AICB: Towards an Artificially Intelligent Central Banker

Caio Simon[1]

[1] Universitat Pompeu Fabra
caio.simon01@estudiant.upf.edu

Contents

1 Introduction

Within the confines of economics, a rich corpus of study has developed over decades focusing on the theory and application of monetary policy. While doubtlessly a subject of much importance to the functioning of the economy at large, the setting in which central bankers make their decisions is also intrinsically interesting. They must decide on policy in an uncertain environment where their choices have large consequences, and these decisions are predicated on the management of the expectations of the public. This environment places the Central Bank (CB) into a perilous environment where credibility is of paramount importance and can quickly be lost. To succeed, the policymaker must play a game of words and intertemporal commitments dilemmas that form an interesting playing ground not only for humans, but also for the burgeoning field of Reinforcement Learning (RL) which is concerned with developing optimal policies for sequential decision-making settings from data. Part literature review and part empirical study, my goal with this paper is to unite the fields of economics and RL to explore the effectiveness of modern algorithms on the conduction of monetary policy.

The outcome of this paper is a simple model of the economy that RL algorithms can operate on, with empirical experiments included and benchmarked against reference policies. Interest in the application of RL to this environment is that it presents many challenges that current algorithms must cope with. The environment features partial observability as the CB is not able to directly observe private sector expectations. Additionally, as will be shown, the setting under which CBs make decisions is one resembling a game, which touches on the complexities involving multi-agent reinforcement learning (MARL). Lastly, the paper is the first, to my knowledge, to integrate forward guidance and narrative economics into RL for economics.

In the next section, I cover the requisite economic scholarship to obtain a good understanding of the challenges in monetary policy and draw parallels to the troubles faced by RL, with a short description of how multi-agent reinforcement learning fares currently. Then, I describe a theoretical model I adapted from existing literature, which relaxes some assumptions made in classic papers in order to make the model more tractable for empirical studies. Afterwards, I train an RL model on this environment and benchmark it against common heuristics found in economics. A discussion follows at the end, detailing steps for future research and implications for both the economic and RL domains.

2 Previous literature

2.1 Monetary Policy Background

In this section, I explore the rich literature that has been built around monetary policy, detailing how economists have thought about the problems inherent in setting rates and picking optimal paths. In subsection 2.1.1 I investigate how economists have used control theory and optimisation in classic papers from the period of 1977-2003. Here, I also expose the conundrums behind monetary policy and explore its game-like quality which renders optimal solutions often hard to find. Then in subsection 2.1.2 I look at how the remit of scholarship expanded to study the unconventional tools employed in the 2010's, focusing especially on the narrative side of economics. Reading is recommended, but not necessary.

2.1.1 Classic papers (1977-2003)

At first glance, monetary policy seems like a problem that could be solved by control theory. It is a sequential decision making process where the agent (the CB) must set interest rates in order to keep the economy around a target. Often, modern CBs such as the Federal Reserve are given a so-called "dual mandate," where they must keep the economy within an inflation target (2% in the long-run for the Federal Reserve) while also aiming for full employment — defined as the "highest level of employment or lowest level of unemployment that the economy can sustain in a context of price stability" (Board, 2025). This is an environment that control theory, which specialises in the analysis of dynamical systems, seems to be applicable to.

However, as first shown by Kydland and Prescott (1977) optimal control theory is unsuitable for monetary policy if we assume that private agents have knowledge of the structure of the economy and the policymaker's optimisation process. In the words of Kyndland and Prescott, central bankers do not face a "game against nature but, rather, a game against rational economic agents." The situation leads to equilibria that are consistent — in that they maximise a utility function in every period and continue to do so in the future — but often suboptimal due to the private market anticipating future actions and doing their fair share of optimisation themselves. This was best shown in the now-famous Barro and Gordon (1983) paper, where the authors extend the analysis by framing the problem of monetary policy as a tension between discretion, where policymakers can select the optimal action in every period, against a rule-based approach that they must commit to. In their world, a CB faces a cost function $\mathcal{C}$ that punishes the existence of inflation but rewards unexpected inflation (as unexpected inflation reduces real prices and increases output), like so [1]:

$$\mathcal{C} = \frac{a}{2}(\phi_t)^2 - b_t(\phi - \phi^e), \text{a and } b_t > 0 \tag{1}$$

Where inflation is represented by ϕ and the expectations thereof, ϕ^e. The first term punishes the existence of inflation, while the second rewards unexpected inflation. b_t is a benefit parameter that has mean $\bar{b}$ and variance σ_b^2. The CB's objective is to minimise the expected sum of costs given some discount rate, γ.

[1] I take a slightly involved approach to this literature review in order to give space to the equations found in classical economics papers to show that, in fact, they resemble the setup found in RL quite closely, but with some key differences that are explained in the text. All equations are taken from the original papers, but the symbols are sometimes changed in order to fit with modern customs.

Clearly this setup is the same that is often assumed when one solves a dynamic programming problem, and a pattern that was borne out in economics:

$$Cost = \mathbb{E}[\mathcal{C}_t + \gamma \cdot \mathcal{C}_{t+1} + \gamma^2 \cdot \mathcal{C}_{t+2} + ...] = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t \mathcal{C}_t] \tag{2}$$

Where both approaches differ is in the environment's behaviour. Whereas often in dynamic programming the environment is somewhat static, in the paper the authors assume that private agents have *rational expectations*, which implies they know the economy and the policymaker's optimisation problem. Under these conditions the policymaker has three main options: to optimise at every period, following discretionary monetary policy, commit to a rule that anchors expectations at 0, or use the commitment to a rule as a ruse and cheat to generate unexpected inflation from which it benefits. The authors show that, in this world of rational expectations private agents anticipate the incentive to cheat, and so do not anchor their expectations to 0 but rather to the true expected level of inflation. The result is a level of the cost below the socially optimal level of inflation (0, in the model) and a persistent inflation bias.

Strategy	**Policymaker cost** $\boldsymbol{\mathcal{C}_t}$	**Inflation** $\boldsymbol{\phi}$	**Inflation expectations** $\boldsymbol{\phi^e}$
Cheating (with people expecting the rule)	$-\frac{1}{2}\frac{\bar{b}^2}{a}$	$\frac{\bar{b}}{a}$	0
Rule	0	0	0
Best enforceable rule	$\frac{1}{2}\frac{\bar{b}^2}{a}\frac{1-\bar{\gamma}}{1+\bar{\gamma}}$	$\frac{\bar{b}}{a}\frac{1-\bar{\gamma}}{1+\bar{\gamma}}$	$\frac{\bar{b}}{a}\frac{1-\bar{\gamma}}{1+\bar{\gamma}}$
Discretion	$\frac{1}{2}\frac{\bar{b}^2}{a}$	$\frac{\bar{b}}{a}$	$\frac{\bar{b}}{a}$

Table 1: Outcomes described by Barro and Gordon under different central bank strategies, sorted by ascending order of cost. The CB has the incentive to cheat consumers and create unexpected inflation but, under rational expectations, agents anticipate this incentive and cease to believe the bank's promises, leading to a suboptimal level of cost and persistent inflation.

Great parallels exist between the usual setup of dynamic programming/control theory and what is usually done in economics but, as shown by the two papers, CBs face an environment where they must take credibility and expectations into account. This setting differs from the usual RL environment and resembles that which is studied by the multi-agent reinforcement learning subfield. Here, the challenge is to search for the sequence of actions that lead to the best equilibria (if they exist), rather than the outright optimal solution, which may be impossible.

