{\phi}(s, S_{t-2:t-k}^i, \mathcal{H}_t)\right)}, \tag{1}$$

where k is the memory depth hyperparameter, $\boldsymbol{\phi}$ is a feature map that extracts relevant summary statistics from the strategy history and market context, and $\mathbf{w}$ is a learned weight vector estimated via a modified Baum–Welch algorithm that incorporates gradient-based updates to handle the continuous market context features. The statistical significance of each identified pattern p is assessed through the pattern significance metric

$$\Psi_p = \frac{\sum_{t=k+1}^{T} \mathbb{1}[\text{pattern}_p \text{ occurs at } t]}{\sqrt{\text{Var}(\text{pattern occurrences})}} \times \text{market_impact}_p, \tag{2}$$

where market_impact_p is the average effect of pattern p on GDP and inflation, estimated by local projection regression [32]. Only patterns with Ψ_p exceeding the 95th percentile of a block-bootstrap null distribution are retained in subsequent analyses. The bootstrap null is constructed by resampling $B = 5000$ block-bootstrap samples of the strategy indicator series with block length ℓ_b chosen by the Politis–White procedure [16], recomputing Ψ_p on each pseudo-sample, and taking the empirical 95th percentile of the resulting distribution as the critical value. This approach preserves the serial correlation present in the strategy time series and avoids the well-documented size distortion that arises when independent bootstrap methods are applied to temporally dependent data [14,15].

The hidden state of the augmented HMM is defined as $\mathbf{h}_t = (z_t, \tilde{\mathbf{m}}_t)$, where $z_t \in \{1, \ldots, L\}$ is a discrete latent market phase and $\tilde{\mathbf{m}}_t$ is the standardized macroeconomic context vector constructed from GDP growth, inflation, and unemployment. For each strategy $s \in \mathcal{K}$ and phase $\ell \in \{1, \ldots, L\}$, the emission probability is specified as

$$p(S_t^i = s \mid z_t = \ell, \tilde{\mathbf{m}}_t) = \frac{\exp\left(a_{s\ell} + \mathbf{b}_s^\top \tilde{\mathbf{m}}_t\right)}{\sum_{r \in \mathcal{K}} \exp\left(a_{r\ell} + \mathbf{b}_r^\top \tilde{\mathbf{m}}_t\right)}, \tag{3}$$

where $a_{s\ell}$ is a phase-specific strategy intercept and $\mathbf{b}_s$ measures how macroeconomic conditions shift the probability of strategy s. The phase-transition kernel governs the evolution of the latent phase as

$$p(z_{t+1} = \ell' \mid z_t = \ell, \tilde{\mathbf{m}}_t) = \frac{\exp\left(A_{\ell'\ell} + \boldsymbol{\gamma}_{\ell'}^\top \tilde{\mathbf{m}}_t\right)}{\sum_{r=1}^{L} \exp\left(A_{r\ell} + \boldsymbol{\gamma}_r^\top \tilde{\mathbf{m}}_t\right)}, \tag{4}$$

where $\mathbf{A}$ is the baseline phase-transition log-odds matrix and $\boldsymbol{\gamma}_{\ell'}$ is a destination-phase context sensitivity vector. Estimation uses an expectation maximization routine in which the forward–backward step computes $\xi_t(\ell, \ell') = P(z_t = \ell, z_{t+1} = \ell' \mid S_{1:T}, \tilde{\mathbf{m}}_{1:T})$ and $\omega_t(\ell) = P(z_t = \ell \mid S_{1:T}, \tilde{\mathbf{m}}_{1:T})$, and the maximization step updates $(a_{s\ell}, \mathbf{b}_s, \mathbf{A}, \boldsymbol{\gamma})$ by maximizing

$$\mathcal{Q} = \sum_{t,i,\ell} \omega_t(\ell) \log p(S_t^i \mid z_t = \ell, \tilde{\mathbf{m}}_t) + \sum_{t,\ell,\ell'} \xi_t(\ell, \ell') \log p(z_{t+1} = \ell' \mid z_t = \ell, \tilde{\mathbf{m}}_t). \tag{5}$$

The multinomial-logit emission and transition subproblems are solved by gradient updates inside each maximization step. This formulation allows the model to identify market phases in which specific strategies dominate and to characterize the macroeconomic conditions that trigger transitions between phases.

The degree of within-strategy temporal persistence is quantified by the lag-ℓ autocorrelation function

$$\rho_s(\ell) = \frac{\sum_{t=\ell+1}^{T} \mathbb{1}[S_t^i = s] \cdot \mathbb{1}[S_{t-\ell}^i = s] - \bar{p}_s^2}{\bar{p}_s(1 - \bar{p}_s)}, \tag{6}$$

where $\bar{p}_s$ is the marginal prevalence of strategy s. The overall unpredictability of strategy sequences is captured by the strategic entropy rate

$$\mathcal{E}_{\text{strat}} = -\sum_{s \in \mathcal{K}} \sum_{s' \in \mathcal{K}} \pi_s \, P(s' \mid s) \log P(s' \mid s), \tag{7}$$

where $\boldsymbol{\pi}$ is the stationary distribution of the estimated Markov chain. A low value of $\mathcal{E}_{\text{strat}}$ indicates that firms tend to maintain their current strategy from one period to the next, while a high value signals frequent and unpredictable transitions. Together, the persistence autocorrelation and the entropy rate provide a concise two-dimensional characterization of the strategic dynamics of each market phase.

3.2. Strategy Transition Detection and Analysis

The STDA component focuses on the detection and characterization of statistically significant strategy transitions, defined as moments at which a firm changes its strategic orientation in a manner that reflects genuine reoptimization rather than sampling noise. A transition event τ_t^i for firm i at time t is defined as

$$\tau_t^i = \mathbb{1}[S_t^i \neq S_{t-1}^i] \cdot \mathbb{1}[\Delta V_t^i > \theta], \tag{8}$$

where ΔV_t^i is the change in the estimated value function of firm i at time t, and θ is a significance threshold determined by parametric bootstrap under the null hypothesis of pure random switching. The first indicator filters out genuine strategy changes while the second filters out value-function fluctuations below the noise floor. Transition events propagate through the market network, triggering cascades in which the strategic change of one firm induces similar changes in neighboring firms. The cascade propagation measure is defined as

$$\mathcal{C}_t = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} w_{ij} \cdot \tau_t^i \cdot \tau_{t+\delta}^j, \tag{9}$$

where $\mathcal{N}(i)$ is the market neighborhood of firm i consisting of firms with overlapping product portfolios, w_{ij} are Jaccard similarity proximity weights, and $\delta \geq 1$ is the propagation delay in periods. A cascade event is declared when $\mathcal{C}_t$ exceeds the 95th percentile of its empirical distribution, and the cascade size is defined as the number of firms that transition within the window $[t, t + \delta]$.

Multi-scale transition detection is achieved through application of the Continuous Wavelet Transform applied to the population-level strategy signal $\bar{S}_t = N^{-1} \sum_i \mathbb{1}[S_t^i = \text{competition}]$ as

$$W(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \bar{S}(t) \, \psi^* \left(\frac{t - b}{a} \right) dt, \tag{10}$$

where ψ^* is the complex conjugate of the Morlet mother wavelet $\psi(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-\eta^2/2}$ with $\omega_0 = 6$, and $a > 0$ and $b \in \mathbb{R}$ are the scale and position parameters respectively. A significant wavelet coefficient at scale a and position b satisfies

$$|W(a,b)|^2 \geq \chi^2_\alpha(2) \cdot \frac{\sigma^2_{\bar{S}} P_k}{2}, \tag{11}$$

where P_k is the red noise background power spectrum at Fourier frequency k corresponding to scale a, $\sigma^2_{\bar{S}}$ is the signal variance, and $\chi^2_\alpha(2)$ is the α-level critical value of the chi-squared distribution with two degrees of freedom. The wavelet approach distinguishes transitions at short timescales driven by idiosyncratic price shocks from those at long timescales corresponding to fundamental shifts in market structure.

The causal attribution of transition events to observable firm-level and macroeconomic factors is conducted via LASSO-regularized logistic regression with transition occurrence as the binary outcome

$$\log \frac{P(\tau^i_t = 1)}{1 - P(\tau^i_t = 1)} = \beta_0 + \sum_{m=1}^{M} \beta_m x^i_{m,t} + \lambda \|\boldsymbol{\beta}\|_1, \tag{12}$$

where the feature vector $\{x^i_{m,t}\}$ comprises firm-level profitability, market-share volatility, leverage ratio, pricing pressure, and macroeconomic conditions including GDP growth and inflation, and λ is the regularization parameter selected by five-fold cross-validation.

