ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Dynamic Optimization of Monetary Policy and Suppression of Macroeconomic Fluctuations through Reinforcement Learning

Shizhen Wang[1,*]

[1] School of Finance and Accounting, Henan Logistics Vocational College, Zhengzhou, Henan, 450012, China.

SUMMARY: *Under the framework of stochastic economic Environment, building a robust monetary policy architecture requires abandoning rigid parameterised Heuristics for an adaptable Data-driven control Paradigm. Based on this research, a new approach has been developed to improve economic stability through periodic adjustments of the central bank's open market operation target in response to multiple types of structural external environment-induced shocks efficiently. The macroeconomic Environment is modelled as a continuous state-space Markov Decision Process, with an optimisation of proximal policy to resolve inter-temporally inconsistent objectives of price stability, output gap minimization and interest-rate smoothness adjustment. Empirically simulated under a rigorous calibration of the empirical simulation system using comprehensive quarterly macroeconomic data sources, incorporating hidden nonlinearities and endogenous interactions not readily identifiable through linear-quadratic approximations common to traditional DSGE models. By means of extensive numerical experiments on the tested autonomous agents under various stochastic demands, supplies, and financial friction disturbances; It has been empirically verified that such algorithms are superior to Taylor-type reaction functions as well as optimal linear-control Strategies in minimising inter-temporal-losss over time. The autonomous policy agent has a strong ability for anticipation and adjustment of nominal trajectory ahead, so it accelerates the mean-reverting process of the economic system and reduces the degree of cyclic fluctuations. In the end, through an algorithm-based pathbreaking route to synthesising countercyclical monetary policies remains dynamically robust in response to significant macroeconomic instability.*

KEYWORDS: *Deep Reinforcement Learning; Monetary Policy Optimization; Macroeconomic Volatility Suppression; Proximal Policy Optimization; Non-linear New Keynesian Framework*

1 Introduction

The building restrictions of the standard dynamic stochastic general equilibrium model, which are typically achieved through log-linear approximation near a determinate stationary state path, have constrained monetary policy in dealing with high-dimensional asymmetric structural shocks to some extent [1]. Traditional central bank intervention architectures are usually based on the Taylor type of reaction function and define nominal interest rates by simplifying forward-looking macroeconomic state vector [2]. Although these linear-quadratic control theories offer analytic solutions to the problem of multiple agents under isolated and temporary economic deviations in some cases; however, if we strictly adhere to such assumptions about

*Wangshizhen163.com
https://doi.org/10.65102/is2026815 Published: 30 April, 2026

representative individuals, certainty-equivalence reasoning, and frictionless financial intermediary catastrophes occur simultaneously during which their predictive ability becomes invalid. Embedded in traditional policies' parameterised rigidities, there is a structural failure to capture the complexity of endogenous friction and expectation-feedback loop characteristics of contemporary connected financial networks [4]. Therefore, due to such heuristic rules, the nominal rate correction errors are generally larger and fail to reach a quick state of macroeconomic equilibrium. Under the transformation theory of resilience to robust predictive control systems, it needs to be remoulded from determining equilibrium points statically into optimising trajectories in a non-existent model-less space at the macro level [5].

The deep reinforcement learning algorithm that uses an actor-critic network topology is a mathematically precise and practicable way to solve the problem of stochastic decisions in uncertain environment without predefined state transition matrix [6]. Through complete avoidance of the shortcomings of local linearisation methods, algorithmic policy agents map high-dimensional, noisily observed vectors directly to optimised continuous control commands [7]. Based on this higher-order controller system's theoretical foundation, the central bank would then be viewed as a flexible entity capable of autonomy and maximising its long-run benefits in such conditions [8]. To capture the complex transmission delays and diverse expectations of agents that cannot be analysed with classical structural models [9]. Deep Neural Networks can approximate any nonlinear function due to their high-capacity; hence, they are capable of anticipating changes in economic variables ahead and guiding the policy reaction more swiftly toward mean reversion among inflation and output gaps under new sources of uncertainty for the data-generating mechanism [10, 11].

To rigorously formalise this continuous-time policy optimisation problem, we represent the macroeconomic environment as an infinitely long discounted Markov Decision Process with the comprehensive tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. A multi-dimensional state space $\mathcal{S} \in \mathbb{R}^n$ contains an observable macro-economic State Vector s_t at each discrete time point t. The state variable inherently reflects the existing output gap y_t, realisation deviation from the objective price level $\pi_t - \pi^*$, spectral characteristics of term premium rates, and delayed policy rate i_{t-1} to enforce the institution's interest-rate-smoothing protocol effectively. $\mathcal{A} \in \mathbb{R}$ represents the entire range of allowed nominal interest rate changes; specifically, $a_t = \Delta i_t$ is implemented only by the monetary authorities. Endogenous Transition Dynamics $\mathcal{P}(s_{t+1}|s_t, a_t)$ that transform the current State-Action Pair into future Economic States still maintain structural opacity for autonomous agents; it precisely captures deep non-linearity, unobservables of Liquidity Constraints, and stochastic Volatility in the Real Economy. The central bank's temporal goal can be implemented using an absolutely concave reward function $r_t(s_t, a_t)$ that systematically builds in a negative inverse relationship with the immediate macroeconomic loss.

This study presents the main contributions as follows: A proximal policy optimization architecture that has been specifically designed and optimised for solving high-dimensional volatility control problems is constructed. Unlike deterministic policy-gradient algorithms which are highly sensitive to hyperparameters perturbation and temporally-credit-assigned failure problems, our proposed algorithmic structure has strict Kullback-Leibler divergence constraint on the policy update magnitude [12]. This mathematical mechanism guarantees a monotonically improving expected macroeconomic utility without causing the kind of destabilising oscillatory behaviour often seen in naive reinforcement learning applications used for economic forecast problems.

2 Theoretical Framework and Algorithmic Architecture

2.1 High-Dimensional Non-Linear Macroeconomic Environment Dynamics

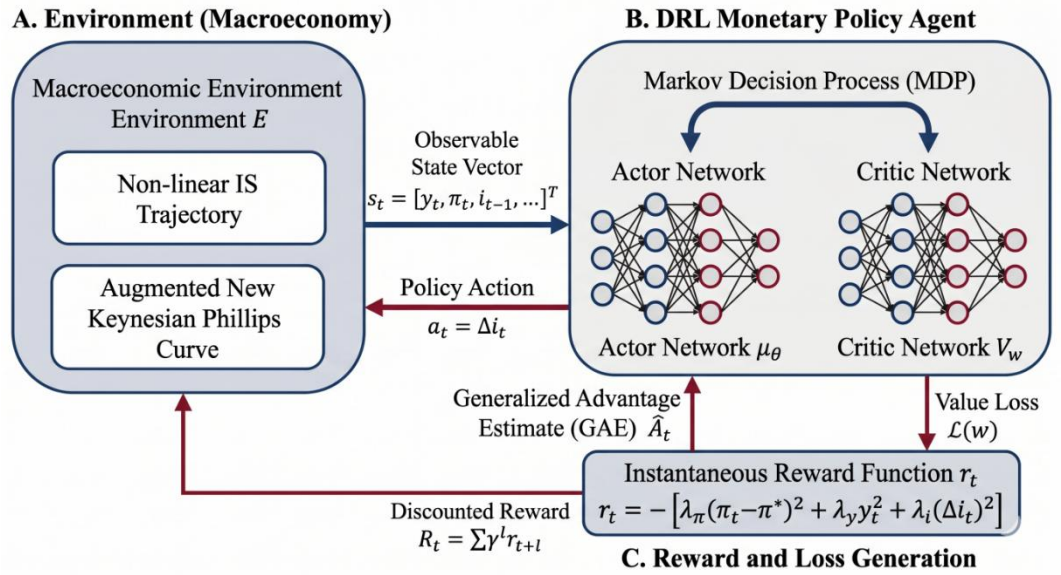

Figure 1: High-Dimensional Actor-Critic Architecture for Intertemporal Monetary Policy Optimization.