The Barro and Gordon paper increased interest in this push and pull between discretion and rules in monetary policy. This culminated in **TaylorDiscretionPolicyRules1993<empty citation>**, one of the most cited papers in the field. In this work, Taylor surveys a number of attempts to craft suitable policy rules usually by means of econometrics. He synthesised the information and put forward his own formulation, seen in equation 3, which was simple but also remarkably effective in describing Federal Reserve policy.

$$i = \phi + \frac{1}{2}y + \frac{1}{2}(\phi - 2) + 2 \tag{3}$$

where i is the nominal interest rate set by the CB, ϕ the inflation rate over the previous four quarters, and y the percentage deviation of real GDP from a target. As Taylor himself argued, the CB should not set interest rates mechanically according to the rule; it disregards the context of the economy's situation and focusses on too narrow a set of variables, but it is a heuristic that has stood the test of time. Even modern econometric studies often find it useful to benchmark their performance against a Taylor rule, as this study will also do.

After Taylor's paper, a string of important contributions came to the fore, such as Clarida et al. (1999) who derive Taylor-type rules more formally from microeconomic foundations under some realistic assumptions. A significant application of control theory to the problem was also made by Benigno and Woodford (2003), who applied approximate linear-quadratic control to solve the issue. They circumvented the issue posed by Kydland and Prescott (1977) and Barro and Gordon (1983) by dealing only with situations where the policymaker could credibly commit to a policy. As such, the game-based solutions became less hostile to control-theory techniques. Using this approach, they find that the optimal policy is not as strict as a rule, but rather a "flexible inflation target" that allows for shocks and disturbances, but which anchors expectations around an optimal point. Their finding, while not as easily applicable as a Taylor rule (it requires the estimat of many some parameters, which can be hard) is much closer to the way in which CBs actually set interest rates.

2.1.2 Unconventional Policy

The 2008 financial crisis drove CBs to the limit of their conventional toolkit of interest rate changes. To further stimulate the economy in times of trouble, they had to rely on unconventional tools like quantitatice easing, negative interest rates, and forward guidance Bhattarai and Neely (2022), the latter of which is analysed in this paper. The crux of forward guidance is for the central bank to release statements committing itself future interest rate levels. It is an attempt to influence expectations and shape long-run interest rates, which CBs do not have direct control over. As this unfolded, a new field dubbed “narrative economics” was born, whose objective is to analyse the policy effects of speeches delivered by the CB employing both traditional economic models, (Bundick & Smith, 2020; Campbell et al., 2019; Hagedorn et al., 2019; McKay et al., 2016) and new machine learning techniques analysing real speeches such as Ashwin (2022), Shiller (2017), and Zahner (2023). The approach introduces an added level of realism to economic analysis which was not present in the earlier papers. Indeed, CBs are not simply rate-setting machines, but also institutions that have to contend with real human fears and behaviour; communicating properly with the market both reduces uncertainty and serves as a powerful tool for monetary policy — provided that central banks can credibly commit to future policy, as was derived in earlier classical papers.

2.2 Reinforcement Learning

Reinforcement learning encompasses a family of techniques that attempt to learn an optimal decision-making policy from data and interaction with the environment. While it has seen success in playing Atari games (Mnih et al., 2013), Go (Silver et al., 2016), and LLMs (Lambert, 2026), it has only recently been applied to monetary policy or economics more generally. Hinterlang and Tänzer (2021) used neural networks to estimate transition dynamics from historical data and proceeded to employ a Deep Deterministic Policy Gradient (DDGP) algorithm to find an optimal monetary policy. However, their approach leaves no space for fundamental concepts in economic theory such as inflation expectations because they cannot be easily observed and estimated using neural networks. A possible concern with their approach is the handling of the Lucas critique (Lucas, 1976), which posits that evaluation of policies based solely on historical data is naïve. Agents would have changed their expectations and behaviours should policies have been different in counterfactual scenarios. By estimating transition dynamics from historical data, the authors could crystallise the parameters and behaviours held during the observation period and propagate it into the backtest they perform with the RL-derived policy. They address this by evaluating their policies on more micro-founded models of the economy after having trained them, though their more complex policies could not be evaluated due to the structure of off-the-shelf models used by the authors.

A second attempt at bringing RL to monetary policy was made by Flak (2025). His approach was less reliant on historical data as he first created a theoretical model of the economy dubbed the “IS-TR-PC,” which is a variation on macroeconomic models commonly taught at undergraduate level. As noted by the author, the simplest version of the model could be solved with control-theory techniques because it is a version of a Linear-Quadratic-Gaussian problem, whose solution takes the shape of a Linear Quadratic Regulator. Yet non-linearities are added to the model to introduce further realism and make it into an environment apt for RL. The author found a policy in this toy-environment using the PPO algorithm, and then proceeded to use the approach in a more realistic simulator built by the Federal Reserve called the FRB/US model for tests. The policies found by PPO were shown to be superior to benchmark Taylor rules. While the paper does include a thoughtful mechanism for inflation expectations are updated, no paper has as of yet incorporated forward guidance into the model, leaving an important part of a central bank’s job untouched.

Lastly, it is important to survey the broader theoretical field of RL and how its contributions can be mapped onto the monetary policy problem. Mazumdar et al. (2019) show that policy gradient methods often fail to converge to Nash equilibria in linear-quadratic games. This is important, as the environment faced by a central bank has been theorised to resemble a game against rational economic agents (Kydland & Prescott, 1977) and often include elements of a linear-quadratic problem (Benigno & Woodford, 2003; Clarida et al., 1999; Flak, 2025). Works that applied RL to monetary policy circumvented this issue by relaxing the assumption of rational expectations. This relaxation makes the environment less tortuous for the agent and, arguably, could resemble real households better than assuming full rationality. Private agents are not usually explicitly modeled in these papers (though there are few to choose from) as optimisers in their own right. Their expectations usually take on a form prescribed by the modeller, but a proper MARL simulation of the economy would be a promising future step in research. For that to work in partially observable environments, update rules like the ones presented in Srinivasan et al. (2018) could be used to improve the performance of actor-critic methods.

3 The environment

3.1 The "laws of motion"

A simple and learnable macroeconomic environment, similar to standard undergraduate textbook treatments and to the framework of Flak (2025), is used to model the economy. The model focuses on two aggregate variables: real output (GDP), denoted by y_t, and inflation, denoted by ϕ_t.

IS-implied output At each period t, monetary conditions determine a notional level of output implied by a standard investment–savings (IS) relationship. This level captures the output that would prevail in the absence of adjustment frictions and is given by

$$\tilde{y}_t = y^* - \sigma\big(r_t - r_t^*\big) + \varepsilon_t^y, \tag{4}$$

where y^* denotes potential output, r_t is the ex ante real interest rate, $\sigma > 0$ governs the sensitivity of aggregate demand to real interest rates, and ε_t^y is a zero-mean demand shock. r_t^* is the natural real rate, which follows a random-walk process as seen in Holston et al. (2017).

$$r_t^* = r_{t-1}^* + \varepsilon_t^{(r^*)} \tag{5}$$

GDP dynamics Actual GDP adjusts gradually toward the IS-implied level. This adjustment is modeled as a partial-adjustment process:

$$y_t = \rho\, y_{t-1} + (1-\rho)\, \tilde{y}_t, \qquad 0 < \rho < 1. \tag{6}$$

The exponential moving average formulation introduces autocorrelation to the model, preventing wild swings in GDP, and allows monetary policy from past periods to have effects in subsequent time steps, which aids in realism.