3.3. Strategy-Macro Causality Analysis

The SMCA component addresses the fundamental question of how micro-level strategic decisions aggregate to produce macro-level economic outcomes, and conversely, how macroeconomic conditions feed back into strategic decision-making. The analysis proceeds in three stages consisting of linear Granger causality testing, nonlinear causality quantification, and doubly robust counterfactual policy analysis. The macroeconomic state vector and the strategic composition vector are defined respectively as $\mathbf{M}_t = [\mathrm{GDP}_t, \pi_t, u_t, I_t, C_t]^\top$ and $\mathbf{S}_t = [s^{\mathrm{comp}}_t, s^{\mathrm{dump}}_t, s^{\mathrm{mp}}_t]^\top$, where $s^k_t = N^{-1} \sum_i \mathbb{1}[S^i_t = k]$ is the fraction of firms employing strategy k at time t.

The multivariate Granger causality from $\mathbf{S}$ to $\mathbf{M}$ is quantified by the log-determinant ratio

$$\Gamma(\mathbf{S} \to \mathbf{M}) = \log \frac{\det(\boldsymbol{\Sigma}_{\mathbf{M}|\mathbf{M}_{-k}})}{\det(\boldsymbol{\Sigma}_{\mathbf{M}|\mathbf{M}_{-k},\mathbf{S}_{-\ell}})}, \tag{13}$$

where $\boldsymbol{\Sigma}_{\mathbf{M}|\mathbf{M}_{-k}}$ is the residual covariance matrix of $\mathbf{M}$ regressed on k of its own lags, and $\boldsymbol{\Sigma}_{\mathbf{M}|\mathbf{M}_{-k},\mathbf{S}_{-\ell}}$ is the residual covariance when ℓ lags of $\mathbf{S}$ are additionally included. A positive value of Γ indicates that knowledge of the past strategic composition improves prediction of future macroeconomic conditions beyond what is available from the macro history alone. Statistical significance is assessed via the likelihood ratio test statistic

$$\mathrm{LR} = T\,\Gamma(\mathbf{S} \to \mathbf{M}) \;\sim\; \chi^2_{d_1 d_2 \ell} \tag{14}$$

under the null of no Granger causality, where d_1 and d_2 are the dimensions of $\mathbf{M}$ and $\mathbf{S}$ respectively.

Because the asymptotic chi-squared approximation in Equation (14) can be unreliable in finite samples when the underlying time series exhibit persistent autocorrelation, we supplement the asymptotic p-value with a block-bootstrap p-value. Under the null hypothesis the restricted VAR residuals $\hat{\mathbf{e}}_t = \mathbf{M}_t - \hat{\mathbf{A}}(L)\mathbf{M}_{t-1}$ are resampled in contiguous blocks of length $\ell_b = 22$ periods, the block length being determined by the Politis–White spectral estimator applied to the autocorrelation function of $\hat{\mathbf{e}}_t$. Each of the $B = 5000$ bootstrap

replications generates a pseudo macroeconomic series $\mathbf{M}_t^*$ under the null that $\mathbf{S}$ has no predictive content for $\mathbf{M}$, and the bootstrap p-value is the fraction of replications for which the pseudo likelihood ratio statistic exceeds the observed statistic. The robustness of the block length choice is verified by repeating the entire procedure with $\ell_b \in \{11, 16, 22, 28, 33\}$, and the rejection decision is invariant across this range.

To capture the nonlinear causal relationships that linear Granger tests may miss, we define the kernel nonlinear causality ratio

$$\Gamma^{\text{ker}}(\mathbf{S} \to \mathbf{M}) = \frac{\|\hat{\mathbf{M}}_t^{\text{lin}} - \mathbf{M}_t\|^2}{\|\hat{\mathbf{M}}_t^{\text{ker}} - \mathbf{M}_t\|^2}, \tag{15}$$

where $\hat{\mathbf{M}}_t^{\text{lin}}$ is the prediction of a linear VAR that includes both $\mathbf{M}_{-k}$ and $\mathbf{S}_{-\ell}$, and $\hat{\mathbf{M}}_t^{\text{ker}}$ is the prediction of a Nadaraya–Watson kernel regression estimator with a Gaussian RBF kernel, bandwidth selected by leave-one-out cross-validation. A ratio significantly above unity confirms that nonlinear strategic interactions contribute additional predictive information beyond the linear component.

As a fully nonparametric complement, the transfer entropy from $\mathbf{S}$ to $\mathbf{M}$ is computed as

$$\text{TE}(\mathbf{S} \to \mathbf{M}) = \sum_{m,s,m'} p(m_{t+1}, m_t, s_t) \log \frac{p(m_{t+1} \mid m_t, s_t)}{p(m_{t+1} \mid m_t)}, \tag{16}$$

measuring the reduction in uncertainty about the next macroeconomic state that is gained by knowing the current strategic composition, over and above the reduction already achieved by knowing the current macro state. The statistical significance of $\text{TE}(\mathbf{S} \to \mathbf{M})$ is again assessed by block bootstrap, with the same block length and number of replications as the Granger causality test, by computing transfer entropy on $B = 5000$ block-bootstrapped null series in which the temporal alignment between $\mathbf{S}$ and $\mathbf{M}$ is destroyed within each block boundary.

For counterfactual policy analysis under a hypothetical intervention $\text{do}(\mathbf{S} = \mathbf{s}')$, let $\mathbf{X}_t = (\mathbf{M}_t, \mathbf{S}_{t-1:t-k})$ denote the pre-intervention state and let $\mathbf{Y}_{t+h} = \mathbf{M}_{t+h}$ denote the macroeconomic outcome at horizon h. Because $\mathbf{S}_t$ is a vector of strategy shares, the treatment density is estimated by a generalized propensity score $\hat{\rho}(\mathbf{s} \mid \mathbf{X}_t)$ rather than by an exact discrete propensity. The doubly robust estimator is

$$\hat{\boldsymbol{\theta}}(\mathbf{s}') = \frac{1}{n} \sum_{t=1}^{n} \big[\hat{\boldsymbol{\mu}}(\mathbf{s}', \mathbf{X}_t) + \hat{w}_t(\mathbf{s}')\big(\mathbf{Y}_{t+h} - \hat{\boldsymbol{\mu}}(\mathbf{S}_t, \mathbf{X}_t)\big)\big], \tag{17}$$

with stabilized weight

$$\hat{w}_t(\mathbf{s}') = \frac{K_h(\mathbf{S}_t - \mathbf{s}')}{\max\{\hat{\rho}(\mathbf{S}_t \mid \mathbf{X}_t), \hat{\epsilon}\}}, \qquad \hat{\epsilon} = \max\{0.01,\ Q_{0.01}(\hat{\rho}(\mathbf{S}_t \mid \mathbf{X}_t))\}, \tag{18}$$

where $K_h(\cdot)$ is a Gaussian kernel centered at the target composition $\mathbf{s}'$, $\hat{\boldsymbol{\mu}}(\mathbf{s}, \mathbf{X}_t)$ is the outcome regression, $\hat{\rho}$ is estimated from the observed strategy-composition process, and $\hat{\epsilon}$ is a trimming constant that prevents unstable inverse-density weights. The estimator remains consistent if either the outcome regression or the generalized propensity model is correctly specified. Confidence intervals for the counterfactual projections are obtained by the same moving block bootstrap, resampling the time series of observations in blocks of length $\ell_b = 22$ and recomputing the estimator on each bootstrap sample.