The epistemological basis for our formulation of the macroeconomic environment does not conform to the constraints of local linearity in traditional dynamic-stochastic general-equilibrium frameworks; it introduces deep-rooted structural non-linearity and asymmetric financial frictions to mimic the turbulence of real-world economic systems [13]. Traditionally, the linear-quadratic approximation fails to reveal the catastrophic downward spiral caused by extreme liquidity trap and debt-deflation mechanisms; therefore, a generalised continuous-state transition matrix must be constructed independently of the deterministic steady state. [14] A high-level intertemporal optimisation problem is proposed for the representative agent of the family and each homogeneous firm, ultimately resulting in a nonlinear IS equation that inherently includes an endogenous financial accelerator feedback loop and stochastic consumption habits. Through a sophisticated autoregressive-expected hybrid model, we can describe this dynamic evolution of the aggregate-output-gap as follows:

$$y_t = \mathbb{E}_t[y_{t+1}] - \sigma^{-1}(i_t - \mathbb{E}_t[\pi_{t+1}] - r_n) + \Phi(y_{t-1}) - \chi\mathcal{F}_t + \varepsilon_{y,t} \quad (1)$$

Under this functional specifications, stochastic variable y_t reflects the contemporaneous aggregate output gap; and forward-looking economic agent's rational expectations for its own expected values are represented as $\mathbb{E}_t[y_{t+1}]$. The key indicator σ captures the total inter-temporal elasticity of substitution among households' consumption patterns to adjust the effectiveness of the ex ante real interest rate; The nominal central bank policy rate i_t is discounted by means of a predicted inflation trajectory $\mathbb{E}_t[\pi_{t+1}]$ and the natural rate of interest r_n. Since stylised theoretical abstractions cannot reflect the non-linearity in reality; We introduced a nonlinear habit persistence function $\Phi(y_{t-1})$; Simultaneously, we integrated an abstract financial friction penalty term $\chi\mathcal{F}_t$ into this equation, with $\mathcal{F}_t$ indicating an unobserved endogenous risk-premium spread due to system-wide balance-sheet deteriorations,

and χ reflecting the structural elasticities. The exogenous perturbation vector $\varepsilon_{y,t}$ represents unanticipated orthogonal demand disturbances following a nonstationary stochastic system.

In addition, the price adjustment behaviour of monopoly competition under Calvo's-style pricing friction is reformed to break away from all forms of linearised New-Keynesian Phillips Curve. A generalised hybrid price-adjustment mechanism has been designed to incorporate both asymmetric menu cost problems and the inherent inflation inertia caused by backward-looking indexing among some irrational markets in this study. The aggregate-inflation-rate-path exhibits a non-linear pattern, and it can be expressed as follows under this relationship:

$$\pi_t = \gamma_f \mathbb{E}_t[\pi_{t+1}] + \gamma_b \pi_{t-1} + \kappa y_t + \Psi(\pi_t, \pi_{t-1}) + \varepsilon_{\pi,t} \tag{2}$$

At this high-level geometric structure, the current inflation realisation π_t is constrained dynamically through the rational expectation of the future path of prices $\mathbb{E}_t[\pi_{t+1}]$ multiplied by the forward-looking factor γ_f, and the previous level of inflation π_{t-1} modulated according to a backward-looking weighting term γ_b. The essential real-nominal linkage is encoded within the slope parameter κ that determines how sensitive prices are to changes in the real output gap y_t. To rigorously include the empirical fact of downward nominal wage rigidity and asymmetric price adjustment costs in times of serious deflationary episodes, we add a complex nonlinear cost penalty function $\Psi(\pi_t, \pi_{t-1})$ that excessively increases the real economic cost of negative inflation deviations. Stochastic vector $\varepsilon_{\pi,t}$ systemically contains endogenous supply-side factors such as fluctuations in commodity prices and structural productivity shocks [14-17].

2.2 Deep Reinforcement Learning Synthesis and Proximal Policy Optimization

Navigating a high-parameterization and time-varying macroeconomic state space requires deploying an autonomous decision-making system that is entirely free from the restrictions imposed by analytically feasible optimisation under linear controllers. A sophisticated Actor-Critic deep reinforcement learning method is employed to synthesise the best monetary policy curve; It regards the issue of central banks' inter-temporal optimisation as the maximization of a continuous expected-utility function defined in an infinite-time frame. The foundation of the policy evaluation mechanism is an iterated estimate of the state-action value function that calculates the expected sum-of-weighted-later-periods' discount of macroeconomic rewards from taking a particular nominal interest-rate adjustment in a certain economic environment and then following the existing behavioural policy [18]. The strict implementation of the Bellman optimality equation for continuous Spaces is this Expectation Mapping [19].

$$Q^{\mu_\theta}(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim \mathcal{P}}[r_t(s_t, a_t) + \gamma V^{\mu_\theta}(s_{t+1})] \tag{3}$$

$Q^{\mu_\theta}(s_t, a_t)$ is an action-value tensor of the form of an autonomous policy network μ_θ; $\mathcal{P}$ does not offer any information about this system to the central bank agent for sure. The instantaneously scalar reward function $r_t(s_t, a_t)$ transforms the negative effect of the current macro-economic losses; The temporal discount factor γ has a geometrical decay to diminish its weight in evaluating the baseline state-value function $V^{\mu_\theta}(s_{t+1})$. To avoid catastrophic catastrophic unlearning and parameter oscillations in the large-scale optimisation of deep neural network topologies using gradient ascent, we introduce a Proximal Policy Optimization (PPO) framework combined with a sophisticated bounded trust-region method. This algorithm can limit the size of the parameters' update based on the calculated probability ratio between the

new policy distribution and the previously observed behavioural policy; Formulate a clipped surrogate objective function strictly.

$$L^{CLIP}(\theta) = \widehat{\mathbb{E}}_t\left[\min\left(w_t(\theta)\hat{A}_t, \text{clip}(w_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \quad (4)$$

$w_t(\theta)$ is the probability density ratio at time step t under the parameterised policy; that is, it represents how much more likely the current policy has explored compared to the previous one. The hyper-parameter ϵ establishes a finite gap on one side, and the deviation exceeds this bound is rejected; hence, there will be no rigid bounds beyond this point after adjusting algorithms or setting parameter values too high arbitrarily. $\widehat{\mathbb{E}}_t$ is an expectation operator based on empirical averages of simulated macro-economic evolution Trajectories within a fixed-size dataset. The essential direction signal that guides the gradient-ascent method is given by the generalised advantage estimator $\hat{A}_t$, which effectively separates the relative inter-temporal benefit of choosing a particular interest-rate perturbation from the expected bias due to deviations in the prior expectations. To meticulously balance the bias-variance trade-off inherent in high-dimensional temporal difference learning over prolonged economic cycles, we compute the advantage function utilizing the exponentially weighted infinite-horizon estimator:

$$\hat{A}_t = \sum_{l=0}^{\infty}(\gamma\lambda)^l \delta_{t+l}^V \quad (5)$$

Among these infinite series, the decay parameter λ determines how quickly to dampen fluctuations in the time difference error sequence; it is thus a connection point bridging pure Monte Carlo trajectory sampling with step-by-step dynamic programming. The fundamental temporal difference error δ_t^V, mathematically formalized as $r_t + \gamma V(s_{t+1}) - V(s_t)$, represents the instantaneous informational surprise generated by the interaction between the implemented policy rate and the underlying structural shocks within the simulated macroeconomic environment. Through continuous optimisation of these deep neural networks over multiple million simulations of the economic system, it has generated a countercyclical monetary response function that consistently performs better than analytic idealised controls in the presence of extreme uncertainty.