The central bank's policy toolkit affects y through the real rate of interest, r_t, which decomposes via the Fisher equation into $r_t = i_{eff,t} - \phi_{t+1}^e$, with $i_{eff,t}$ being the effective rate of interest at time t (which will be further explained later) and ϕ_{t+1}^e representing inflation expectations for period $t+1$.

Phillips curve The second key variable, inflation ϕ is set by a linearised Phillips curve of the form

$$\phi_t = \phi_{t+1}^e + \kappa\Big(\frac{y_t - y^*}{y^*}\Big) + \varepsilon_t^\phi \tag{7}$$

which conceptualises inflation as a combination of inflation expectations, ϕ_{t+1}^e, and deviations of GDP from its natural rate, $(\frac{y_t - y^*}{y^*})$. The slope of the Phillips curve is κ, and ϵ_t^ϕ is a zero-mean shock with standard deviation σ_ϕ. From this equation, it becomes clear that the CB's main tool in inflation control is to affect y, as inflation expectations are set by the market itself. If ϕ_{t+1}^e deviates significantly from target inflation, ϕ^*, then actual inflation also responds. This creates a dynamic where the bank needs to keep expectations anchored close to their target if they wish to prevent spiraling into an inflationary cycle.

Inflation expectations Past reinforcement learning literature has largely avoided incorporating rational inflation expectations, as fully rational private agents tend to become ruthless optimizers, turning the environment into a game. Following the approach of Flak (2025), I implement a relaxed form of inflation expectations, described by:

$$\phi^e_{t+1} = (1 - \vartheta_{t-1})\,\phi^* + \vartheta_{t-1}\,\phi_{t-1}, \tag{8}$$

where the expectation weight ϑ_{t-1} evolves according to

$$\vartheta_{t-1} = \begin{cases} \tilde{\vartheta}\left[1 - e^{-\overline{k}\,(\phi_{t-1}-\phi^*)}\right], & \text{if } \phi_{t-1} > \phi^*, \\ \tilde{\vartheta}\left[1 - e^{-\underline{k}\,(\phi_{t-1}-\phi^*)}\right], & \text{if } \phi_{t-1} \leq \phi^*. \end{cases} \tag{9}$$

This formulation has two attractive properties. First, the "unanchoredness" of expectations represented by ϑ_{t-1} grows sharply if the central bank allows inflation to deviate significantly from its target ϕ^*, meaning the CB must devote more attention to inflation management the more unanchored inflation expectations become. Second, it replicates the behaviour seen in Banerjee and Mehrotra (2023) where, empirically, inflation expectations behave differently when they are below and above the target. This can be achieved by setting $|\underline{k}| < \overline{k}$.

3.2 The Central Bank's actions

In this model of the economy, the central bank has two actions at its disposal in every time period, both of which are intended to influence the effective rate of interest, $i_{eff,t}$. The first action is to simply change the nominal rate of interest i_t. This is the standard tool in the monetary policy toolkit and, for simplicity, the bank's options for interest rate changes are set to be between -3% and 3% in every period. Then, the central bank is also allowed to issue one of four forward guidance statements. The bank can issue a hawkish statement, implying an increase in interest rates; a dovish statement, implying a decrease in interest rates; a hold statement that signifies rates will remain steady; and an option to release no statement, where the market is left to infer the next action without guidance. The forward guidance creates a promise for period $t + 1$ regarding what the central bank will do in future, which affects the decisions of the private market.

The meanings of each statement are not set in stone but rather evolve with the actions taken by the CB. The market observes the development of promises and their subsequent realisations and updates its expectations in a Bayesian manner, though with some modifications. In this conceptualisation, each type of statement s has its own vector $\boldsymbol{\alpha_s}$, which stores the prior and information garnered throughout the episode. As shown in Figure 1, priors are given to each statement in the initial state which place high probability mass in areas that are consistent with the statements. This helps steer the meaning of forward guidance while allowing for updating beliefs.

Formally, let $\mathbf{p}_s = (p_{s,1}, p_{s,2}, \ldots, p_{s,K})$ be the vector of probabilities for K possible interest rate moves Δi given a statement s. The distribution is defined as:

$$\mathbf{p}_s \sim \text{Dirichlet}(\boldsymbol{\alpha}_s) \tag{10}$$

Through Dirichlet-Categorical conjugacy, the market's point estimate for the probability of any specific move k is given by the mean of the Dirichlet distribution:

$$\hat{p}_{s,k} = \frac{\alpha_{s,k}}{\sum \alpha_{s,j}} \tag{11}$$

And the expected rate change becomes the normalised dot product between the possible rate moves, represented by vector $\mathbf{v}$, and $\hat{p}_{s,k}$:

$$\mathbb{E}[\Delta i_{t+1} \mid \boldsymbol{\alpha}_s] = \frac{\mathbf{v} \cdot \boldsymbol{\alpha}_s}{\sum_{j=1}^{K} \alpha_{s,j}} \tag{12}$$

The updating process is straightforward due to the conjugacy of the Dirichlet and Categorical distributions. After issuing a statement s in period t, the market observes the realised interest rate change in period $t + 1$ and evaluates whether the central bank's action was consistent with the direction implied by

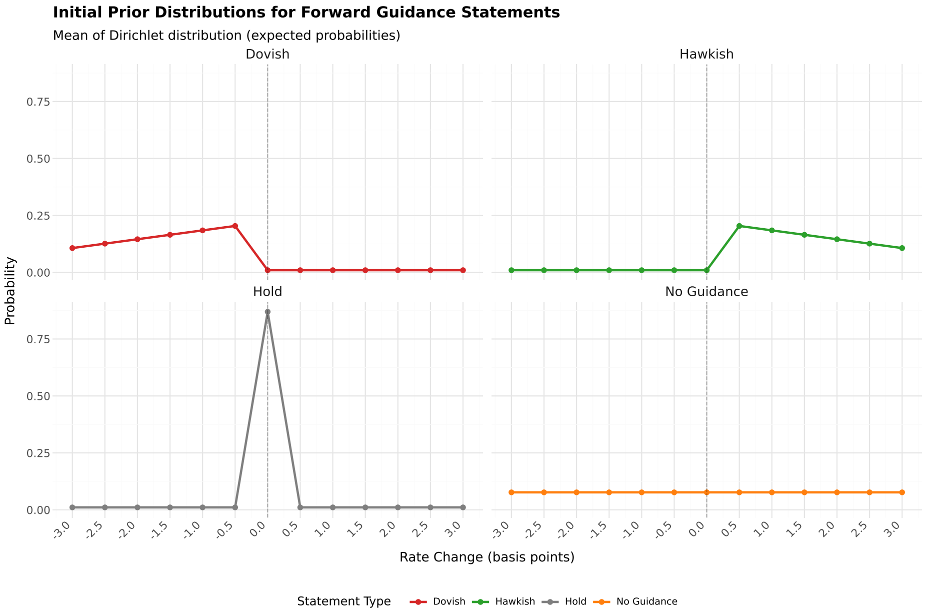


Figure 1: Market priors following a Dirichlet distribution.

the statement. If the realised move aligns directionally with the guidance, the promise is considered kept; otherwise, it is considered broken.

Formally, the alignment is determined by the realized interest rate change Δi_{t+1} relative to the statement s issued at time t:

- **Hawkish (s_H):** Kept if $\Delta i_{t+1} > 0$; broken if $\Delta i_{t+1} \leq 0$.
- **Dovish (s_D):** Kept if $\Delta i_{t+1} < 0$; broken if $\Delta i_{t+1} \geq 0$.
- **Hold (s_O):** Kept if $\Delta i_{t+1} = 0$; broken if $\Delta i_{t+1} \neq 0$.
- **No Statement (s_N):** Serves as a neutral baseline. The market updates its beliefs based on the realized action to learn the bank's default policy tendency, but since no commitment was made, the promise cannot be "broken."