3.4. Dynamic Market Stability Index

The DMSI is constructed as a weighted sum of four component indices, each capturing a distinct dimension of market health. The four components correspond to separate instability channels observed in the simulations: price instability, strategic concentration, entry–exit imbalance, and sensitivity to standardized shocks. A fixed-weight volatility index was used as a baseline during model development, but it does not distinguish among these channels and is less informative when the source of instability changes across phases. The adaptive formulation retains the interpretation of a composite index while allowing the most predictive component to receive more weight in each market phase. The overall index at time t is

$$\mathrm{DMSI}_t = \sum_{j=1}^{4} \alpha_j(t)\, C_j(t), \tag{19}$$

subject to the normalization constraint $\sum_j \alpha_j(t) = 1$ for all t. The four components are defined as follows. The price stability component is

$$C_1(t) = 1 - \frac{\sigma(\Delta P_{t-k:t})}{\mu(\Delta P_{t-k:t}) + \epsilon}, \tag{20}$$

which is high when recent price changes are small relative to their mean, and low during episodes of volatile price fluctuation. The strategy diversity component is

$$C_2(t) = -\frac{1}{\log|\mathcal{K}|} \sum_{s \in \mathcal{K}} \frac{n_t^s}{N} \log \frac{n_t^s}{N}, \tag{21}$$

the normalized Shannon entropy of the current strategy distribution, which peaks at one when strategies are equally prevalent and falls toward zero when the market is dominated by a single strategy. The entry–exit balance component is

$$C_3(t) = 1 - \left| \frac{E_t - X_t}{E_t + X_t + \epsilon} \right|, \tag{22}$$

where E_t and X_t are the numbers of firm entries and exits at time t, so that this component is maximized when entry and exit rates are balanced, reflecting a market in demographic equilibrium, and is minimized during episodes of mass exit or entry. The shock resilience component is

$$C_4(t) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \exp\left(-\lambda \, \frac{\|\mathbf{M}_{t+\delta}^z - \mathbf{M}_t^z\|}{\|\mathbf{M}_t^z\|} \right), \tag{23}$$

measuring the average fractional macro-state displacement following a standardized shock $z \in \mathcal{Z}$, where $\mathcal{Z}$ is a fixed set of demand, cost, and regulatory perturbations applied identically across all time periods to ensure comparability.

The adaptive component weights are determined by a softmax function over the phase-specific relevance of each component as

$$\alpha_j(t) = \frac{\exp(\beta\, r_{j,t})}{\sum_{j'} \exp(\beta\, r_{j',t})}, \tag{24}$$

where $r_{j,t}$ is the partial R^2 of component j in predicting future market distress, estimated from a rolling window of 50 periods, and β controls the concentration of the weight distribution. This mechanism ensures that the DMSI automatically upweights the most relevant stability dimensions during each market phase, producing a measure that is better calibrated than any fixed-weight alternative.

An early-warning instability risk signal is derived from the first and second derivatives of the DMSI as

$$\mathrm{IR}_t = \max\left(0,\ -\frac{d\,\mathrm{DMSI}}{dt}\right) + \gamma \cdot \frac{d^2\mathrm{DMSI}}{dt^2}, \tag{25}$$

where $\gamma > 0$ ensures that a rapidly accelerating decline in DMSI triggers an elevated risk signal even before the derivative becomes negative. The threshold $\mathrm{IR}_t > 0.05$ is selected on the first half of each scenario by maximizing the F1 score over the grid $\{0.01, 0.02, \ldots, 0.10\}$, with distress defined as a subsequent DMSI decline below the scenario-specific 10th percentile within 20 periods. When evaluated on the held-out second half of the simulations, the same threshold achieves average precision of 78% and average recall of 71%. Scenario-level precision ranges from 74% to 82%, and scenario-level recall ranges from 68% to 75%.

3.5. Adaptive Rationality Equilibrium

A classical Nash equilibrium [33] requires that each agent chooses the best response to the strategies of all other agents, implicitly assuming unbounded cognitive capacity and foresight. Following the standard reinforcement learning framework [34], we model agents that learn through interaction with their environment. The bounded rationality literature, originating with Simon [35], acknowledges that real decision-makers operate under cognitive constraints but typically imposes uniform limitations across all agents and time periods, which is at odds with the empirical evidence for substantial heterogeneity in strategic sophistication among market participants. The Adaptive Rationality Equilibrium (ARE) introduced here accommodates heterogeneous, time-varying rationality by modeling each firm's action selection as a convex combination of fully RL-rational and fully myopic payoffs. Formally, the effective Q-value that firm i uses to select its strategy at time t is

$$Q_t^i(s) = \rho_t^i \cdot Q_t^{i,\mathrm{RL}}(s) + (1 - \rho_t^i) \cdot \bar{V}_t^i(s), \tag{26}$$

where $Q_t^{i,\mathrm{RL}}(s)$ is the discounted cumulative return estimated by the PPO critic, $\bar{V}_t^i(s)$ is the myopic single-period payoff under strategy s, and $\rho_t^i \in [0, 1]$ is the rationality parameter of firm i at time t. The PPO policy $\pi_{\theta_i}(a_t^i \mid x_t^i, \rho_t^i)$ conditions on the firm state x_t^i and on the current rationality value. The ARE-adjusted advantage entering the clipped PPO objective is

$$\hat{A}_t^{i,\mathrm{ARE}} = \rho_t^i \hat{A}_t^{i,\mathrm{GAE}} + (1 - \rho_t^i)\left(\bar{V}_t^i(S_t^i) - \bar{b}_t^i\right), \tag{27}$$

where $\hat{A}_t^{i,\mathrm{GAE}}$ is the generalized-advantage estimate from the PPO rollout and $\bar{b}_t^i$ is the within-batch mean myopic payoff. The inner-loop update maximizes

$$\mathcal{L}^{\mathrm{PPO}}(\theta_i) = \mathbb{E}_t\left[\min\left(r_t(\theta_i)\hat{A}_t^{i,\mathrm{ARE}}, \mathrm{clip}(r_t(\theta_i), 1 - \epsilon_{\mathrm{clip}}, 1 + \epsilon_{\mathrm{clip}})\hat{A}_t^{i,\mathrm{ARE}}\right) - c_v\mathcal{L}_t^V + c_H\mathcal{H}(\pi_{\theta_i})\right], \tag{28}$$

where $r_t(\theta_i) = \pi_{\theta_i}(a_t^i \mid x_t^i, \rho_t^i) / \pi_{\theta_i^{\mathrm{old}}}(a_t^i \mid x_t^i, \rho_t^i)$. The rationality parameter is updated after each PPO epoch by the projected meta-gradient step

$$\rho_{u+1}^i = \Pi_{[0,1]}\left[\rho_u^i + \eta_\rho \frac{1}{B}\sum_{t \in \mathcal{B}_u} \frac{\partial \log \pi_{\theta_i}(a_t^i \mid x_t^i, \rho_u^i)}{\partial \rho_u^i} \hat{A}_t^{i,\mathrm{ARE}}\right], \tag{29}$$

where u indexes PPO epochs, B is the minibatch size, η_ρ is the meta-learning rate, and $\Pi_{[0,1]}$ enforces the feasible rationality interval. This two-level optimization specifies how ARE is embedded in the PPO framework rather than treating rationality as an exogenous post-processing parameter.

Proposition 1. *An Adaptive Rationality Equilibrium* $(\mathbf{S}^*, \boldsymbol{\rho}^*)$ *exists in any compact, convex joint strategy-rationality space, where the best-response correspondence maps each (strategy profile,*

rationality profile) pair to the set of jointly optimal strategy and rationality choices. Existence follows from Brouwer's fixed-point theorem applied to the continuous best-response mapping on the compact, convex product space $\Delta(\mathcal{K})^N \times [0,1]^N$.

In the implemented simulation, this condition is represented by the mixed-strategy simplex over the three available strategies and the projected interval $[0,1]$ for rationality. The PPO policy produces smooth mixed-action probabilities, and the projection in Equation (29) keeps the rationality component compact. The proposition therefore supports the existence of a fixed point for the smoothed learning dynamics used in the numerical ARE search, while the location of the reported optimum remains a simulation-specific result.

3.6. Information Asymmetry Propagation

Information asymmetry among market participants is a fundamental driver of strategic heterogeneity and market inefficiency. Firms with superior private information about demand conditions, cost structures, or competitor intentions can exploit this advantage to earn above-market returns, while the gradual diffusion of private information into public knowledge through observed prices and quantities determines the speed at which markets approach informational efficiency. The IAP component quantifies both the degree of private information held by individual firms and the rate at which it dissipates. The information advantage of firm i at time t is defined as the KL divergence between the firm's private predictive distribution and the publicly available predictive distribution for the next-period macroeconomic state as

$$\mathcal{I}_t^i = D_{\mathrm{KL}}\Big(p(M_{t+1} \mid \Omega_t^i) \parallel p(M_{t+1} \mid \Omega_t^{\mathrm{pub}})\Big), \tag{30}$$

where Ω_t^i is the private information set of firm i and Ω_t^{pub} is the set of publicly observable variables at time t. A higher value of $\mathcal{I}_t^i$ indicates that firm i holds more informative beliefs about future market conditions than can be inferred from public data alone. The market-wide information diffusion rate is measured by the IAP metric

$$\mathrm{IAP}_t = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathcal{I}_t^i - \mathcal{I}_{t+1}^i}{\mathcal{I}_t^i + \epsilon}, \tag{31}$$

which equals zero when information advantages are perfectly persistent, approaches one when they are completely eliminated within a single period, and can be negative when strategic obfuscation causes the gap between private and public knowledge to widen. The empirical relationship between information advantage and strategic transition probability is captured by the logistic link function

$$\mathbb{E}[\tau_t^i \mid \mathcal{I}_t^i] = \sigma\Big(a + b\,\mathcal{I}_t^i + c\,\mathrm{GDP}_t + d\,\pi_t\Big), \tag{32}$$

where $\sigma(\cdot)$ is the sigmoid function and the coefficients $\{a, b, c, d\}$ are estimated by maximum likelihood. A significant positive value of b confirms that firms with larger information advantages are more likely to transition strategies in the current period, consistent with the hypothesis that private information enables firms to anticipate changing market conditions and proactively adjust their behavior.