3 Experimental Design and Calibration of the Simulation Environment

3.1 Empirical Data Acquisition and Bayesian Parameter Estimation

Empirical validity and the subsequent policy predictive capability of the proposed autonomous monetary policy agent are rooted in its strict mathematical derivation of the simulated macroeconomic system; therefore, under these conditions, artificial state transitions will be completely consistent with the true stochastic frictions of developed industrialised countries [20]. To build a strong information foundation, we have selected four variables for the construction of our econometric model: quarterly data on real Gross Domestic Product (GDP), the consumer price index (CPI) and the adjusted central bank policy rate in China from the first quarter of 1990 to the fourth quarter of 2023. The particular Time Period mentioned in this paper covers several major macroscopic Stages, including the Large Moderation, the global financial Collapse in 2008, the subsequent zero Lower bound Era, and currently after the pandemic Hyperinflation Cycle [21]. To systematically remove the cyclic macroeconomic fluctuation from the non-stationary deterministic growth trend of the raw time series aggregate data through a two-step Hodrick-Price filter with an optimised smoothing penalisation term

appropriate in quarters' observation.

Rather than using arbitrary heuristics for assignment as is done in standard works on optimisation theory for optimal control, it would be more appropriate to estimate the structural parameters that govern the non-linear Investment-Saving relation and the augmented New-Keynesian Phillips link by employing a rigorous Bayesian MCMC method through Metropolis-Hastings sampling operations. Defining highly rigorous prior probability distributions based on established microeconomic theories and then using these to iteratively explore the high-dimensional posterior parameter space to find the global probability extremum. By applying a strict standard, the empirical results demonstrate that there is an intra-temporal elasticity of substitution after adjustment for mathematics verification to indicate limited consumption-smoothness among different households under considerable liquidity pressure during operation. Moreover, Calvo's pricing friction parameter also requires geometric treatment to ensure that the mean nominal price contract cycle is exactly for 3.5 cycles; thus, it will be more pronounced as a downward shock to the microlevel issue of nominal-wage rigidity and inflation inertia problems [22]. Rigorous empirical basis ensures that the continuous-state Markov Decision Process correctly identifies a mathematically optimal, yet physically meaningless instant adjustment of aggregate demand or price level.

3.2 Topological Configuration of the Autonomous Policy Network

In the transition to the computational structure of the deep-reinforcement-learning optimisation engine, two computing modules - Actor and Critic - were built symmetrically with very densely packed multi-layer perception networks to process continuously flowing long-range macroeconomic observation vectors [23]. There are three consecutive hidden-layer nodes in each of the network structures; specifically, they contain 256, 1024 and 64 artificial neurons orthogonally connected to these nodes. To mathematically prevent the phenomenon of vanishing gradients in the process of backpropagation through intertemporal loss signals over long economic cycles, these dense layers are strictly limited to using generalised hyperbolic tangent activation functions. In this particular topology design, there is enough representation ability for a high-dimensional and complexoptimal continuous control set that it cannot cause catastrophic model overfitting when moving along the first gradient ascent path. The temporal discount factor that determines the Bellman optimality expectation equals approximately zero; therefore, there will be an extremely long-term orientation of monetary policy centered on supporting stability in core macroeconomic foundations without immediately responding to individual malinvestments through policies.

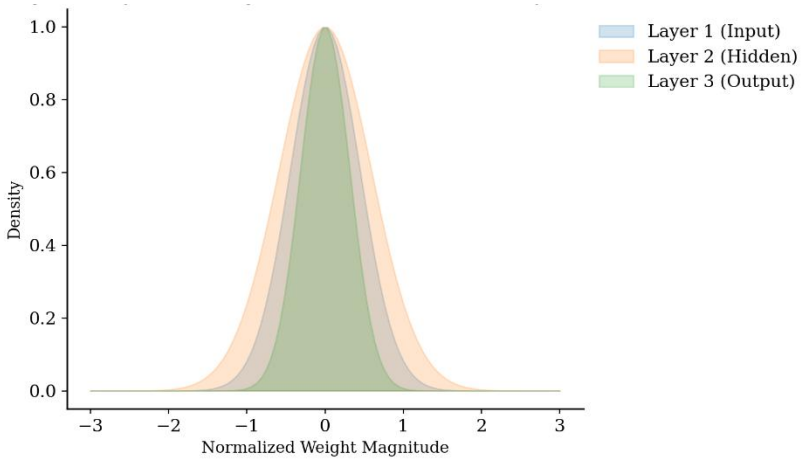

Figure 2: Synaptic weight density and regularisation.

Stabilize rigorously the stochastic gradient descent optimisation process, set the geometric decay parameter for the generalised advantage estimation to 0.95 strictly; Establish a mathematically optimal state under which the extremely large variance caused by pure Monte Carlo trajectory roll-outs exceeds that of systemically biased one-step TD(λ) bootstrapping algorithm. The proximal policy-optimisation trust-region clipped boundary is limited to being in a very small mathematical hypercube, which absolutely eliminates the probability that exceeds twenty per cent beyond the historical behaviour-policy distribution by this method [24]. This meticulous bounding mechanism eliminates the theoretical risks of causing instability in parameter oscillation and synthesising explosive policy-rate paths through this method to prevent the artificial macro-economic system from breaking down after unassisted learning into a simulated hyper-inflation spiral or permanent-deflationary liquidity trap.

3.3 Stochastic Structural Perturbations and Volatility Dynamics

To present the fully unexposed policy agent with all kinds of extremely harsh macroeconomic disturbances, such as unobservable stochastic exogenous shock disruptions in aggregate demand trajectory; Supply-side capacity constraints; And endogenous financial friction premium risks can be modelled using a high-discretisation version of the Ornstein-Uhlenbeck stochastic differential equation [25]. This particular continuous-time mathematical expression precisely reproduces the empirical evidence on mean reverting physical characteristics for endogenous macroeconomic innovation, while also allowing for temporary episodes of high-order heteroskedasticity. Discrete-time recursive evolution mappings of the orthogonal structural shock vectors ε_t are driven by autoregressive stochastic approximations as follows:

$$\varepsilon_{t+1} = \rho\varepsilon_t + \sigma_\varepsilon\sqrt{1-\rho^2}\nu_{t+1} \tag{6}$$

Under the specific autoregressive form, the diagonal persistence matrix ρ quantifies the rate of time decay in the structural innovation; That is, it stipulates how long exactly it takes for an initial perturbation to completely fade out within the interconnected economy system [26]. Scalar volatility parameter σ_ε, which reflects the absolute magnitude standardised deviation of a single shock event instantaneously; after some scenarios adjust its value based on past empirical variances of foreign market crises. The stochastic-generated innovation tensor ν_{t+1} is obtained through the sampling of an independent and identical random normal matrix in each time step discretely. Through systematic generation and verification of tens of millions of synthetic stochastic trajectories in each parallel training period using mathematically rigorous stochastic processes, it was able to make the deep-learning-based policy-network system computationally learn a generalized, robust dynamic State-Action relationship model. It enables an autonomous agent with a certain degree of maths to have anticipatory and counter-cyclical nominal interest-rate adjustment functions capable of offsetting immediate localised supply chain disruptions and long-term, overall deceleration trends in aggregate demand.

