Optimal Monetary Policy using Reinforcement Learning

Natascha Hinterlang[a,*], Alina Tänzer[b]

[a]*Deutsche Bundesbank, Mainzer Landstraße 46, 60325 Frankfurt, Germany*
[b]*Goethe University Frankfurt, Theodor-W.-Adorno Platz 3, 60323 Frankfurt*

Abstract

This paper introduces a reinforcement learning based approach to compute optimal interest rate reaction functions in terms of fulfilling inflation and output gap targets. Our data-driven approach incorporates nonlinear economy structures and uncertainty about these. We use quarterly U.S. data from 1987:Q3-2023:Q2 to estimate (nonlinear) transition equations, train optimal policies and perform counterfactual analyses to evaluate them, assuming that the transition equations remain unchanged. All of our resulting policy rules outperform other common rules as well as the actual federal funds rate. Given a neural network representation of the economy, our optimized nonlinear policy rules reduce the central bank's loss by over 27 %.

Keywords: Optimal Monetary Policy; Reinforcement Learning; Artificial Neural Network; Machine Learning; Reaction Function (JEL: C45, C61, E52, E58)

Acknowledgments

The paper represents the authors' personal opinions and does not necessarily reflect the views of the Deutsche Bundesbank or the Eurosystem. Any errors are ours. We would like to thank Klaus Adam, Ajit Desai, Jesús Fernández-Villaverde, Refet Gürkaynak, Uwe Hassler, Josef Hollmayr, Andreas Joseph, Sören Karau, Malte Knüppel, Philipp Lieberknecht, Anika Martin, Alexander Meyer-Gohde, Emmanuel Mönch, Frank Schorfheide, Nikolai Stähler, Harald Uhlig, Volker Wieland, participants at the 23rd INFER Annual Conference, the International Conference on Economic Modeling and Data Science 2021, the 30th Annual Symposium of the Society for Nonlinear Dynamics and Econometrics, the 28th International Conference Computing in Economics and Finance and the 24th Applied Economics Meeting for helpful comments and discussions. During the preparation of this work, the author used ChatGPT 4o to improve language. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication. Declarations of interest: none.

[*] Corresponding author. *Email address*: natascha.hinterlang@bundesbank.de

Preprint submitted to Economic Modelling *October 2, 2024*

Optimal Monetary Policy using Reinforcement Learning

Abstract

This paper introduces a reinforcement learning based approach to compute optimal interest rate reaction functions in terms of fulfilling inflation and output gap targets. Our data-driven approach incorporates nonlinear economy structures and uncertainty about these. We use quarterly U.S. data from 1987:Q3-2023:Q2 to estimate (nonlinear) transition equations, train optimal policies and perform counterfactual analyses to evaluate them, assuming that the transition equations remain unchanged. All of our resulting policy rules outperform other common rules as well as the actual federal funds rate. Given a neural network representation of the economy, our optimized nonlinear policy rules reduce the central bank's loss by over 27 %.

Keywords: Optimal Monetary Policy; Reinforcement Learning; Artificial Neural Network; Machine Learning; Reaction Function (JEL: C45, C61, E52, E58)

1. Introduction

A simple linear rule can effectively describe actual monetary policy decisions in the early 1990s, as demonstrated by Taylor (1993). The evaluation of such rule-based policies in terms of optimality and robustness has become a central topic in the literature (see, e.g., Taylor and Williams, 2010). This paper introduces a machine learning-based approach to a monetary policy optimization problem under the central bank's imperfect knowledge.

In the first step, we estimate macroeconomic transition equations for inflation and the output gap using quarterly U.S. data from 1987:Q3 to 2023:Q2. We consider two cases: a linear economy, estimated via a structural vector autoregression (SVAR), and a nonlinear economy, approximated by artificial neural networks (ANNs). The latter can capture unspecified nonlinearities due to their universal approximator property, improving the data fit by 21% compared to the SVAR representation. Notably, the ANN model enhances the description of inflation developments since COVID-19.