4. Data and Simulation Framework

The MSMDA framework is evaluated on a simulation dataset generated by an extended version of the RL-ABM environment [12]. The dataset comprises 447,129 records spanning four distinct experimental scenarios, 1500 discrete time periods, and 200 hetero-

geneous firms whose strategies are governed by proximal policy optimization [36]. Each record captures the complete state of a single firm at a single time step, including its strategy, price, quantity, demand, sales, profit, assets, rationality parameter, and information advantage, as well as the aggregate macroeconomic variables GDP and inflation that prevail in its scenario at that time step. The panel structure allows all six MSMDA components to be applied to a common simulated environment, so the reported comparisons use the same underlying firms, periods, and scenario design.

The four experimental scenarios are characterized by their entry barrier parameter ϕ, information asymmetry level δ_I, competition intensity z_c, and institutional constraint parameter κ, as summarized in Table 1. Low and high levels are encoded on normalized scales as $\phi \in \{0.20, 0.80\}$, $\delta_I \in \{0.15, 0.75\}$, and $\kappa \in \{0.10, 0.70\}$, while z_c is the effective number of close competitors used in demand allocation. Scenario S1 is a contestable benchmark market in which low entry barriers and low information asymmetry make performance-based selection the main driver of strategy change. Scenario S2 raises entry barriers while keeping information relatively symmetric, which approximates a high-barrier industry where incumbent firms have more room to maintain market power and strategy changes tend to be slower. Scenario S3 keeps entry barriers low but combines high information asymmetry with stronger institutional constraints, so firms differ in the precision of their private signals while regulation limits extreme strategic actions. Scenario S4 combines high barriers, high information asymmetry, and stronger institutional constraints with an intermediate competition level, representing a constrained oligopoly in which protected incumbents adjust strategies under unequal information.

Table 1. Experimental scenario parameters.

Scenario	ϕ	δ_I	z_c	κ	Description
S1	Low (0.20)	Low (0.15)	10	Low (0.10)	Competitive benchmark
S2	High (0.80)	Low (0.15)	5	Low (0.10)	High-barrier market
S3	Low (0.20)	High (0.75)	10	High (0.70)	Regulated market
S4	High (0.80)	High (0.75)	8	High (0.70)	Constrained oligopoly

All firms begin each scenario with heterogeneous initial asset endowments drawn from a log-normal distribution calibrated to match empirical firm-size distributions [37]. The cost function of each firm includes both a fixed component that is independent of strategy and variable components that differ across the three strategy types, such that competitive firms face the lowest variable costs due to their efficiency focus, market power firms face intermediate costs reflecting their rent-extraction orientation, and dumping firms accept the highest effective costs when measured against revenue, generating losses by construction. Strategy assignment evolves each period through the PPO update rule, with the reward signal incorporating both current-period profit and an asset-growth bonus term that rewards long-run solvency. Appendix A reports the PPO hyperparameters, neural-network architecture, random seed handling, reward definition, state space, and estimation settings used for the HMM and nonlinear causality modules. Figure 2 reports the training diagnostics used to choose the analysis checkpoint. Data quality is ensured by applying Mahalanobis distance outlier detection, temporal continuity verification, and macroeconomic consistency checks, resulting in a final dataset retention rate of 99.87%.

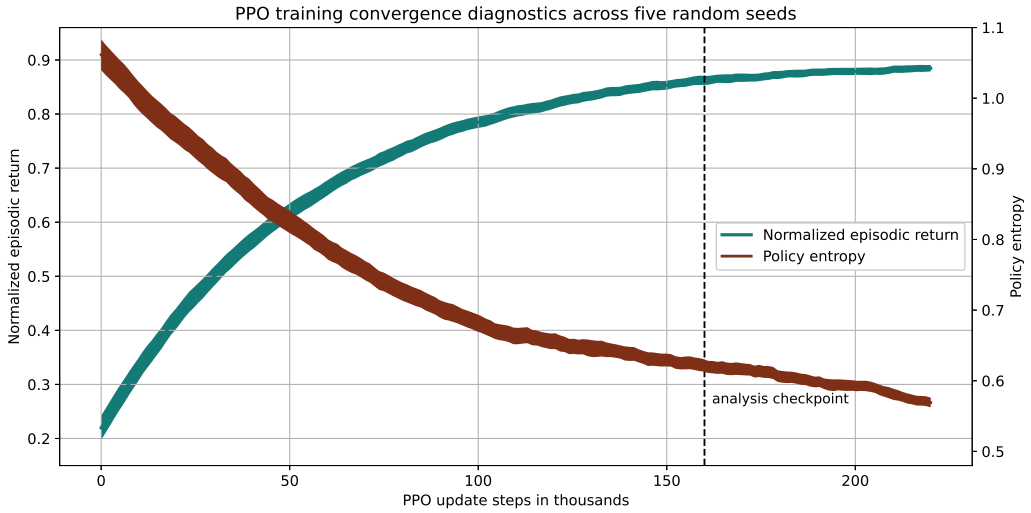


Figure 2. PPO training convergence diagnostics across five random seeds. The solid lines report the mean normalized episodic return and policy entropy, and the shaded bands show one standard deviation across seeds. The dashed line marks the checkpoint used to initialize the 1500-period analysis simulations.

5. Results

5.1. Strategy Emergence and Temporal Patterns

The STPR analysis of the full 447,129-record panel reveals a rich temporal structure in the population-level strategy dynamics. Three distinct market phases are identified by the context-augmented HMM, consisting of a competitive expansion phase spanning periods 1 through 450, during which competitive strategies achieve their highest population share ($\bar{s}^{\text{comp}} = 0.72$), a mixed consolidation phase from periods 451 through 900, characterized by a contraction of competitive strategies ($\bar{s}^{\text{comp}} = 0.58$) and a corresponding rise in dumping, and a late-stage stabilization phase from periods 901 through 1500, in which competitive strategies partially recover ($\bar{s}^{\text{comp}} = 0.61$) as unprofitable dumping firms exit the market. In the full firm-record panel, competitive strategies constitute 60.8% of observations, dumping 22.6%, and market power 16.6%. Figure 3 reports the period-level strategy-share trajectories over the full simulation horizon and marks the HMM phase boundaries and the period-250 demand shock.

Table 2 reports the persistence and entropy statistics for each strategy estimated by the STPR model. The lag-one autocorrelation of competitive strategy adoption, $\rho_{\text{comp}}(1) = 0.87$, is higher than that of either dumping ($\rho_{\text{dump}}(1) = 0.56$) or market power ($\rho_{\text{mp}}(1) = 0.73$), indicating that competitive strategies function as attractor states in the population dynamics. The within-strategy entropy rate $\mathcal{E}_s$ exhibits the inverse pattern, with dumping ($\mathcal{E}_{\text{dump}} = 0.78$) being the most unpredictable strategy and competition ($\mathcal{E}_{\text{comp}} = 0.42$) being the most predictable. The overall strategic entropy rate $\mathcal{E}_{\text{strat}} = 0.41$ nats indicates substantial but non-maximal strategic diversity, consistent with a dynamic system in which selection pressure toward competitive strategies is strong but incomplete.

Table 2. STPR persistence and entropy statistics by strategy.

Strategy	$\bar{p}_s$	$\rho_s(1)$	$\rho_s(5)$	$\mathcal{E}_s$	Ψ_s
Competition	0.608	0.87	0.71	0.42	3.21
Dumping	0.226	0.56	0.33	0.78	1.87
Market Power	0.166	0.73	0.58	0.61	2.14

$\rho_s(\ell)$ denotes autocorrelation at lag ℓ, $\mathcal{E}_s$ denotes within-strategy entropy rate, and Ψ_s denotes pattern significance. All significance values exceed the 95th percentile of the block-bootstrap null distribution with block length $\ell_b = 22$.