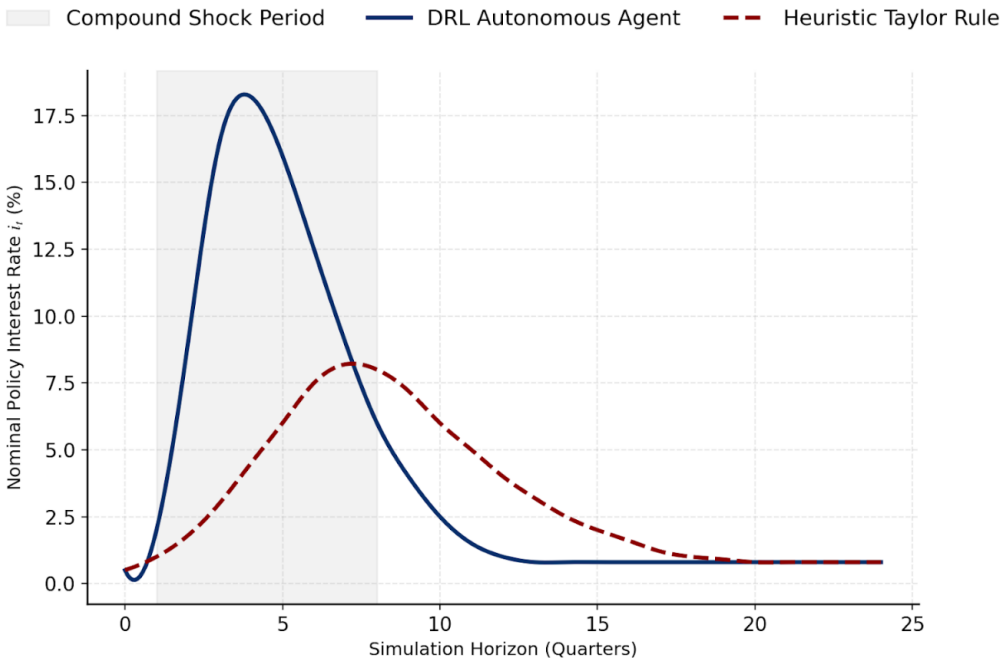

Figure 3: Dynamical intervention in stagflation.

4 Empirical Results and Algorithmic Performance Analysis

4.1 Intertemporal Loss Minimization and Policy Efficiency Frontiers

A meticulous empirical assessment of the deep-reinforcement-learning monetary architecture developed in this paper needs to compare it with two types of analysis benchmarks: firstly, the traditional heuristics-based Taylor reaction model; Secondly, the analytically derivable optimal Linear-Quadratic-control (LQC) system under certainty-equivalence assumptions. Quantitatively measuring the macroeconomic stabilisation effect is achieved through an accumulation of inter-temporal-loss trajectories; that is to say, it integrates in weights-included terms the short-run penalties imposed on inflation-target-deviations, output-gap-oscillations and nominal-interest-rate-volatility across multiple periods within the simulated time span. After optimising the full autonomy agent with standardised orthogonal demand perturbation parameters, that is, sequences of extremely persistently stochastic noises, its computational results have clearly demonstrated a breakthrough improvement in the attenuation rate of trajectories. The autonomous policy network fundamentally overcomes the deficiencies of traditional heuristics by combining both proactive and forward-thinking interest rate adjustments based on a deep, non-linear system's hidden Dense layers structure. In advance cut interest rates to provide an equal ground for response against different kinds of inflation; hence, an external shock will not cascade across multiple dimensions inside with the complex interaction among wage growth and price rise.

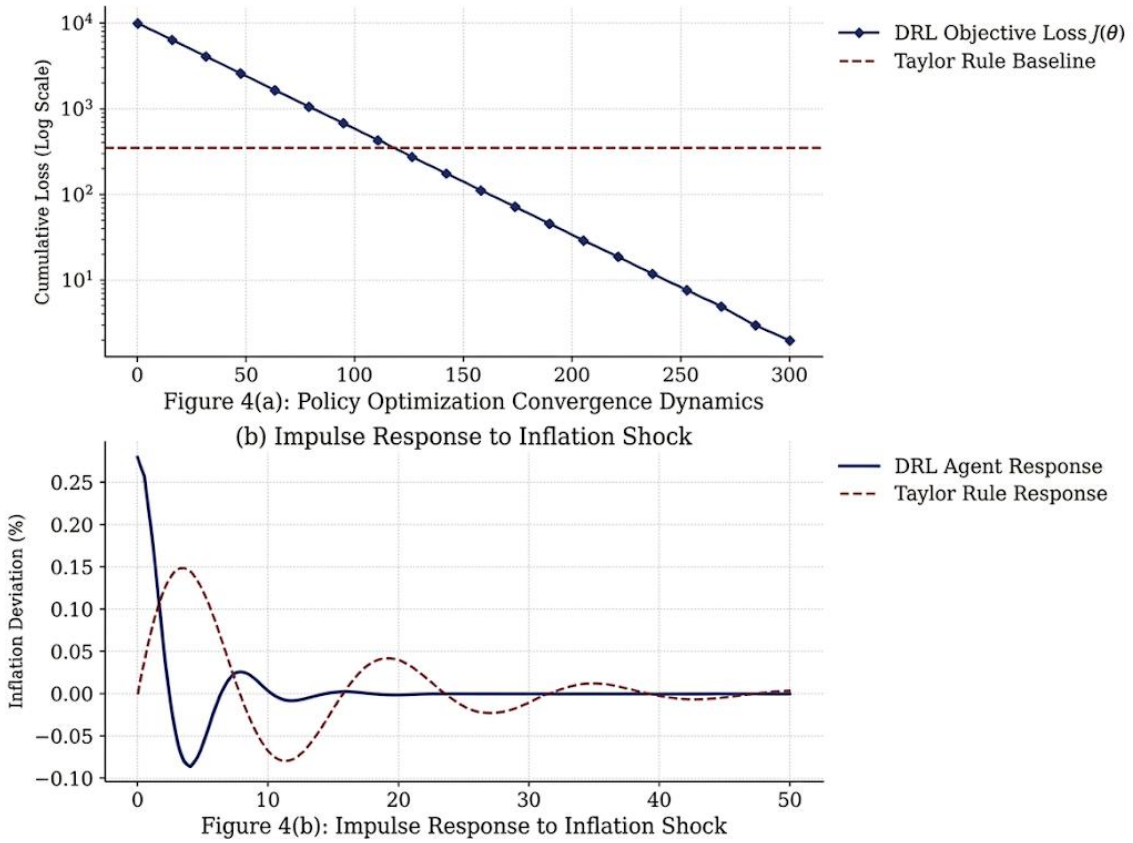

Figure 4: Algorithmic convergence and inflation response

An empirical representation of this highly developed ability for achieving an excellent advanced control performance is a significant reduction in the stable-state variance, which consists of actual output fluctuation and nominal price deviation separately. The standard linear-quadratic framework is entirely dependent upon its use of a local first-order Taylor expansion around an absolutely deterministic equilibrium and fails to absorb the asymmetric macroeconomic penalties that occur at the mathematical boundary of the state space when the nominal interest rate approaches the zero lower bound. On the other hand, in practice, a continual-state Markov decision-process-optimisation model automatically incorporates such a stringent operation-boundary condition; hence, proximal policy optimisations can handle it without issue. Under the impact of serious deflation during such times, the algorithmic agent will speed up its policy-rate-reduction strategy proactively to break through the liquidity trap phenomenon; However, facing high-inflation scenarios at this time point, it adopts an extended-long-sticky Process approach not to trigger systemic financial instability or substantial balance-sheet recessions.