To capture the reputational cost of breaking forward guidance, updates to beliefs are asymmetric. Specifically, belief updates are three times as strong when a promise is broken than when it is kept. Let u denote the update intensity:

$$u = \begin{cases} 1 & \text{if the promise is kept,} \\ 3 & \text{if the promise is broken.} \end{cases} \tag{13}$$

Let e_k be a one-hot vector indicating the realised interest rate move k (i.e., $e_{k,j}$ if $k = j$, and 0 otherwise). The belief vector associated with statement s is then updated according to:

$$\boldsymbol{\alpha}_s^{\text{new}} = \boldsymbol{\alpha}_s^{\text{old}} + u\,\mathbf{e}_k \tag{14}$$

.

This rule implies that each observation incrementally reshapes the market's interpretation of forward guidance statements, with broken promises accelerating belief revisions and thereby weakening the credibility of the central bank's communication over time. Importantly, as the episode develops, the reputational damage of broken promises attenuates due to the properties of the Dirichlet. To prevent this from happening, I institute a rolling memory window whereby the market forgets its oldest update after n time steps, but leaves the prior intact.

Finally, the market expectations for future policy in time $t+1$ is given by

$$i^e_{t+1} = i_t + \mathbb{E}[\Delta i_{t+1} \mid \boldsymbol{\alpha}_s], \tag{15}$$

Which is combined with the current policy rate to form the effective interest rate at time t, a convex combination of both rates:

$$i_{eff,t} = \omega i_t + (1-\omega) i^e_{t+1}, \qquad \omega \in [0,1]. \tag{16}$$

The parameter ω governs the relative importance of current monetary policy versus expectations in determining aggregate demand. Higher values of ω imply that the economy responds primarily to the contemporaneous policy rate, while lower values reflect a stronger influence of forward guidance and expected future policy actions.

To my knowledge, this belief-updating process is unseen in the literature, but represents a version of bounded rationality, which is a realistic and dynamic way of modelling the effects of forward guidance, with reputation being implicitly taken into account throughout the process.

3.3 The objective

At every time step, the central bank observes a reward R_t. Its objective is to find a policy π that maximises the expected sum of rewards subject to a discount rate γ (set at 0.99).

$$\max \mathbb{E} \sum_{t=0}^{\infty} \gamma^t R_t \tag{17}$$

Two reward functions were used to optimise this objetive. Originally, the reward function used was the one in (18), which was meant to encourage the central bank to be close to inflation and GDP targets, with larger deviations being penalised at a quadratic rate[2]. This reward function style has also been used in papers that formulate interest-rate control as a linear-quadratic problem and solve it with control theory methods. However, the non-linearities added throughout the process (such as in inflation expectations) turn this into a harder problem apt for deep RL.

$$R^{\text{Test}}_t = -\frac{1}{2}[(\phi_t - \phi^*)^2 + {\frac{y_t - y^*}{y^*}}^2] \tag{18}$$

Unfortunately, the reward function turned out to be disastrous for training. The model would rapidly hike interest rates, crashing the economy, or it would lower them into the negative numbers, creating hyperinflation. The situation was so difficult that an episode would usually last only eight steps — which makes the learning process highly inefficient. To aid learning, I developed a more complex loss function to discourage large rate changes and extreme levels of the interest rate, which can be seen in equation (20):

$$R^{\text{Train}}_t = -\frac{1}{2}[(\phi_t - \phi^*)^2 + {\frac{y_t - y^*}{y^*}}^2 + \frac{3}{10}(\Delta i)^2 + \frac{3}{10}(d_t)^2] \tag{19}$$

where the interest rate deviation from a threshold of extreme values is set as:

$$d_t = \begin{cases} (i_t - 0.10) \times 100 & \text{if } i_t > 0.10 \\ (-0.01 - i_t) \times 100 & \text{if } i_t < -0.01 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

While (20) was used for training, equation (18) was still used for benchmarking and testing the policy after training. In effect, (20) was a form of surrogate loss to speed up training.

This completes the description of the environment[3]. While the laws of motion of the environment are complex and involve a litany of equations[4], the key point is that the setting asks for a balance between GDP

[2] The deviations of inflation and GDP from target are scaled up by a factor of 100 such that they are interpreted as percentage points, not decimals.

[3] The values that the parameters of the model take on are detailed in the appendix. They were selected heuristically.

[4] For which I apologise. Economics papers are usually full of these equations.

and inflation targeting. To succeed, the central banking agent will need to learn to both set interest rates and manage expectations with its words. The model includes two axes of credibility — one for promise-breaking in its forward guidance and one in its commitment to target inflation — that mirrors the reputational dynamics found in the real world, and which the agent must keep abreast of.

4 Training and Results

The RL algorithm chosen for this task was Proximal Policy Optimisation (PPO) for its ability to handle policies that require multiple inputs per time step and its relatively light hyperparameter list. The model was trained for two-million steps with the only change from the default PPO settings from Stable Baselines 3 being an increase in the entropy coefficient to 0.01 to encourage exploration. All observations and rewards were normalised.

The agent was not allowed to observe all relevant variables — making this a partially observable environment. It was allowed to see key economic variables that a normal central banker would have access to, such as inflation, GDP (and lagged observations of the same), the interest rate, its last statement, and the market-defined interest rate expectations, i^e_{t+1}. It was not allowed to observe r^*_t, ϑ_{t-1}, however, or any of the explicit probabilities that define the distribution of variables for determining i^e_{t+1}. This increases the difficulty of the problem, but resembles the uncertainty faced by central bankers.

In order to ascertain just how good the model is, a comparison to other methods must be made. Luckily, Taylor rules offer a simple and intuitive benchmark with which to compare. I constructed three candidate Taylor rules. They all follow the formula:

$$i_t = \lambda i_{t-1} + (1-\lambda) \cdot (\phi + \frac{1}{2} y_{t-1} + \frac{1}{2}(\phi_{t-1} - 2) + 2) \tag{21}$$

Here λ is a smoothing parameter, which encourages the interest rate to change slowly.

4.1 Benchmarks

Pure Taylor rule with guidance The first candidate rule is a "pure Taylor rule" — by which I mean it is meant to replicate the original's construction — where λ is set to 0. This makes for a capricious, highly volatile rule which can take big cuts of the interest rate if deemed necessary. As for its forward guidance behaviour, at every time step it sends out a signal consistent with what it did in the current round. That is, if it performed a hawkish action in the current period (raised rates), then its forward guidance implies a raise in rates in the following period.

Smooth Taylor rule with guidance The second candidate, a "smooth Taylor rule" has a λ parameter which was heuristically set to 0.75. In contrast with the pure Taylor rule, this one is much more cautious and measured. It rarely makes big interest moves, preferring to change them by only half a percent in most cases. Its forward guidance behaviour is the same as the pure Taylor rule.

Smooth Taylor rule without guidance Finally, the third rule follows the exact same style as the smooth Taylor rule but instead of offering forward guidance, it opts always for the "No Guidance" option, which is a more classical type of Taylor rule, as the author did not specify any forward guidance behaviour. This helps us ascertain just how much forward guidance matters in this economy.

4.2 Evaluation

The economic environment is inherently noisy. To account for this stochasticity, I evaluate the policies using Monte Carlo (MC) simulations rather than a single trajectory, allowing for a more robust assessment of performance across 1,500 independent runs, each with a maximum of 250 steps.