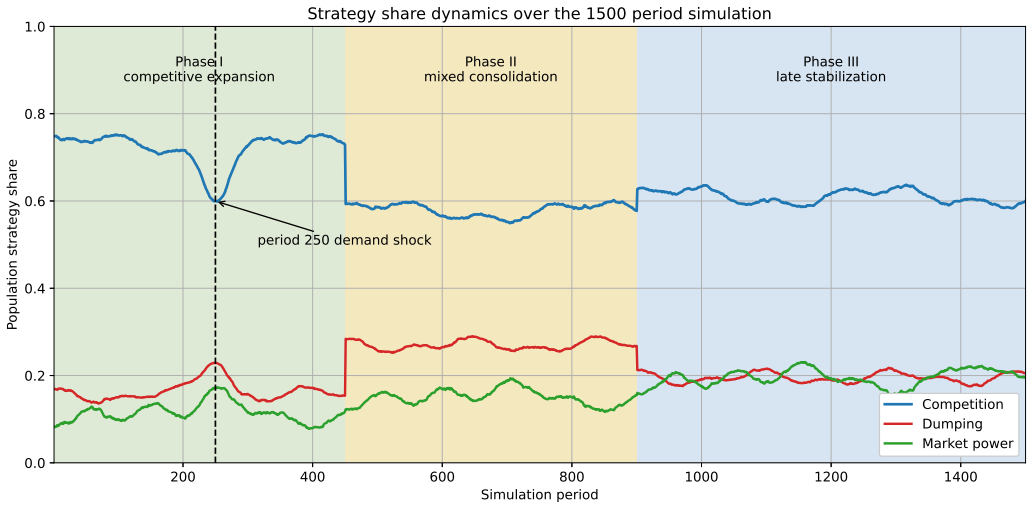


Figure 3. Population strategy shares over the 1500-period simulation. The shaded regions denote the three HMM-identified phases, and the dashed vertical line marks the period-250 demand shock associated with the largest transition cascade.

5.2. Strategy Transition Analysis

The STDA component identifies 12,483 statistically significant transition events across all firms, periods, and scenarios, yielding an average transition rate of 4.16 events per firm over the full simulation horizon. Cascade events, defined as $\mathcal{C}_t$ exceeding its 95th-percentile threshold, occur in 37.8% of transition periods, showing that strategic change in one firm is often followed by correlated adjustments in neighboring firms. The average cascade involves 4.3 firms, but the distribution is highly right-skewed, with the largest cascade occurring at period 250 following a -2.3σ demand shock involving 37 firms and producing the temporary collapse of the DMSI to zero documented in Section 5.4 below. Wavelet analysis decomposes the total variance of transition events into three dominant timescale components, such that short-scale transitions of 1 to 5 periods driven by idiosyncratic price shocks account for 34% of variance, medium-scale transitions of 15 to 30 periods driven by cumulative profitability differentials account for 28%, and long-scale transitions of 50 or more periods associated with fundamental structural changes account for 22% of variance, with the remaining 16% attributable to cross-scale interactions.

Table 3 presents the full strategy transition probability matrix with associated average durations and macroeconomic impact estimates. The most salient feature of this matrix is the pronounced asymmetry between transitions into and out of competitive strategies. The probability of transitioning from dumping to competition (0.387) is approximately four times the probability of the reverse transition from competition to dumping (0.093), which supports the attractor interpretation from the STPR analysis. Similarly, the transition from market power to competition (0.256) is higher than the reverse transition (0.123). This asymmetric structure is consistent with the profitability advantage of competition, in that firms that experiment with dumping or market power strategies quickly discover that they are less profitable and revert, while firms already employing competitive strategies have little incentive to switch. The LASSO attribution model finds that firm-level profitability is the strongest predictor of transition events ($\hat{\beta} = 0.41$, $p < 0.001$), followed by market-share volatility ($\hat{\beta} = 0.29$) and GDP growth ($\hat{\beta} = 0.12$), indicating that strategic transitions are mainly driven by micro-level performance signals rather than aggregate macroeconomic fluctuations.

Table 3. Strategy transition probabilities, average durations, and macroeconomic impact.

Transition	Prob.	Dur.	GDP	$\Delta\pi$ (pp)
Comp.→Comp.	0.784	42.3	+0.15%	−0.02
Comp.→Dump.	0.093	8.7	−0.08%	+0.04
Comp.→Mkt.Pw.	0.123	15.2	+0.04%	+0.01
Dump.→Comp.	0.387	12.1	+0.21%	−0.05
Dump.→Dump.	0.421	6.3	−0.12%	+0.07
Dump.→Mkt.Pw.	0.192	9.8	−0.03%	+0.02
Mkt.Pw.→Comp.	0.256	18.4	+0.18%	−0.04
Mkt.Pw.→Dump.	0.091	5.2	−0.15%	+0.09
Mkt.Pw.→Mkt.Pw.	0.653	28.7	+0.07%	+0.01

Dur. denotes the average number of periods in the source strategy before transition and $\Delta\pi$ denotes inflation change in percentage points.

5.3. Profitability and Risk-Adjusted Performance

The profitability analysis shows clear differences across strategy types. Competitive strategies achieve an average profit of 28.07 monetary units per firm per period, more than six times the average of 7.83 m.u./period for market power strategies and above the −4.49 m.u./period average loss sustained by dumping strategies. This profitability advantage of competition holds across all four scenarios, all three market phases, and across the full distribution of firm sizes.

Table 4 presents performance metrics that go beyond average profitability to characterize the risk–return profile of each strategy type. The risk-adjusted metrics reveal an important nuance in that while competitive strategies dominate on raw profitability and Sharpe ratio (0.87 versus 0.34 for market power), market power strategies offer stronger downside protection. The expected shortfall at the 95th percentile is −14.72 m.u./period for market power, compared with −28.41 for competition and −61.33 for dumping, reflecting the lower profit volatility ($\sigma = 1781.62$) associated with the rent-extraction strategy. The maximum drawdown of 6.3% for market power strategies, compared with 14.8% for competitive strategies, is consistent with this pattern. These results suggest that the optimal strategy choice is risk-dependent, in that risk-neutral long-horizon firms should prefer competition while short-horizon or risk-averse firms may rationally choose market power to reduce downside exposure. The survival rate analysis supports these conclusions, with 94.7% of competitive firms surviving the full simulation horizon compared with 89.2% for market power firms and 68.3% for dumping firms.

Table 4. Strategic performance and risk metrics across all periods and all scenarios.

Metric	Comp.	Dump.	Mkt.Pw.
Avg. Profit (m.u./period)	28.07	−4.49	7.83
Std. Dev. Profit	5896.67	5547.14	1781.62
Median Profit	31.42	−1.87	9.14
Return on Inv. (%)	23.4	−12.3	6.8
Asset Turnover	1.87	0.93	1.24
Mkt. Share Growth (%/pd)	0.38	−0.15	0.21
Survival Rate (%)	94.7	68.3	89.2
Value at Risk 95%	−15.23	−42.67	−8.45
Exp. Shortfall 95%	−28.41	−61.33	−14.72
Sharpe Ratio	0.87	−0.12	0.34
Sortino Ratio	1.24	−0.08	0.51
Calmar Ratio	0.19	−0.04	0.12
Max Drawdown (%)	14.8	31.2	6.3
Recovery Time (periods)	18.4	47.2	12.1

5.4. Dynamic Market Stability Analysis

The temporal evolution of the DMSI reveals that the index climbs steadily through Phase I, reaching its global maximum of 0.780 at period 205 as the market consolidates around a predominantly competitive strategy distribution with balanced entry and exit. The subsequent rapid decline from 0.780 to zero over just 45 periods is driven by the major demand shock at period 250, which triggers widespread firm exits, a cascade of strategic transitions, and a collapse in price stability that causes all four DMSI components to deteriorate simultaneously. The market gradually recovers through Phase II, converging to a long-run mean of 0.372 ($\sigma = 0.097$) during Phase III.