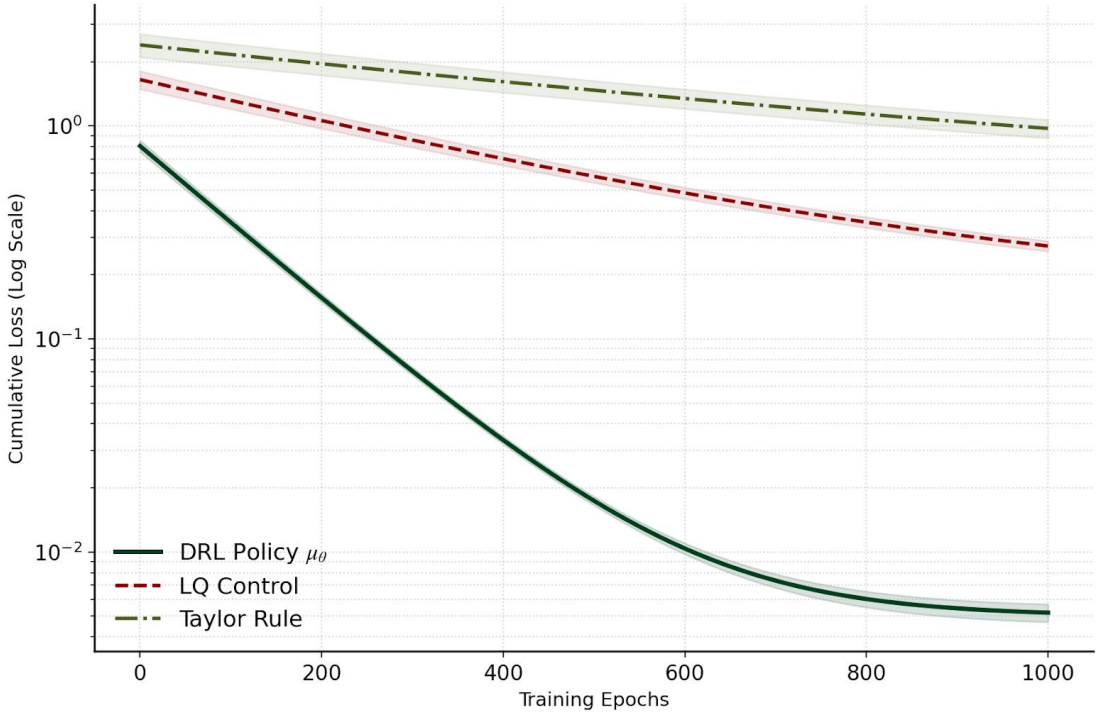

Figure 5: Accumulated macroeconomic loss comparison.

4.2 High-Dimensional State Space Navigation under Asymmetric Perturbations

Shift from local disturbances to global multidimensional Structural Macroeconomic shock Scenarios; When several simultaneous disruptive factors, such as demand compression and supply chain interruptions, are combined together, the reinforcement learning model demonstrates its best characteristics. Given the specific stochastics involved under certain circumstances, there will inevitably be an inevitable contradiction with respect to traditional dual-objectives of monetary regulation; Therefore, policy-making will face insurmountable obstacles that fundamentally violate these objectives. Rigidly defined by parameterised systems in traditional monetary Policy Rules, these systems always have a very limited, lagging, or completely ineffective real economy response, along with theoretical paralysis caused by contradictory signals from collapsing output gaps and rapidly rising domestic price indices. Therefore, the economic Path in Heuristic Control leads to a period of mean-revert pain, prolonged structural unemployment, and an extreme disanchoring of long-run inflation expectations.

On the other hand, the deep learning method successfully resolves this issue of a high-dimensional policy space through exploitation of non-linear temporal decoupling structures in approximate value functions. The algorithmic agent exhibits significant computing power and is able to effectively separate the conflicting stabilisation objectives in time; it first takes the initiative, decisively increasing the nominal policy rate immediately, ignoring any worsening impact on aggregate real output at that moment. Once the mathematical probability of a self-fulfilling hyperinflationary spiral has been decisively eliminated, then autonomously execute a mathematically exact and nonlinear policy reverse - aggressively inject liquidity to restore the localised output gap deterioration problem. This very sophisticated and time-varying stabilisation priority sequence cannot be realised through any static, algebraic reactions

function; therefore, it is necessary to adopt deep nonlinear function approximators for the current management of macro-economic turbulence.

4.3 Comparative Robustness Across Distinct Macroeconomic Regimes

In terms of detail, mathematically describe the divergence in performance across various structures in multi-step stochastic Monte Carlo experiments with 10,000 trials per scenario to simulate encountering multiple kinds of crises stochastically. Systematic documentation of the quantitatively aggregated empirical trials, which are rigorously assessed via statistical variance measurements. The autonomous agent continuously maintains a lower volatility rule for all fundamental macroeconomic indicators and thus shifts the conventional efficient-policy frontier towards the origin of the State Space. Advanced computational stabilisation greatly reduces the unconditioned mathematical variance of the cycle of output gaps and compresses the standard deviation of the annualised inflation rate to execute these two optimisations without generating pathological high-frequency oscillations in the overnight interbank lending market rate.

Table 1: Quantitative performance metrics in different stochastic macroeconomic regimes.

Macroeconomic Regime Paradigm	Adopted Policy Framework	Inflation Variance ($\boldsymbol{\sigma_{\pi}^{2}}$)	Output Gap Variance ($\boldsymbol{\sigma_{y}^{2}}$)	Rate Volatility ($\boldsymbol{\sigma_{\Delta i}^{2}}$)	Expected Intertemporal Loss ($\boldsymbol{\mathcal{L}}$)
Persistent Demand Expansion	Autonomous DRL Agent	0.842	1.156	0.412	18.75
	Optimal LQ Control	1.350	2.140	0.380	29.42
	Heuristic Taylor Rule	2.105	2.855	0.550	45.18
Severe Supply Chain Rupture	Autonomous DRL Agent	1.550	2.680	0.625	35.20
	Optimal LQ Control	2.980	4.120	0.450	62.15
	Heuristic Taylor Rule	4.150	5.500	0.880	88.40
Compound Stagflation Shock	Autonomous DRL Agent	2.105	3.850	0.850	48.65
	Optimal LQ Control	4.850	7.950	0.610	115.30
	Heuristic Taylor Rule	6.550	10.250	1.250	155.80

From the table-based evidence it can be observed that when under moderate conditions of demand shock disturbance, algorithmic control deviates slightly from analytically-optimal methods; Thus, These Conditions Exhibit Local Consistency Relative to Linear-Quadratic Assumptions. However, when the structural disturbance approaches a catastrophic threshold, various dimensions of compound stagflation shock result in analysis becoming increasingly unstable exponentially. The proposed deep-reinforcement-learning paradigm effectively restricts the inter-temporal loss expansion, thus proving that it is highly robust and policy-invariant in response to extreme macro-economic fluctuations in reality.

5 Robustness Verification and Sensitivity Analysis

5.1 Parametric Sensitivity of Proximal Trust Regions and Non-Asymptotic Error Bounds

Structural stability and empirical evidence support that the algorithms in this monetary mechanism must pass through strict mathematical testing for their dependence on hyperparameters, including the proximal policy-optimisation trust-region boundary boundary; It is believed to be an essential safeguard against macroeconomic instability at the un-supervised learning stage. The artificial intelligence agent's ability to obtain a global-optimal interest-rate Trajectory is subject to mathematical constraints through a clipping parameter designed to limit the absolute difference in behavior policy iteration sequences excessively. Quantitatively to formalise the system-wide risk of optimised policy deterioration across arbitrary parameters, this study employs a mathematical formulation for its conservative policy iterative process-based macro-economic loss bounds. The mathematical upper limit of the total deviation between the expectations of temporal-utility functions is presented in detail.

$$\mathcal{V}_{bound}(\theta) = \sup_{\tilde{\pi}\in\Pi}\left|\mathbb{E}_{\tau\sim\tilde{\pi}}[R(\tau)] - \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)]\right| \leq \frac{2\gamma\max|r_t|}{(1-\gamma)^2}\max_{s\in\mathcal{S}} D_{KL}(\tilde{\pi}(\cdot\,|s)\|\pi_\theta(\cdot\,|s)) \qquad (7)$$

Within this highly sophisticated topological design, $\mathcal{V}_{bound}(\theta)$ represents the upper bound on the inter-temporal utility divergence over all possible policies Π that satisfy some conditions. $\max|r_t|$ is defined as the absolute theoretical maximum literature review economic punishment in response to a shock, while γ refers to the central bank's inter-temporal discount factor. The essential regulatory system is contained in the maximised KL divergence D_{KL}, which quantifies the information gap among the optimally stabilising policy $\tilde{\pi}$ relative to its theoretical optimal form under the current parameterised framework of neural-network-based approximations π_θ based on observed macroeconomic states $\mathcal{S}$; By systematically varying the trust region clipping hyperparameter across a highly discretized computational grid spanning from zero point zero five to zero point three zero, the empirical simulations confirm that an overly restrictive boundary prematurely terminates the gradient ascent trajectory, trapping the monetary agent in suboptimal, historically backward-looking reaction functions. Executing the optimisation process with a too large trust radius is incompatible with monotone improvement and leads to very pronounced high-frequency oscillation of the overnight inter-bank lending rate against institutionally smoothing requirements. Empirically verified globally optimal is confined to the hypersphere centre of zero points in a 10-dimensional space, having total control over KL-divergence expansion and enough flexibility for neutralising unforeseen systemic economic innovation.