The relative performance of each policy can be easily ascertained by comparing the distributions of the sum of rewards in the MC trajectories, shown in Figure 2. Encouragingly, the PPO policy bests all other policies. For a random trajectory, the PPO policy's sum of rewards outperforms the pure Taylor and the smooth Taylor with no guidance in over 90% of trajectories, while for a smooth Taylor rule with guidance, the RL agent surpasses it 75% of the time. Interestingly, if we disregard the PPO policy for a moment, we can also observe that forward guidance yielded a significant performance gain for the smoothed policy when compared to its silent counterpart. Yet, forward guidance was not enough to boost the pure Taylor rule, whose distribution of rewards resembles that of the mute smoothed policy. This suggests an environment where both interest rate policy and forward guidance have to work in conjunction to achieve success.

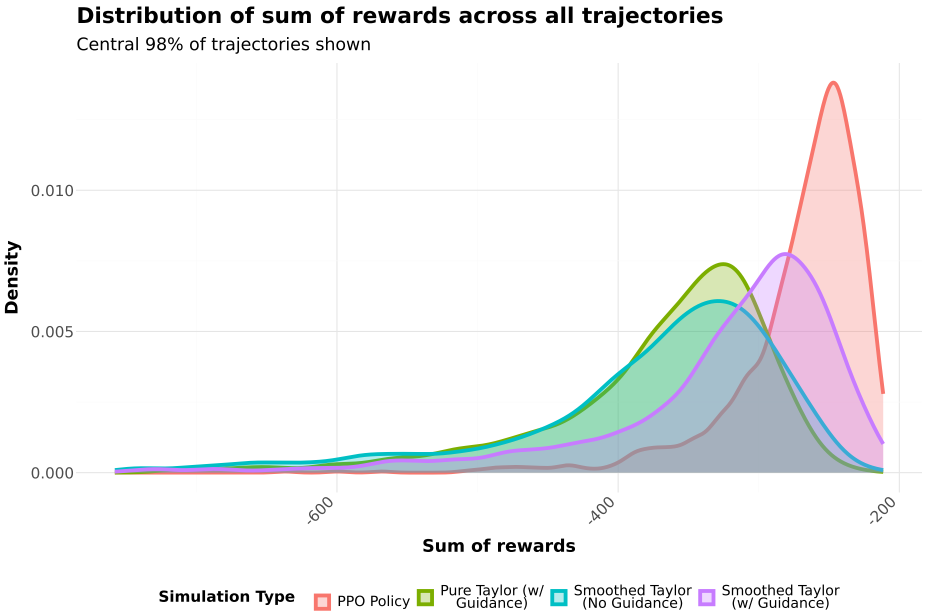
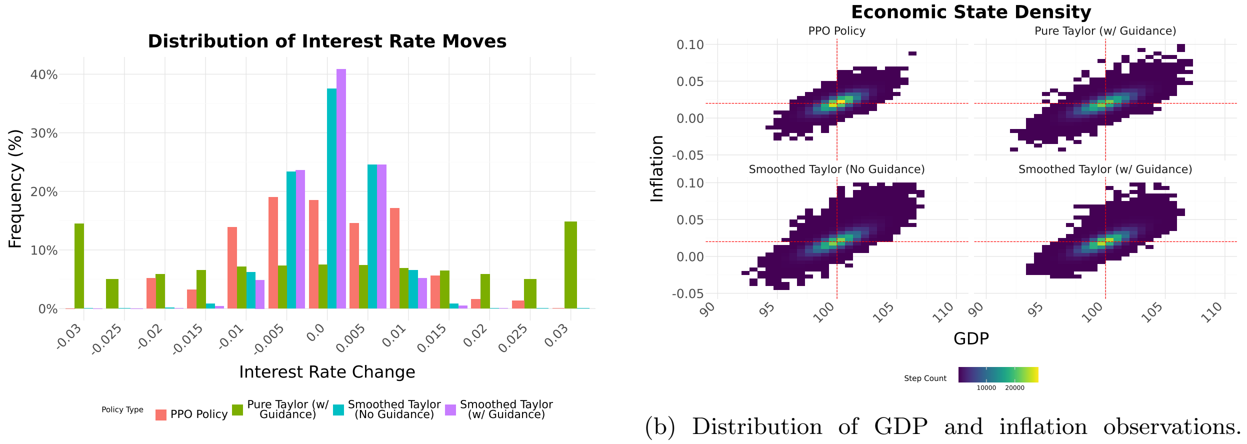


Figure 2: Distribution of sum of rewards across MC trajectories for each policy. Dataset-wide top and bottom 1% of observations not shown to control for outliers.

The PPO agent's performance is predicated on an interest rate control policy that is both less extreme than the pure Taylor rule — it rarely ever makes an interest rate change past 150 basis points — and more responsive than the pair of smoothed Taylor policies, which are exceedingly cautious. From panel 3a, it is clear that it does not shy away from changing rates, which helps it keep the economy within a noticeably tighter range than the other policies, as shown in the right panel.

The RL agent's situational awareness is also consistent with economic intuition. From Figure 4 we can see that the agent elects to increase rates (a hawk, in economic jargon), decrease them (a dove), or hold them steady in appropriate situations. When inflation and GDP are below target, its probability of decreasing rates is much higher, while the opposite holds true when the variables are above target. The 3D probability plot shows that the policy forms a clear wall demarcating regions where it stimulates and those where it dampens the economy. Moreover, there seems to be a slight bias towards increasing output in the model's policy. In Figure 3a we can see that the model picks a rate cut of 0.005 more often than it does a hike by a commensurate amount. This pattern is also clear in Figure 4, with significant probability of cutting rates being present in situations where inflation is above target but the economy is below its potential — a situation called "stagflation".



(a) Distribution of interest rate changes

(b) Distribution of GDP and inflation observations. Dashed lines mark target inflation and GDP.

Figure 3: Interest rate adjustment behavior across policies. The left panel shows the distribution of rate changes, while the right panel displays the distribution of states in the economy.