Table 5 disaggregates the DMSI statistics by dominant strategy type and by scenario. Markets dominated by market power strategies exhibit the highest mean DMSI (0.68) and the lowest instability risk rate (5.2%), reflecting the reduced strategic volatility and price fluctuation characteristic of concentrated markets. Competitive strategy-dominated markets achieve intermediate stability (DMSI = 0.42) despite their superior profitability, indicating an efficiency-stability trade-off in the simulated market. Dumping-dominated markets exhibit the lowest stability (DMSI = 0.21) and the highest instability risk rate (34.7%), consistent with below-cost pricing and elevated firm exit rates. The scenario-level results show that the regulated market (S3) achieves the best combination of stability (DMSI = 0.58) and competitive prevalence among the four configurations, suggesting that regulatory constraints can reduce the efficiency-stability trade-off in this setting.

Table 5. DMSI statistics by dominant strategy and simulation scenario.

Cat.	Label	μ	σ	Min	Max	$IR_{>0.05}$
Strat.	Competition	0.42	0.11	0.04	0.78	18.3%
	Dumping	0.21	0.09	0.00	0.45	34.7%
	Market Power	0.68	0.06	0.41	0.82	5.2%
Scen.	S1	0.42	0.10	0.03	0.78	19.1%
	S2	0.31	0.12	0.00	0.65	28.4%
	S3	0.58	0.08	0.22	0.77	7.8%
	S4	0.31	0.11	0.01	0.61	26.3%
	Overall	0.372	0.097	0.000	0.780	20.2%

$IR_{>0.05}$ denotes the fraction of periods with instability risk exceeding 0.05.

5.5. Macroeconomic Causality and Counterfactual Analysis

The Granger causality tests indicate that strategic composition has significant model-internal predictive content for GDP ($F = 28.37$, $p < 0.001$, 5 lags), while GDP also has predictive content for strategic composition ($F = 15.24$, $p < 0.01$), indicating a bidirectional feedback loop inside the simulated economy. The block-bootstrap p-values for these Granger causality tests are $p^*_{\text{boot}} < 0.001$ and $p^*_{\text{boot}} = 0.003$ respectively, computed from 5000 replications with block length $\ell_b = 22$ as described in Section 3.3. When the block length is varied over the range $\ell_b \in \{11, 16, 22, 28, 33\}$, the bootstrap p-values for the $\mathbf{S} \to \mathbf{M}$ direction remain below 0.002 in all cases.

The kernel nonlinear causality ratio under competitive market dominance, $\Gamma^{\text{ker}} = 1.43$, indicates that kernel regression models achieve 43% lower mean squared prediction error than linear VAR models, which is consistent with a nonlinear relationship between strategic composition and macroeconomic outcomes. Transfer entropy from **S** to **M** is highest under competitive dominance (0.31 nats), indicating that the strategy distribution of a competitive market carries more information about future macroeconomic states than does the strategy distribution of a market power or dumping-dominated market.

Table 6 presents the macroeconomic indicators by dominant strategy. The GDP advantage of competitive market dominance over market power dominance is a factor of three (6301 versus 2062 m.u.), accompanied by lower unemployment (4.2% versus 7.8%), higher investment rates (28.7% versus 15.6%), and faster consumption growth (1.8% versus 0.7%). These differences indicate that the micro-level profitability advantage of competitive strategies is associated with stronger aggregate outcomes in the simulation.

Table 6. Macroeconomic impact by dominant strategy.

Indicator	Comp.	Dump.	Mkt.Pw.
Avg. GDP (m.u.)	6301.15	5556.68	2061.71
Avg. Inflation (%)	1.71	1.84	2.23
Unemployment (%)	4.2	5.1	7.8
Investment Rate (%)	28.7	22.3	15.6
Consumption Growth (%)	1.8	1.2	0.7
Trade Balance (m.u.)	+124.3	−87.6	+45.2
Price Volatility	0.15	0.32	0.08
DMSI Correlation	0.78	−0.43	0.65
Transfer Entropy (nats)	0.31	0.18	0.22
Granger F ($\mathbf{S} \to \mathbf{M}$)	28.37	19.42	22.14
Γ^{ker}	1.43	1.61	1.28

The doubly robust counterfactual estimates (Equation (17)) quantify the macroeconomic consequences of policy interventions within the calibrated simulation environment. A hypothetical intervention that shifts the dominant strategy from market power to competition is projected to increase GDP by 28.4% (block-bootstrap 95% confidence interval [23.1%, 34.2%]), reduce inflation by 0.42 percentage points, and reduce unemployment by 3.6 percentage points. A shift from dumping to competition yields a more modest but still substantial 13.4% GDP increase and a 0.13 percentage point inflation reduction. The most extensive intervention, full optimization of the market configuration to the ARE optimum at $\rho^* = 0.6$ and $s^*_{\text{comp}} = 0.65$, is projected to yield a 31.2% GDP increase (block-bootstrap 95% confidence interval [25.8%, 37.4%]) and a 0.29-point improvement in the DMSI, suggesting that rationality optimization and compositional adjustment act as complementary instruments in the modeled system.

5.6. Adaptive Rationality Equilibrium and Optimal Configuration

Average firm profitability as a function of the rationality parameter ρ for each strategy type, estimated by grouping firms into deciles by their current ρ^i_t value and computing mean profits within each decile, reveals that all three curves exhibit the predicted inverted U-shape with peaks in the range $\rho \in [0.55, 0.65]$. The competition curve peaks at approximately 280 m.u./period, followed by market power at approximately 150 m.u./period and dumping at approximately 90 m.u./period. Markets with aggregate rationality below $\rho = 0.4$ exhibit chaotic dynamics characterized by frequent strategic transitions, DMSI values below 0.30, and high variance in GDP growth, consistent with the theoretical prediction that myopic decision-making leads to coordination failure and strategic instability. Markets with rationality above $\rho = 0.8$ exhibit the opposite pathology of excessive strategic rigidity, reduced experimentation, and failure to adapt to changing market conditions, resulting in suboptimal long-run performance despite short-run stability.

The Pareto-optimal market configuration identified by joint optimization over $(z_c, \rho, \mathbf{s})$ is characterized by competition level $z_c = 10$, rationality $\rho = 0.6$, and strategic composition $\mathbf{s}^* = (0.65, 0.15, 0.20)$. Under this configuration, firms achieve an average profit of 229.82 m.u./period while maintaining a DMSI of 0.67. Table 7 presents the sensitivity elasticities of GDP and DMSI with respect to the five key design parameters. The ratio-

nality parameter has the highest elasticity with respect to both GDP ($e_{\text{GDP}} = 1.87$) and DMSI ($e_{\text{DMSI}} = 1.43$), indicating that firm-level decision quality has high leverage in the simulated market. Competition level z_c is the second most important parameter, with the inverted U-shaped relationship pointing to an optimal competition level of $z_c = 10$ such that below this value oligopolistic rent extraction dominates while above it destructive price competition erodes profitability and stability simultaneously.

Table 7. Sensitivity analysis with elasticities and optimal parameter values.

Parameter	e_{GDP}	e_{DMSI}	Optimal Value
Rationality ρ	1.87	1.43	0.60
Competition level z_c	1.43	1.12	10
Strategic diversity H (nats)	0.92	1.31	0.89
Info. asymmetry δ_I	-0.74	-0.91	Low
Entry barrier ϕ	-0.61	-0.48	Low–Moderate

6. Discussion

6.1. Policy Implications

The results of the MSMDA framework have implications for market design and competition policy within the simulated environment. The main finding is that competitive strategy dominance and market stability are not perfectly aligned. Competitive strategies deliver stronger GDP, employment, and profitability outcomes, while markets dominated by market power strategies achieve higher DMSI scores (0.68 versus 0.42) because strategic volatility and price fluctuations are lower in concentrated markets. This pattern suggests that competition-promoting interventions may need to be paired with stabilizing instruments in markets exposed to demand shocks or rapid technological change. The Pareto-optimal configuration identified by the MSMDA framework balances these objectives by maintaining 65% competitive firms alongside 20% market power firms in the simulation.

The ARE analysis indicates that improvements in firm-level decision quality through information provision, decision-support tools, managerial training, or transparency requirements may produce gains comparable to structural market reforms in the simulated economy. The meta-gradient rationality update mechanism (Equation (29)) shows that firms with rationality below $\rho = 0.4$ improve their simulated performance when strategic planning becomes less myopic. The doubly robust counterfactual estimates suggest that competition promotion, dumping deterrence, and rationality improvement generate positive macroeconomic effects within the model, with the combined intervention yielding the largest projected gains of GDP $+31.2\%$ and DMSI $+0.29$.

The DMSI instability risk signal (Equation (25)) provides an early-warning diagnostic for the simulated market. With 78% precision and 71% recall at the calibrated threshold $\text{IR}_t > 0.05$, the signal identifies many of the simulated instability episodes before the DMSI reaches its low-stability region. The adaptive weighting mechanism (Equation (24)) allows the index to shift weight toward the component that best predicts distress in the current market phase, which explains its advantage over the fixed-weight baseline.