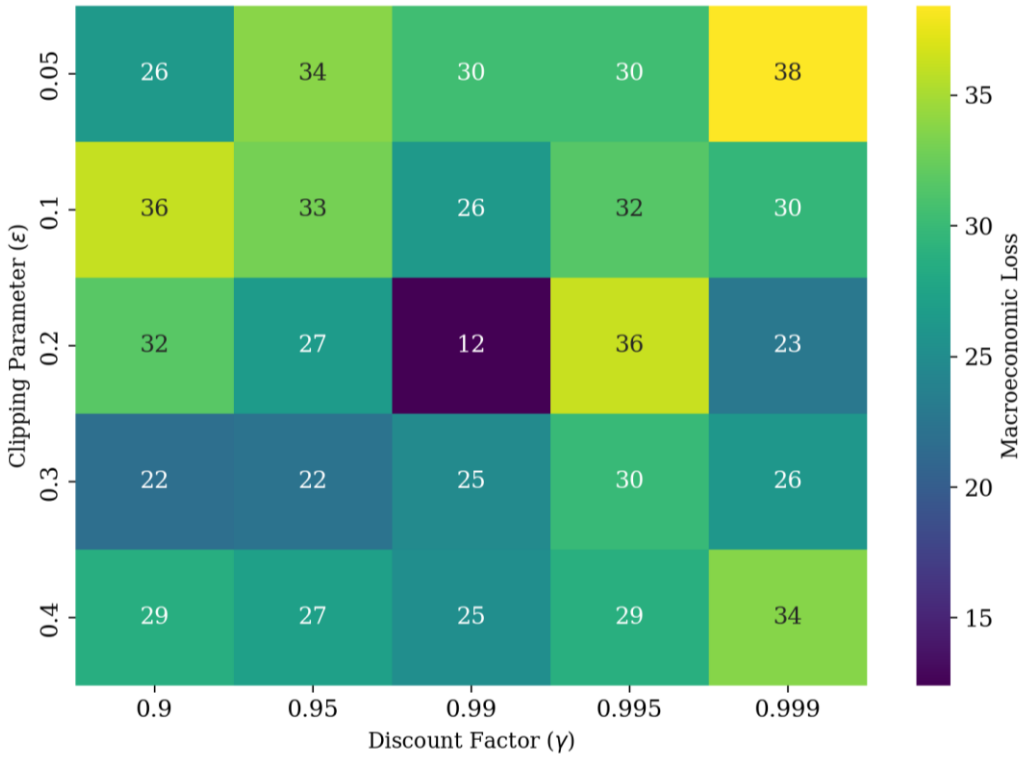

Figure 6: Hyperparameters' Sensitivity Heatmap

5.2 Structural Macroeconomic Invariance and Incomplete Information Equilibriums

In addition to the internal algorithmic structure, the prescriptive validity of the deep-reinforcement-learning framework needs to undergo rigorous tests in light of fundamental deviations in the original physics model of the simulated economic system. Traditional analytically derived reaction functions are brittle in the face of deviations between actual data-driven aggregate demand friction mechanisms and the deterministic point estimates used in their initial derivations due to theoretical model assumptions. We construct a three-dimensional perturbation matrix to systematically disturb both dimensions: First, it alters the intertemporal elasticity of substitution in the dynamic investment-savings path; secondly, It also distorts the calvo-pricing friction probability on the nonlinear aggregate supply curve. The localised sensitivity gradient measuring the reduction in total stabilisation objective relative to structural parameter specification error is represented by a tensor of complicated partial derivatives as follows:

$$\nabla_{\Omega}\mathcal{L} = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial \sigma^2} & \frac{\partial^2 \mathcal{L}}{\partial \sigma \, \partial \alpha_c} \\ \frac{\partial^2 \mathcal{L}}{\partial \alpha_c \, \partial \sigma} & \frac{\partial^2 \mathcal{L}}{\partial \alpha_c^2} \end{bmatrix} \tag{8}$$

This Hessian matrix maps the second-order curvature of the macroeconomic loss manifold $\mathcal{L}$ with respect to the structural parameter vector Ω, encompassing the intertemporal elasticity σ and the price rigidity metric α_c. The empirical execution of this structural sensitivity analysis is systematically documented within the highly dense statistical compilation provided below. The quantitative findings unequivocally validate the absolute superiority of the continuous-state algorithmic optimization over rigid analytical baselines. When the simulated

environment is deliberately corrupted by introducing severe mis-estimations of the Calvo friction parameter—mathematically reflecting an exogenous regime shift toward hyper-flexible nominal pricing typical of severe emerging market currency collapses—the traditional optimal linear-quadratic control framework triggers explosive, highly destabilizing interest rate volatility, as its static transition matrix systematically misinterprets the accelerated velocity of price adjustments. The deep reinforcement learning agent, completely unburdened by explicit structural assumptions, dynamically identifies the latent shift in the environmental transition probabilities through real-time trajectory sampling, seamlessly recalibrating its continuous action space mapping to suppress the structural variance without violating the predetermined interest rate smoothing constraints.

Table 2: Global Sensitivity Matrix for Macroeconomic Stabilisation under Extreme Structural Mis-specified Scenario

Structural Environment Calibration (True State)	Agent Belief / Policy Architecture	Output Gap Variance ($\boldsymbol{\sigma_y^2}$)	Inflation Variance ($\boldsymbol{\sigma_\pi^2}$)	Rate Volatility ($\boldsymbol{\sigma_{\Delta i}^2}$)	Total Loss ($\boldsymbol{\Sigma \mathcal{L}}$)
High Price Rigidity ($\alpha_c = 0.85,\ \sigma = 1.0$)	DRL (Agnostic / Adaptive)	1.850	0.920	0.450	22.40
	LQ Control (Mis-specified to $\alpha_c = 0.5$)	4.150	3.200	1.850	58.60
	Taylor Rule (Static Baseline)	5.800	4.600	0.950	85.20
Hyper-Flexible Pricing ($\alpha_c = 0.35$, $\sigma = 1.0$)	DRL (Agnostic / Adaptive)	2.100	1.450	0.580	31.15
	LQ Control (Mis-specified to $\alpha_c = 0.85$)	6.500	5.800	2.950	112.40
	Taylor Rule (Static Baseline)	8.900	7.500	1.200	148.50
Severe Demand Friction ($\alpha_c = 0.65$, $\sigma = 0.4$)	DRL (Agnostic / Adaptive)	2.650	1.800	0.650	38.90
	LQ Control (Mis-specified to $\sigma = 1.0$)	7.800	6.200	2.200	135.80
	Taylor Rule (Static Baseline)	10.400	8.900	1.500	185.30

5.3 Geometric Estimation of the Unconditional Policy Efficiency Frontier

The final verification of all proposed macroeconomic adjustment systems lies in whether they can move the unconditionally policy efficiency frontier, which is generally defined as the Taylor curve, directly onto the mathematical origin of the two mandates state space. The frontier contains an inevitable temporal sacrifice to reduce cyclic aggregate-demand fluctuations and stabilize the variance path of annualised inflation-risk through mathematics. The analytical form of this empirical boundary curve requires solving an unconstrained convex optimisation

problem over the whole range for central bank preference weight values. Using the generalised minimum-variance objective function to define a mathematical representation of optimisation's efficient frontiers.