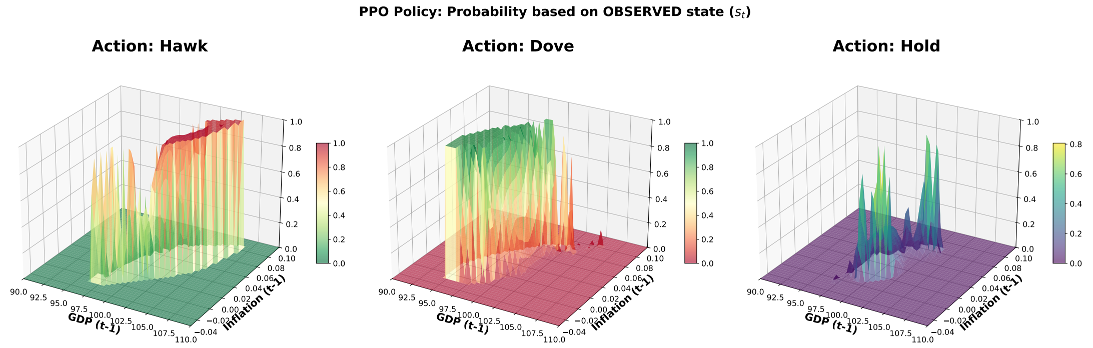
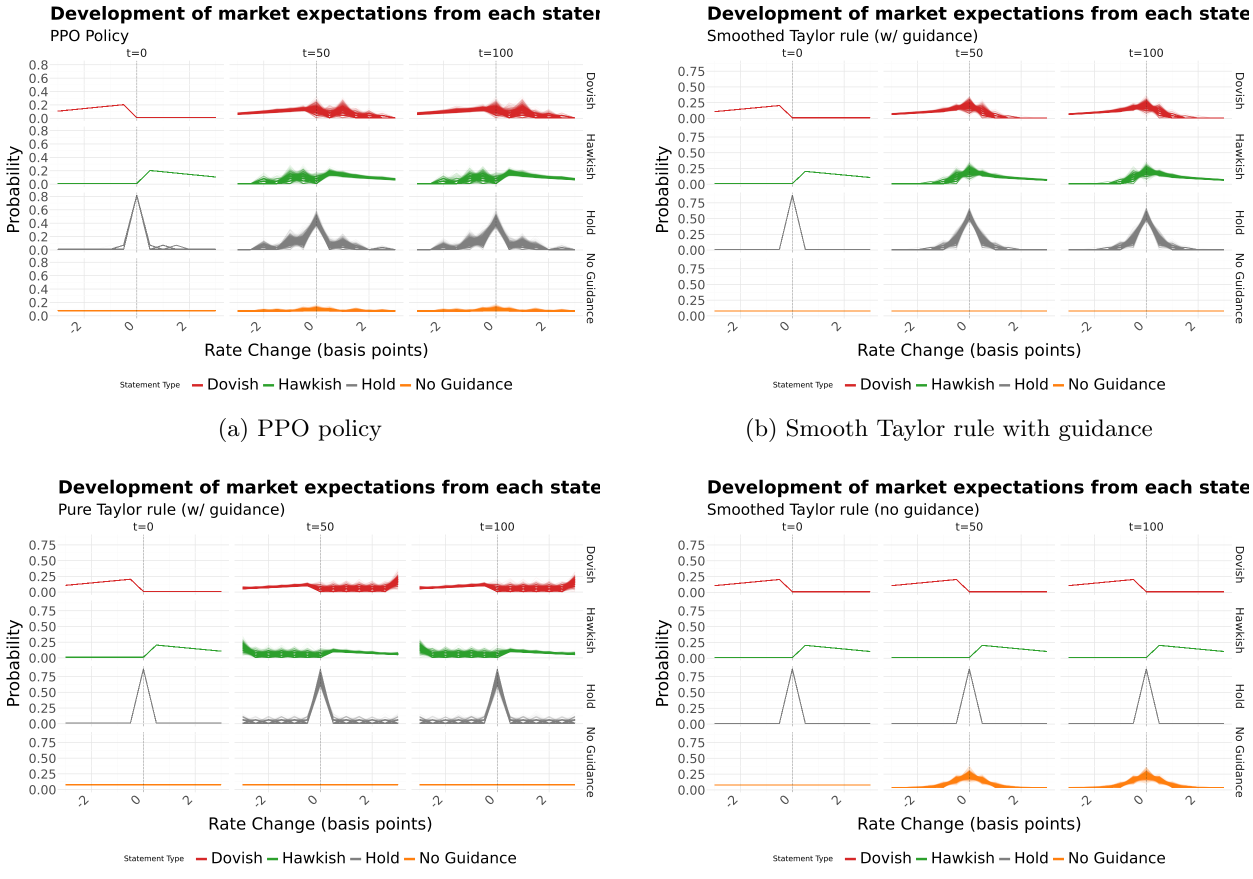


Figure 4: Probability of increasing rates (hawk), decreasing (dove), or holding them steady across states of the economy.

The other tool that central bankers have at their disposal — their words — also show an interesting, albeit puzzling pattern. We can understand how each policy's words affected expectations by plotting the Dirichlet distributions of each statement, which shows us the probability the market assigns to each interest rate move, also taking into account the prior information. And, through the MC draws, we can appraise the uncertainty and regularity around these distributions. Figure 5 shows these distributions for each statement type and policy at different checkpoints. Here we can see that the smoothed Taylor policy with guidance (panel 5b) maintains the shape of the prior fairly well. Most of the probability mass seems to go towards the "hold" action, which it performs very often, and there is also a measure of cheating the market that it performs. For example, for the "hawk" action, which in principle implies an increase in rates, a moderate *decrease* was actually expected by the time $t = 50$, blindsiding the market — and for the "dove" action the pattern reverses.

In the more extreme pure Taylor policy, cheating the market became an expectation. This is because it predicts it will do in the next step what it did in the current but, due to its volatile nature, the Taylor rule sees itself constantly see-sawing between rate hikes and cuts. The result is a policy that creates maximum confusion in the market, to the point that the signal in the central banker's words are diluted.

What emerges from the forward guidance mechanism is a puzzle induced by the environment. If we look at the reward distribution, it seems that forward guidance had a substantial positive impact in the smooth Taylor rule's performance when compared to its silent counterpart. Also, judging by its outperformance of the pure Taylor rule with behaviour that was altogether more consistent and trustworthy, and it seems that keeping the market's trust is a boon for performance. Yet the best performing policy, the one found by PPO, seems to have a forward guidance behaviour akin to that of the pure Taylor rule — capricious and unpredictable — which calls into question the limits of forward guidance within this environment.