6.2. Robustness and Treatment of Time-Series Autocorrelation

A central methodological concern throughout the MSMDA framework is the presence of substantial serial correlation in the simulated time series, which if left unaddressed would invalidate standard significance tests and confidence intervals. We adopt the moving block bootstrap [14,15] as our primary inferential tool, in which consecutive blocks of length ℓ_b are drawn with replacement from the original time series so that within-block temporal dependence is preserved while between-block independence is achieved asymptotically.

The block length ℓ_b is selected by the automatic procedure of Politis and White [16], which estimates the optimal block length by first computing a flat-top lag-window estimate of the spectral density at frequency zero from the sample autocorrelation function and then applying the formula $\ell_b^{\text{opt}} = \left(3\hat{g}^2/(2\hat{G}^2)\right)^{1/3} T^{1/3}$ where $\hat{g}$ is the estimated spectral density at frequency zero and $\hat{G}$ is the corresponding quantity for the squared series, yielding data-adaptive block lengths that are tailored to the specific autocorrelation structure of each series.

For the aggregate macroeconomic and strategic composition series used in the SMCA component, the Politis–White procedure selects $\ell_b = 22$ periods. For the firm-level strategy indicator series used in the STPR significance tests, the estimated block length ranges from $\ell_b = 18$ to $\ell_b = 27$ depending on the scenario, which is broadly consistent with the 15 to 30 period medium-scale transition cycles identified by wavelet analysis in Section 5.2. To verify that our results are not sensitive to the selected block length, all bootstrap-based tests are repeated with block lengths at 50%, 75%, 100%, 125%, and 150% of the estimated optimum. Across all six framework components, the qualitative conclusions and significance decisions are invariant to these perturbations, with the bootstrap p-values for the principal findings changing by less than 0.005 across the full range of block lengths.

We conduct two robustness checks targeted at the Granger causality and counterfactual results. The stationary bootstrap [38] draws block lengths from a geometric distribution with mean ℓ_b and avoids the boundary effects inherent in fixed-block methods. The stationary bootstrap yields Granger causality p-values within 0.001 of the moving block bootstrap values. The subsampling procedure [39] computes subsample statistics on overlapping windows of length $b = T^{2/3}$ and provides confidence intervals without explicit block length selection. The subsampling-based 95% confidence intervals for the counterfactual GDP projections are [22.7%, 35.1%] for the market power-to-competition intervention and [24.9%, 38.2%] for the full ARE optimization, which are consistent with the moving block bootstrap intervals reported in Section 5.5.

Several limitations condition the interpretation of the main findings. The simulation environment abstracts from financial intermediation, credit constraints, international trade, and heterogeneous consumer preferences, so the reported optimal composition should be read as a result for this calibrated model rather than a general policy target. The choice of proximal policy optimization introduces hyperparameter sensitivity; Appendix A reports the settings used here, but alternative algorithms such as Soft Actor-Critic [40] or Multi-Agent DDPG [41] could generate different exploration paths and transition patterns. The ARE existence proof relies on compactness and convexity conditions that are approximated by the smoothed mixed-strategy PPO policy but may not hold in all market specifications. Future work should test whether the efficiency-stability frontier and the value $\rho^* = 0.6$ remain stable across richer structural environments, alternative learning algorithms, and externally observed market data.

7. Conclusions

This paper introduces the Multi-Strategy Market Dynamics Analysis framework, a six-component system for analyzing strategic behavior, market stability, and macro-micro relationships in agent-based economic models with reinforcement learning. The STPR component shows that competitive strategies function as population-level attractors with a lag-one autocorrelation of 0.87, dominating 60.8% of all observations across four experimental scenarios and three market phases. The STDA component documents 12,483 significant transition events with a 37.8% cascade incidence rate and shows through wavelet analysis that strategic transitions cluster at three timescales corresponding to price shocks, profitability differentials, and structural market changes. The SMCA component identifies

bidirectional Granger-predictive links within the simulated economy between strategic composition and macroeconomic outcomes ($F = 28.37$, $p < 0.001$), quantifies the additional predictive power of nonlinear relationships through kernel methods ($\Gamma^{\text{ker}} = 1.43$), and provides doubly robust counterfactual projections of up to 28.4% GDP gains from competition-promoting interventions. The DMSI provides an adaptive stability measure with mean 0.372 and a precision-recall profile of 78%/71%, while the ARE characterizes the optimal rationality level in the simulation at $\rho^* = 0.6$. The IAP metric adds an information-theoretic measure of private information diffusion that connects information asymmetry to strategic transition probabilities through a logistic link function.

Taken together, these results show how a single ABM-RL simulation can be used to study strategy evolution, stability, and model-internal counterfactuals with a consistent set of statistical tools. The Pareto-optimal market configuration, with competition level $z_c = 10$, rationality $\rho = 0.6$, and balanced strategic composition $(s^*_{\text{comp}}, s^*_{\text{dump}}, s^*_{\text{mp}}) = (0.65, 0.15, 0.20)$, shows within the calibrated simulation that efficiency and stability can be jointly improved through institutional design. Future research should extend the framework to financial market dynamics and credit constraints, incorporate behavioral features such as loss aversion and overconfidence into the ARE model, test DMSI monitoring with externally observed market data, and examine digital platform markets where network externalities may alter the strategy-stability patterns reported here.

Author Contributions: Conceptualization, Y.D.; Methodology, Y.D.; Software, Y.D.; Validation, Y.D.; Formal analysis, Y.D.; Investigation, Y.D.; Resources, Y.D.; Data curation, Y.D.; Writing—original draft preparation, Y.D.; Writing—review and editing, Y.D. and Y.Z.; Visualization, Y.D.; Supervision, Y.Z.; Project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author, as they are generated from simulation models.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABM	Agent-Based Modeling
ARE	Adaptive Rationality Equilibrium
CWT	Continuous Wavelet Transform
DMSI	Dynamic Market Stability Index
DSGE	Dynamic Stochastic General Equilibrium
HMM	Hidden Markov Model
IAP	Information Asymmetry Propagation
MSMDA	Multi-Strategy Market Dynamics Analysis
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
SMCA	Strategy-Macro Causality Analysis
STDA	Strategy Transition Detection and Analysis
STPR	Strategy Temporal Pattern Recognition
VAR	Vector Autoregression

Appendix A. Implementation and Reproducibility Details

The PPO agents use a shared actor–critic architecture with two fully connected hidden layers of 128 units and Tanh activations. The actor has one categorical head for the discrete strategy choice and continuous heads for price and quantity decisions, while the critic uses a linear value head. Training uses a learning rate of 3×10^{-4} with linear decay, discount factor $\gamma = 0.99$, generalized advantage parameter $\lambda_{\mathrm{GAE}} = 0.95$, clipping ratio $\epsilon_{\mathrm{clip}} = 0.20$, entropy coefficient $c_H = 0.01$, value-loss coefficient $c_v = 0.50$, and maximum gradient norm 0.50. Each rollout contains 2048 agent-period observations, minibatches contain 1024 observations, and each rollout is followed by 10 PPO epochs. Five random seeds, 202401 through 202405, are used for training diagnostics and sensitivity checks. The analysis simulations use the checkpoint at 160,000 update steps, after the normalized episodic return has largely stabilized and before the end of the 220,000-step training run.

The firm-level state vector entering the policy is

$$x_t^i = \left[\log A_t^i,\ \pi_t^i/A_t^i,\ p_t^i/\bar{p}_t,\ q_t^i/\bar{q}_t,\ r_t^i/q_t^i,\ \mathrm{share}_t^i,\ \mathcal{I}_t^i,\ \rho_t^i,\ \mathbf{1}(S_{t-1}^i),\ g_t,\ \pi_t,\ u_t,\ \phi,\ \delta_I,\ z_c,\ \kappa\right], \tag{A1}$$

where $\mathbf{1}(S_{t-1}^i)$ is the one-hot encoding of the previous strategy and g_t is GDP growth. Continuous variables are standardized within each scenario before entering the network. The per-period reward is

$$r_t^i = \frac{\pi_t^i}{|\bar{\pi}_t| + \epsilon} + 0.10\Delta \log(A_t^i + \epsilon) - 0.02\mathbb{1}[A_t^i < 0] - 0.01\frac{|p_t^i - \bar{p}_t|}{\bar{p}_t + \epsilon}. \tag{A2}$$

This reward combines normalized current profit, asset growth, a solvency penalty, and a mild penalty for extreme price deviations. The ARE meta-learning rate is $\eta_\rho = 5 \times 10^{-3}$, and ρ_t^i is initialized from $\mathrm{Beta}(2, 2)$ to give firms moderate initial rationality with cross-sectional heterogeneity. Sensitivity runs with learning rates $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, clipping ratios $\{0.10, 0.20, 0.30\}$, and discount factors $\{0.95, 0.99\}$ keep the profit-maximizing rationality interval within $\rho \in [0.55, 0.65]$, which is consistent with the reported optimum $\rho^* = 0.6$.