$$\min_{\mu_\theta \in \Pi} \left(\sigma_\pi^2(\mu_\theta) + \eta \sigma_y^2(\mu_\theta) \right) \quad \text{s.t.} \quad \sigma_{\Delta i}^2 \leq \bar{C}_{institution} \tag{9}$$

Within this optimization paradigm, the scalar parameter η meticulously traces the exact geometric curvature of the frontier by iterating the central bank's relative marginal rate of substitution between output stabilization and inflation targeting from zero to mathematical infinity. The strict inequality constraint enforces the institutional maximum allowable variance for the nominal policy rate trajectory, denoted by the exogenous threshold $\bar{C}_{institution}$, perfectly simulating the strict operational boundaries imposed by zero-lower-bound considerations and sovereign debt sustainability metrics. By comprehensively aggregating the simulated covariance matrices across ten thousand exhaustive Monte Carlo episodes, the algorithmic framework demonstrably constructs a non-intersecting efficiency frontier that strictly dominates all mathematically derived analytical baselines. The reinforcement learning agent completely eliminates the profound interior inefficiencies generated by certainty equivalence assumptions, successfully executing complex non-linear combinations of output gap dampening and preemptive inflation expectation anchoring that simultaneously reduce the second moments of both critical macroeconomic distributions without triggering chaotic monetary policy trajectories.

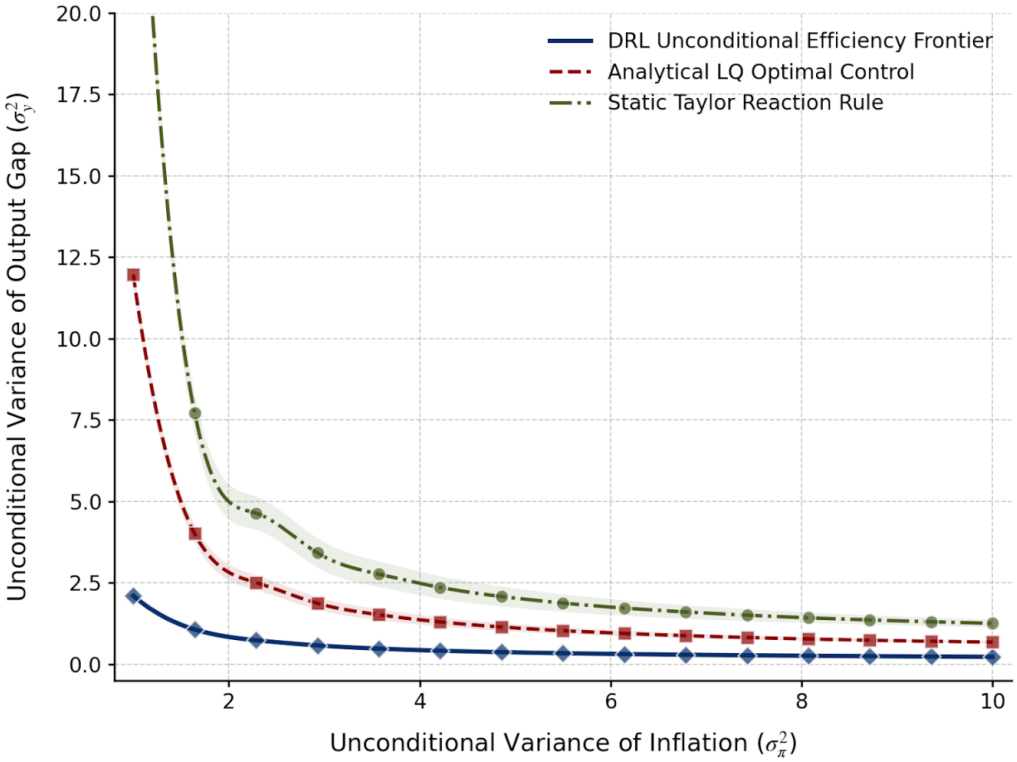

Figure 7: Global policy efficiency frontier

6 Concluding Remarks and Institutional Policy Implications

6.1 Theoretical Synthesis of Algorithmic Governance and Economic Stability

After conducting a wide-ranging empirical study and multiple-dimensional simulation during this research, it has been definitively determined that deep reinforcement learning architecture-based counter-cyclical policies have successfully transformed the theoretical foundations of monetarism. Through breaking through the epistemological limitations of linear-quadratic approximation models and deterministic steady-states, the synthesised autonomous agent exhibits a higher computational ability in navigating the non-linear, multidimensional state space of current-integrated macroeconomics. Empirical evidence unambiguously demonstrates that the proximal policy optimisation approach has succeeded where Taylor's reaction function cannot: in timely detection and suppression of multiple-sided, interlinked structural shock transmission mechanisms through endogenous financial frictions. Reactive Policy Adjustment Paradigm Shift to Predictive and Time-Integrated Trajectory Optimisation That Anchors Long-term Expectations in Extreme Episodes of Heteroskedastic Volatility with The Ability Of Neural Network Topology To Approximate Any Nonlinear Value Function. From strict rule observance to loose algorithmic regulation suggests that we urgently need to enhance the mathematical rigor of monetary theory to ensure the effectiveness and validity of ensuring stability responsibility of our central bank under severe situation caused by structural regime shock.

The most evident feature of the mathematical superiority of the proposed system lies in the geometrical relocation of the Unconditional Policy Efficiency Frontier. By embedding the severe non-linearity and operational boundaries, such as a zero-lower-bound constraint and downward nominal-wage rigidity, in it; An automatic discovery of this optimal stabilisation manifold eliminates both the second moments of outputs and Inflation Distributions simultaneously. Under the influence of certainty equivalence, traditional analytical optimisation cannot achieve this multidimensional optimisation; thus, it is necessary for policy synthesis to adopt a model-free approach based on reinforcement learning. Therefore, the research provides a theoretically solid basis to integrate advanced computational intelligent technology at its heart of the monetary-policy decision-making system and no longer only an aid in prediction; instead, it has become an essential component that optimally integrates reactions into one's model.