(a) PPO policy

(b) Smooth Taylor rule with guidance

(c) Pure Taylor rule with guidance

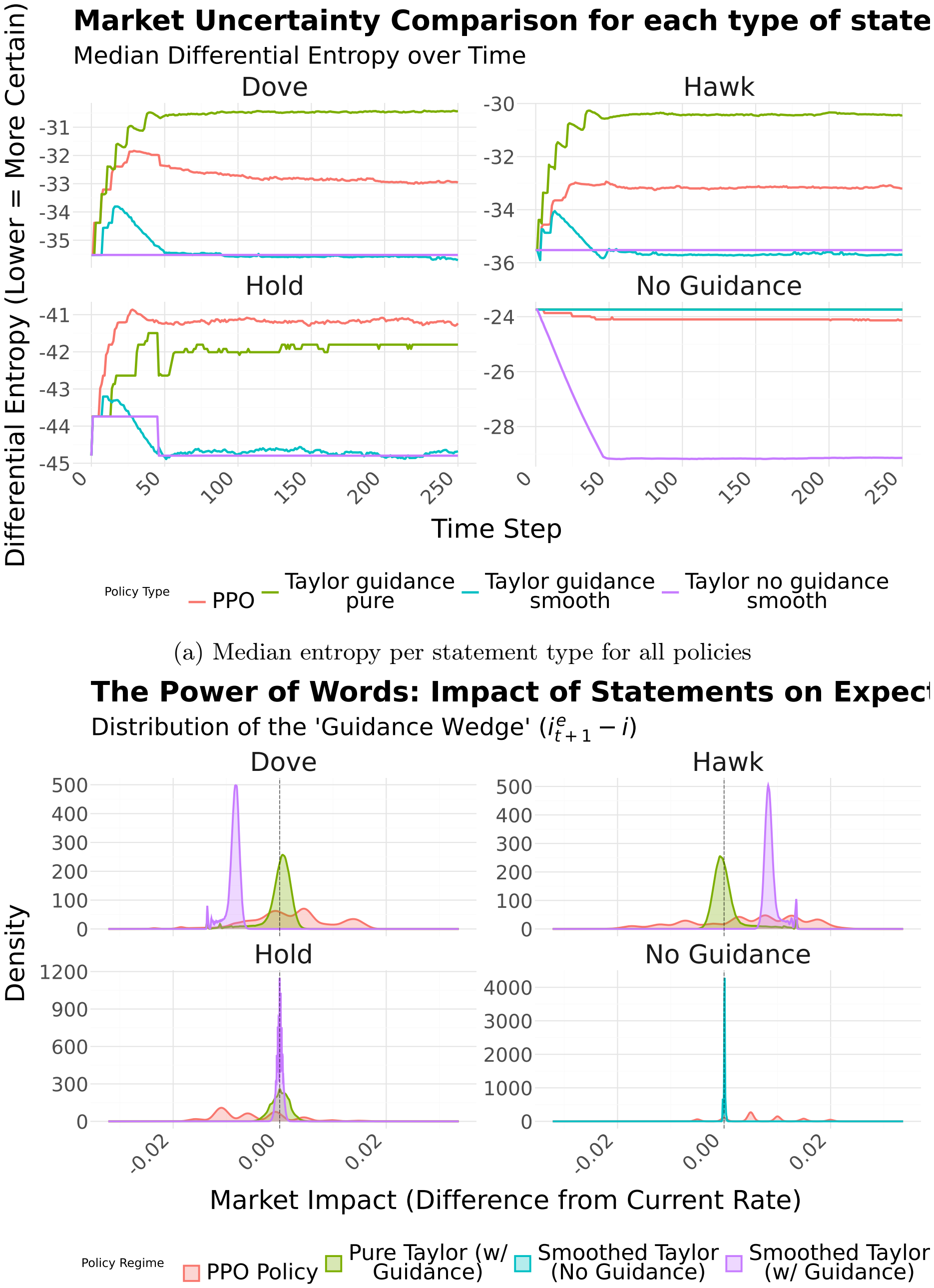
(d) Smooth Taylor rule no guidance

Figure 5: The Dirichlet distributions at different time points for each different policy and statement type

Put another way, the Dirichlet distributions also encode the uncertainty of the market, which we can quantify by way of entropy. As we can see in Figure 6a, both the pure Taylor rule and the PPO policy generally have an entropy-increasing effect on expectations, although the PPO policy's effects is more moderate. This corresponds to an increase in uncertainty which is not shared by the smoothed policies whose style is more measured and consistent.

In another way to deconstruct the pattern, it is possible to calculate the "policy wedge" that the agent receives from its forward guidance. The wedge is defined as the expected interest rate minus the actual policy rate, $i^e_{t+1} - i_t$, with large absolute values indicating that central bankers' words have more power. From the plot 6b, it is evident that the pure Taylor policy completely loses the trust of the market. By shaping its expectations into an almost uniform distribution, it foregoes any power in its communication. The smoothed Taylor rule has a much more arresting power over the market, and certainly benefits from its consistency when compared to other Taylor rules. The PPO policy, however, does not completely lose control over the market. The distribution of the policy wedge is much more diffuse than that of the pure Taylor rule, concentrated at zero. While not at all conclusive, this could suggest that the model "prunes" the expectations of the market in such a way that its strength is reduced but not entirely lost. Much like the "reward hacking" observed in early Deep RL applications (e.g., Atari games), the PPO agent may be exploiting the environment's expectation mechanism. It may find that the initial market expectations are too rigid, leading it to 'prune' or scatter those expectations to grant itself more operational flexibility.

Another possible explanation is simply that the environment is such that, if the interest rate control policy is good enough, the market's interest rate expectations and forward guidance do not matter. The agent may simply decide to follow its interest rate policy and ensure the market understands its behaviour well. This is supported by the rather spread out distribution of the guidance wedge, which suggests no preference for a direction of the expectations. However, the case is weakened because, if the agent only wanted the market to learn from its behaviour with no bias, then it could use the "No Guidance" statement, whose prior distribution is much more malleable. The bias introduced by the model is therefore likely of some attractiveness to the agent.



(a) Median entropy per statement type for all policies

(b) Distribution of the policy wedge: how powerful the central bank's words are

Figure 6: Interest rate adjustment behavior across policies. The top panel shows the distribution of rate changes, while the bottom panel displays the distribution of states in the economy.

To aid in the interpretation of the policy, it is possible to also plot the probability of each statement for every economic state (shown in Figure 7), which reveals significant overlap with Figure 4. It seems that the agent has a high probability of releasing a hawkish statement in areas where it chooses a hawkish action and likewise for dovish actions. However, the probability "wall" for statements is much smaller than for actions, showing perhaps that it does not forecast its next action very well yet, or that it prefers to empower its interest rate cuts and hikes with words only when the situation is in clear need of more power. In a way, it is very similar to the handcrafted policies that it competes against, as it predicts for the next step what it did in the current, though this has some subtleties. Thus, it is clear also why its behaviour falls in between the extremes of the pure and smooth policies — because its own rate-setting behaviour occupies a middle ground between the two behaviours. It also seems to rely the "No Guidance" option in situations of stagflation. We saw previously that in such situations it is likely to pursue a cut in the interest rate, so refusing to release any signal could be used to in effect blunt the effect of the interest rate change.

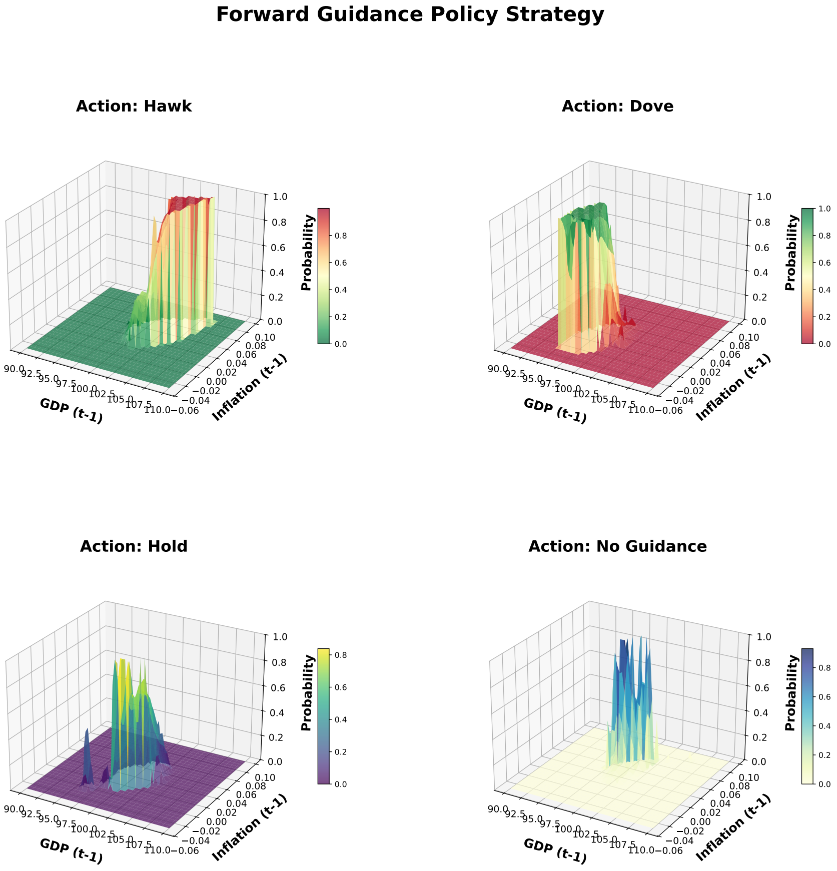


Figure 7: Probability of releasing a statement suggesting a hike, a cut, holding the interest rate, or no guidance at all.

5 Discussion

What emerged from the experiment with RL in an economic simulator was a policy quite proficient in its use of the interest rate as a monetary policy tool. Yet its use of forward guidance, could suggest some error in trying to predict its next move, or it could be intentionally shaping the expectations to its needs. It is still hard to interpret its “thought process” as is often the case with Deep RL.