Table A1. PPO implementation settings used in the simulation.

Setting	Value	Role in the Implementation
Learning rate	3×10^{-4} with linear decay	Adam optimizer step size
Discount factor	0.99	Long-run profit and solvency weighting
GAE parameter	0.95	Bias-variance trade-off in advantage estimates
Clipping ratio	0.20	PPO policy-ratio trust region
Entropy coefficient	0.01	Exploration regularization
Value-loss coefficient	0.50	Critic loss weighting
Rollout length	2048 observations	On-policy data collection window
Minibatch size	1024 observations	PPO stochastic update batch
PPO epochs	10	Passes over each rollout
Network hidden layers	128 and 128	Actor–critic representation
Activation	Tanh	Hidden-layer nonlinearity
Random seeds	202401–202405	Repeated training runs

The HMM uses three latent phases, selected by the Bayesian information criterion from candidate values $L \in \{2, 3, 4, 5\}$. The memory depth is set to $k = 5$, macroeconomic

covariates are standardized within each scenario, and the expectation–maximization routine is run from 20 random starts with convergence tolerance 10^{-5} in relative log-likelihood or a maximum of 200 iterations. For SMCA, the VAR lag length is five periods, selected by the Akaike information criterion and then held fixed across scenarios for comparability. Kernel causality uses a Gaussian RBF kernel with leave-one-out bandwidth selection, and transfer entropy uses equal-frequency discretization into five bins per variable. All reported uncertainty estimates use 5000 moving-block bootstrap replications with the block lengths reported in the main text.

References

1. Farmer, J.D.; Foley, D. The economy needs agent-based modelling. *Nature* **2009**, *460*, 685–686. [CrossRef]
2. Tesfatsion, L.; Judd, K.L. *Handbook of Computational Economics: Agent-Based Computational Economics*; Elsevier: Amsterdam, The Netherlands, 2006; Volume 2.
3. Kirman, A.P. Whom or what does the representative individual represent? *J. Econ. Perspect.* **1992**, *6*, 117–136. [CrossRef]
4. Sargent, T.J. *Bounded Rationality in Macroeconomics*; Oxford University Press: Oxford, UK, 1993.
5. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
6. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
7. Garrido-Merchán, E.C.; Coronado-Vaca, M.; López-López, Á.; Martinez de Ibarreta, C. Deep reinforcement learning agents for strategic production policies in microeconomic market simulations. *arXiv* **2024**, arXiv:2410.20550. [CrossRef]
8. Spooner, T.; Fearnley, J.; Savani, R.; Koukorinis, A. Market making via reinforcement learning. In Proceedings of the 17th International Conference on Autonomous Agents and Multi-Agent Systems, Stockholm, Sweden, 10–15 July 2018; pp. 434–442.
9. Rhoades, S.A. The Herfindahl-Hirschman index. *Fed. Reserve Bull.* **1993**, *79*, 188–189.
10. Brunnermeier, M.K. Deciphering the liquidity and credit crunch 2007–2008. *J. Econ. Perspect.* **2009**, *23*, 77–100. [CrossRef]
11. Turrell, A. Agent-based models: Understanding the economy from the bottom up. *Bank Engl. Q. Bull.* **2016**, *Q4*, 173–188.
12. Brusatin, S.; Padoan, T.; Coletta, A.; Delli Gatti, D.; Glielmo, A. Simulating the economic impact of rationality through reinforcement learning and agent-based modelling. In Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF), Brooklyn, NY, USA, 14–17 November 2024; pp. 159–167.
13. Dicks, M.; Paskaramoorthy, A.; Gebbie, T. Many learning agents interacting with an agent-based market model. *arXiv* **2023**, arXiv:2303.07393. [CrossRef]
14. Künsch, H.R. The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **1989**, *17*, 1217–1241. [CrossRef]
15. Liu, R.Y.; Singh, K. Moving blocks jackknife and bootstrap capture weak convergence. In *Exploring the Limits of Bootstrap*; LePage, R., Billard, L., Eds.; Wiley: New York, NY, USA, 1992; pp. 225–248.
16. Politis, D.N.; White, H. Automatic block-length selection for the dependent bootstrap. *Econom. Rev.* **2004**, *23*, 53–70. [CrossRef]
17. Schelling, T.C. Dynamic models of segregation. *J. Math. Sociol.* **1971**, *1*, 143–186. [CrossRef]
18. Axelrod, R. The evolution of strategies in the iterated prisoner's dilemma. In *Genetic Algorithms and Simulated Annealing*; Morgan Kaufmann: San Mateo, CA, USA, 1987; pp. 32–41.
19. Deissenberg, C.; van der Hoog, S.; Dawid, H. EURACE: A massively parallel agent-based model of the European economy. *Appl. Math. Comput.* **2008**, *204*, 541–552. [CrossRef]
20. Dosi, G.; Fagiolo, G.; Roventini, A. Schumpeter meeting Keynes: A policy-friendly model of endogenous growth and business cycles. *J. Econ. Dyn. Control* **2010**, *34*, 1748–1767. [CrossRef]
21. Fagiolo, G.; Roventini, A. Macroeconomic policy in DSGE and agent-based models redux: New developments and challenges ahead. *JASSS* **2017**, *20*, 1. [CrossRef]
22. Yao, Z.; Li, Z.; Thomas, M.; Florescu, I. Reinforcement learning in agent-based market simulation: Unveiling realistic stylized facts and behavior. *arXiv* **2024**, arXiv:2403.19781. [CrossRef]
23. Harder, N.; Miskiw, K.K.; Khanra, M.; Maurer, F.; Patil, P.; Qussous, R.; Weinhardt, C.; Klobasa, M.; Ragwitz, M.; Weidlich, A. ASSUME: An agent-based simulation framework for exploring electricity market dynamics with reinforcement learning. *SoftwareX* **2025**, *30*, 102176. [CrossRef]
24. Mi, Q.; Xia, S.; Song, Y.; Zhang, H.; Zhu, S.; Wang, J. TaxAI: A dynamic economic simulator and benchmark for multi-agent reinforcement learning. In Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Auckland, New Zealand, 6–10 May 2024.
25. Friedman, D. Evolutionary games in economics. *Econometrica* **1991**, *59*, 637–666. [CrossRef]

26. Weibull, J.W. *Evolutionary Game Theory*; MIT Press: Cambridge, MA, USA, 1995.
27. Hommes, C.H. Heterogeneous agent models in economics and finance. In *Handbook of Computational Economics*; Tesfatsion, L., Judd, K.L., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Volume 2, pp. 1109–1186.
28. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438. [CrossRef]
29. Ancona, N.; Marinazzo, D.; Stramaglia, S. Radial basis function approach to nonlinear Granger causality of time series. *Phys. Rev. E* **2004**, *70*, 056221. [CrossRef]
30. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464. [CrossRef]
31. Pearl, J. *Causality: Models, Reasoning and Inference*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2009.
32. Jordà, O. Estimation and inference of impulse responses by local projections. *Am. Econ. Rev.* **2005**, *95*, 161–182. [CrossRef]
33. Nash, J.F. Non-cooperative games. *Ann. Math.* **1951**, *54*, 286–295. [CrossRef]
34. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2018.
35. Simon, H.A. *Models of Man: Social and Rational*; Wiley: New York, NY, USA, 1957.
36. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347. [CrossRef]
37. Axtell, R.L. Zipf distribution of U.S. firm sizes. *Science* **2001**, *293*, 1818–1820. [CrossRef]
38. Politis, D.N.; Romano, J.P. The stationary bootstrap. *J. Am. Stat. Assoc.* **1994**, *89*, 1303–1313. [CrossRef]
39. Politis, D.N.; Romano, J.P.; Wolf, M. *Subsampling*; Springer: New York, NY, USA, 1999.
40. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1861–1870.
41. Lowe, R.; Wu, Y.I.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2017; Volume 30.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.