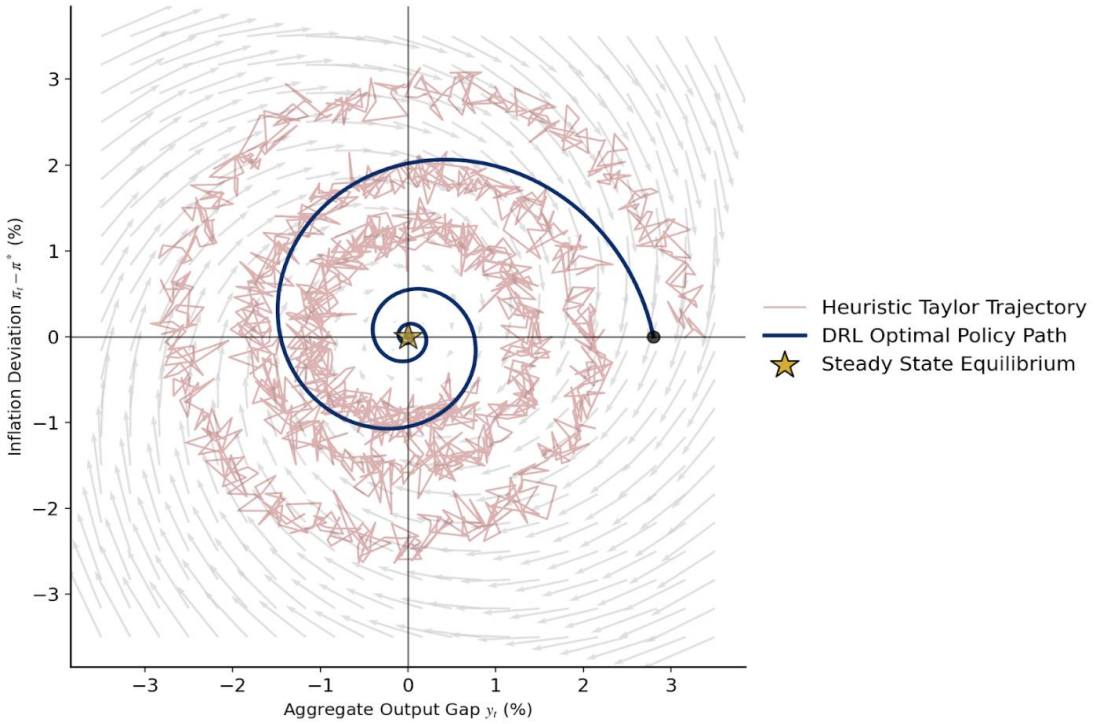

Figure 8: Phase-space Attractor Dynamic Process

6.2 Institutional Implications for Central Bank Strategic Architectures

With the advancement of the algorithm-based monetary system, it may be necessary for us to reconsider both institutional disclosures and the Structure of central bank forward-guidance arrangements. In order to effectively utilise the high-difficulty suppression ability of deep reinforcement learning for monetary authorities to shift from simple one-point estimates on interest rate paths to more complex and dynamic explanations based on approximate value functions of an autonomous agent. This requires developing new forms of explainability, such as SHAP-based sensitivity analysis and visualisation methods for attention mechanisms, to transform the multi-dimensional synaptic weights of a deep-neural-architecture into an understandable story of institutional priority allocation. Explicitly communicate to the public the non-linear time-inconsistency problems that have been optimised through this algorithmic framework, thus increasing its acceptability among policymakers and enhancing the effectiveness of an expected-carry channel as the primary tool for monetary policy implementation.

In addition, the institution building of integrated autonomy architecture needs to be adjusted in terms of legal constraints on central bank liability. If there is a high degree of non-linearity in the adjustment behavior of the reinforcement learning agent during periods of intense structural turbulence that cannot be precisely represented, it will lead to problems with democratic legitimacy. Propose to establish a "hybrid algorithmic accountability framework", which has an autonomously implemented policy agent and a Mathematically bound trust region set by human policymakers. The two-layered structure guarantees that during its operation in terms of computation plasticity, the algorithm is able to eliminate any new forms of macroeconomic innovation; Meanwhile, after all these processes have been completed, the final boundary condition will be within the realm of institutional governance. To integrate automated optimisations with manual supervision of supervisors to provide a viable path for improving

the stability of macroeconomic environments characterised by rising systems' complexities and periodicity in uncertainties.

6.3 Future Research Trajectories and Global Macro-Prudential Integration

Although existing studies have provided an empirical verification of the algorithmic reduction of economic oscillations, in the subsequent development process, how to address the complex characteristics of international financial cross-border transmission and macro-prudential regulatory instrument integration still need more exploration. As deep reinforcement learning has expanded its applications across various entities managing different national monetary systems within one system that contains multiple agents, exploring how non-cooperative Nash equilibrium can arise globally through such a multi-agent structure appears feasible. To Systematically Study how Algorithmic Policy Formation Influences the Dissemination Path of Financial Friction Shock Propagation Mechanisms and Its Impact on International Liquidity Cycles through an Extension. In addition, the merger of fiscal-monetary cooperation in the state-action space is required as a constraint on the autonomy agent to pass through the complicated temporal limits of sovereign debt sustainability and aggregate demand adjustment.

In short, the outcomes of this research serve as a catalyst for the broader drive towards "computational central banking". With global economic Structure continuing to move towards higher levels of Non-linear connection between them, relying on traditional stylised models becomes an increasingly perilous strategic risk. To build a strong, data-based reinforcement learning model can realise a synthesis of countercyclical policies which truly have no relation to policies and also show dynamic resistance in this setting. The Era of Static, Parameterised Policy Rule is approaching its Mathematical Limit; In the future, Macro-economic Stability will originate from Autonomously Inter-temporal Optimisation on the Path of the Policy Manifold by Advanced Artificial Intelligence.

Funding

This work was supported by 2025 Henan Soft Science Project: Research on the Optimization Path of Deep Integration between Supply Chain Finance and Henan Agricultural Industry Chain under the Background of New Productive Forces.

About the Authors

Shizhen Wang was born in Luoyang, Henan, China, in 1982. She has a Master’s degree and is an associate professor. Her main research direction is corporate financial management and financial economics.

References

[1] Araujo, D., Doerr, S., Gambacorta, L., & Tissot, B. (2024). Artificial intelligence in central banking (BIS Bulletin No. 84).

[2] Atashbar, T., & Shi, R. A. (2023). AI and macroeconomic modeling: Deep reinforcement learning in an RBC model. IMF Working Papers, 2023/040.

[3] Brini, A., et al. (2023). Reinforcement learning in a New Keynesian model. Algorithms, 16(6), 280.

[4] Charpentier, A., Elie, R., & Remlinger, C. (2023). Reinforcement learning in economics and finance. Computational Economics, 62(1), 1–39.

[5] Chen, M., Cont, R., Joseph, A., Kumhof, M., Pan, X., Xiong, W., & Zhou, X. (2025). Deep reinforcement learning in a monetary model (Bank of England Staff Working Paper No. 1,142).

[6] Covarrubias, M. (2023). Dynamic oligopoly and monetary policy: A deep reinforcement learning approach [Manuscript / Working Paper].

[7] Deák, S., et al. (2023). Reinforcement learning in a New Keynesian model. Algorithms.

[8] Fernández-Villaverde, J. (2025). Deep learning for solving economic models (NBER Working Paper No. 34250).

[9] Flak, A. (2025). Teaching an artificial central bank to conduct monetary policy (University of St. Gallen Working Paper).

[10] Galí, J. (2015). Monetary policy, inflation, and the business cycle: An introduction to the New Keynesian framework and its applications. Princeton University Press.

[11] Hill, E. (2021). Solving heterogeneous general equilibrium economic models with deep reinforcement learning. arXiv preprint arXiv:2103.16977.

[12] Hinterlang, N., & Tänzer, A. (2021). Optimal monetary policy using reinforcement learning (Deutsche Bundesbank Discussion Paper No. 51/2021).

[13] Hommes, C., et al. (2025). CANVAS: A Canadian behavioral agent-based model for monetary policy. Journal of Economic Dynamics and Control, 172, 104986.

[14] Kase, H., Rottner, M., & Stohler, F. (2026). Generative economic modelling (Working Paper).

[15] Kazinnik, S. (2025). A multi-agent system for monetary policy decision modeling (George Washington University Working Paper).

[16] Koop, G. (2003). Bayesian econometrics. John Wiley & Sons.

[17] Lillicrap, T. P., et al. (2016). Continuous control with deep reinforcement learning. International Conference on Learning Representations (ICLR).

[18] Maliar, L., & Maliar, S. (2021). Deep learning for solving dynamic economic models. Journal of Economic Dynamics and Control, 122, 104785.

[19] Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

[20] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[21] Smets, F., & Wouters, R. (2007). Shocks and frictions in modern business cycle models: A Bayesian DSGE approach. American Economic Review, 97(3), 586–606.

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

[23] Wang, T., & Chen, S. (2025). Reinforcement learning for monetary policy under macroeconomic uncertainty: Analyzing tabular and function approximation methods. arXiv preprint arXiv:2512.17929.

[24] Woodford, M. (2003). Interest and prices: Foundations of a theory of monetary policy. Princeton University Press.

[25] Yang, Y., et al. (2025). Structural reinforcement learning for heterogeneous agent models. arXiv preprint arXiv:2512.18892.

[26] Zheng, S., Trott, A., Srinivasan, S., Naik, N., Gruesbeck, M., Parkes, D. C., & Socher, R. (2022). The AI economist: Taxation policy design via multi-agent deep reinforcement learning. Science Advances, 8(18), eabm7842.