It should be evident that the agent, while very good at solving this environment, cannot be transported to the real world yet. The environment crafted here is exploitable and learnable, while the real world has many more sources of bias and uncertainty than what was given in the model. For example, central bankers often face uncertainty in their measurements, which was not the case here. Moreover, the inflation expectations here were not as difficult to cope with as the rational expectations that are commonly assumed in economics. To become truly impressive, the model has to be trained in more complex economic simulators, such as the ones seen in Flak (2025) and Hinterlang and Tänzer (2021).

Furthermore, even within the environment some more tests could have been performed. The model may be, in a sense, overfitting to the current environment. An interesting experiment would be to raise noise levels and compare a model re-trained on such a world with the one that was trained in this paper. If the performance completely crashes, then it is clearly an overfit which is not suitable to the real world.

Lastly, future work should also include better benchmarks for the policy. The Taylor rules were selected in this scenario because they were easy to implement and offered good performance. However, after a suitably long training period, the RL policy was able to convincingly outperform them. In future, researchers could pit different RL algorithms against each other or specify functional forms for the RL algorithm — by having it tune the parameters of a Taylor rule, for example.

6 Conclusion

Overall, the work demonstrated that RL agents are indeed capable of operating in a complex economic environment. It was able to learn the policy rules that are typically applied in day-to-day central banking and also perform forward guidance. The main contributions of the paper are introducing a novel framework for modelling basic forward guidance in an RL environment and the application of the agent to such a simulator. Future work should focus on extending the environment in order to make it more realistic.

References

Ashwin, J. (2022, December 2). *Multi-Dimensional Uncertainty And Central Bank Communication: What Do Central Bankers Talk About And Why?* 4291983. https://doi.org/10.2139/ssrn.4291983

Banerjee, R., & Mehrotra, A. (2023). Unanticipated and Backward-Looking: Deflations and the Behavior of Inflation Expectations. *International Journal of Central Banking*, *19*(4), 41–83. Retrieved January 14, 2026, from https://ideas.repec.org//a/ijc/ijcjou/y2023q4a2.html

Barro, R. J., & Gordon, D. B. (1983, February). *Rules, Discretion and Reputation in a Model of Monetary Policy.* 1079. https://doi.org/10.3386/w1079

Benigno, P., & Woodford, M. (2003, August). *Optimal Monetary and Fiscal Policy: A Linear Quadratic Approach.* 9905. https://doi.org/10.3386/w9905

Bhattarai, S., & Neely, C. J. (2022). An Analysis of the Literature on International Unconventional Monetary Policy. *Journal of Economic Literature*, *60*(2), 527–597. https://doi.org/10.1257/jel.20201493

Board, F. R. (2025, August 22). *What economic goals does the Federal Reserve seek to achieve through its monetary policy?* Board of Governors of the Federal Reserve System. Retrieved January 11, 2026, from https://www.federalreserve.gov/faqs/what-economic-goals-does-federal-reserve-seek-to-achieve-through-monetary-policy.htm

Bundick, B., & Smith, A. L. (2020). The Dynamic Effects of Forward Guidance Shocks. *The Review of Economics and Statistics*, *102*(5), 946–965. https://doi.org/10.1162/rest_a_00856

Campbell, J. R., Ferroni, F., Fisher, J. D. M., & Melosi, L. (2019). The limits of forward guidance. *Journal of Monetary Economics*, *108*, 118–134. https://doi.org/10.1016/j.jmoneco.2019.08.009

Clarida, R., Gali, J., & Gertler, M. (1999). The Science of Monetary Policy: A New Keynesian Perspective. *Journal of Economic Literature*, *37*(4), 1661–1707. https://doi.org/10.1257/jel.37.4.1661

Flak, A. (2025). Teaching an Artificial Central Bank to Conduct Monetary Policy. Retrieved January 13, 2026, from https://www.alexandria.unisg.ch/handle/20.500.14171/122350

Hagedorn, M., Luo, J., Manovskii, I., & Mitman, K. (2019). Forward guidance. *Journal of Monetary Economics*, *102*, 1–23. https://doi.org/10.1016/j.jmoneco.2019.01.014

Hinterlang, N., & Tänzer, A. (2021). *Optimal Monetary Policy Using Reinforcement Learning*. 4025682. https://doi.org/10.2139/ssrn.4025682

Holston, K., Laubach, T., & Williams, J. C. (2017). Measuring the natural rate of interest: International trends and determinants. *Journal of International Economics*, *108*, S59–S75. https://doi.org/10.1016/j.jinteco.2017.01.004

Kydland, F. E., & Prescott, E. C. (1977). Rules Rather than Discretion: The Inconsistency of Optimal Plans. *Journal of Political Economy*, *85*(3), 473–491. Retrieved January 11, 2026, from https://www.jstor.org/stable/1830193

Lambert, N. (2026, January 2). *Reinforcement Learning from Human Feedback*. arXiv: `2504.12501 [cs]`. https://doi.org/10.48550/arXiv.2504.12501

Lucas, R. E. (1976). Econometric policy evaluation: A critique. *Carnegie-Rochester Conference Series on Public Policy*, *1*, 19–46. https://doi.org/10.1016/S0167-2231(76)80003-6

Mazumdar, E., Ratliff, L. J., Jordan, M. I., & Sastry, S. S. (2019, December 16). *Policy-Gradient Algorithms Have No Guarantees of Convergence in Linear Quadratic Games*. arXiv: `1907.03712 [cs]`. https://doi.org/10.48550/arXiv.1907.03712

McKay, A., Nakamura, E., & Steinsson, J. (2016). The Power of Forward Guidance Revisited. *American Economic Review*, *106*(10), 3133–3158. https://doi.org/10.1257/aer.20150063

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M. (2013, December 19). *Playing Atari with Deep Reinforcement Learning*. arXiv: `1312.5602 [cs]`. https://doi.org/10.48550/arXiv.1312.5602

Shiller, R. J. (2017). Narrative Economics. *American Economic Review*, *107*(4), 967–1004. https://doi.org/10.1257/aer.107.4.967

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Srinivasan, S., Lanctot, M., Zambaldi, V., Perolat, J., Tuyls, K., Munos, R., & Bowling, M. (2018). Actor-Critic Policy Optimization in Partially Observable Multiagent Environments. *Advances in Neural Information Processing Systems*, *31*. Retrieved January 11, 2026, from https://proceedings.neurips.cc/paper_files/paper/2018/hash/e22dd5dabde45eda5a1a67772c8e25dd-Abstract.html

Zahner, J. (2023, November 3). *Talking Fragmentation Away – Decoding the ’Whatever it Takes’ Effect*. 4622208. https://doi.org/10.2139/ssrn.4622208

7 Appendix

Table 2: Model Parameters and Calibration

Parameter Name	**Symbol**	**Value / Definition**
GDP target	y^*	100
Intertemporal rate of substitution	σ	200
Real interest rate	r_t	$i_{\text{eff},t} - \phi^e_{t+1}$
IS shocks	ε^y_t	$\mathcal{N}(0, \sigma^2_y)$
Natural rate of interest (R-star)	r^*_t	$r^*_{t-1} + \varepsilon^{(r^*)}_t$
Phillips curve shocks	ε^ϕ_t	$\mathcal{N}(0, 0.005)$
R-star shock	$\varepsilon^{(r^*)}_t$	$\mathcal{N}(0, 0.001)$
GDP persistence	ρ	0.7
Inflation expectations	ϕ^e_{t+1}	$(1 - \vartheta_{t-1})\phi^* + \vartheta_{t-1}\phi_{t-1}$
Inflation slope (above target)	$\overline{k}$	30
Inflation slope (below target)	$\underline{k}$	-20
Effective rate weight	ω	0.6