In the second step, we model monetary policy as a reinforcement learning (RL) problem, assuming the estimated relations are given. The central bank aims to smooth inflation and the output gap by setting the nominal interest rate according to a reaction function. During RL, the central bank learns its optimal reaction function by interacting with the economic environment, represented by the estimated transition equations. Importantly, we assume that while the central bank observes the current state of the

Preprint submitted to Economic Modelling *October 2, 2024*

economy, it does not know these transition equations. This assumption is motivated by the recent shift of central banks from a forward guidance regime to a more *data-driven* approach, emphasizing the need for monetary policy decisions to wait for the most recent incoming data.[1] The RL method captures this step-by-step explorative behavior under uncertainty about the true underlying model, making it a suitable tool for analyzing optimal monetary policy.

Specifically, RL is at least as flexible as the standard optimal regulator problem. Both approaches can address nonlinearities in the economy and the policy function, as well as uncertainty in the form of shocks and model parameters. However, the RL algorithm goes a step further by assuming that the agent does not know the underlying model at all. Instead, it learns to maximize the expected long-term reward given action and observational variables. The advantage of RL increases with the complexity of the economy, as the approach is prone to the curse of dimensionality, leaving ample room for future research. Our introductory setup is relatively small-scale to provide a proof of concept and facilitate economic interpretability.

In a dynamic historical counterfactual setup, we find that the optimized policy rules yield inflation and output gap series much closer to the targets compared to the actual and prescribed paths of common policy rules from the literature and the Federal Reserve's monetary policy report (MPR). Relying on the nonlinear model transition equations, a nonlinear reaction function performs best in the counterfactual exercise. While the optimized linear rules reduce the central bank's loss by 15%, the nonlinear reaction functions boost the improvement to over 27%. This is because they can react flexibly to different combinations of inflation and output gap values. In particular, partial dependence plots of the optimized rules imply a plateau at values close to the targets and stronger reactions if the output gap is negative and inflation is below target. In a static counterfactual, where we derive the RL-suggested interest rates in response to actual inflation and output gap values, we find evidence that the Federal Reserve held interest rates *too low for too long* prior to the financial crisis. Moreover, the optimized rules suggest an earlier and sharper increase in interest rates since COVID-19.

One caveat of the analysis is that the quantitative results of the historical counterfactual depend on the estimated model equations, which are assumed to be given. Allowing for agents with rational expectations could alter the results since changes in the reaction function would also imply changes in the transition equations' parameters, as pointed out by Lucas (1976). However, since we assume that the central bank does not know the transition equations, it seems implausible that private agents behave more rationally than the central bank. To analyze the sensitivity of the RL-optimized linear reaction functions with respect to model uncertainty, we conduct a model comparison exercise using 11 dynamic stochastic general equilibrium (DSGE) models. The results indicate that the RL-optimized policy rules within the ANN economy are also quite ro-

[1] See, e.g., recent press release `https://www.federalreserve.gov/monetarypolicy/files/monetary20240320a1.pdf`.

bust, yielding comparable stability results measured by unconditional variances to the common policy rules.

Our paper relates to different streams of the literature. Generally, it contributes to the literature on optimal monetary policy reaction functions. By *optimal*, we mean optimal with respect to a given central bank mandate, in contrast to Ramsey optimality (see, e.g., Debortoli et al., 2019 for the link between both). Svensson (1997) and Woodford (2001) discuss the standard approach, where the central bank faces a linear-quadratic (L-Q) optimization problem, i.e., it has a quadratic loss function and linear constraints. Specifically, our paper closely connects to studies that deviate from the L-Q framework by assuming nonlinear constraints (e.g., Orphanides and Wieland, 2000, Schaling, 2004, Dolado et al., 2004, 2005, Adam and Billi, 2006). Moreover, we connect to the literature on monetary policy and information frictions (see, e.g., Sargent et al., 2006, Gaspar et al., 2006, Reis, 2009, Benchimol and Bounader, 2023, and Benchimol, 2024). However, most of these studies (except for Sargent et al., 2006) typically assume perfect knowledge of the central bank and deviate from rational expectations on the private sector side, while we assume imperfect knowledge of the central bank.

Our work also relates to the expanding literature on monetary policy rules versus discretion following Taylor (1993) (see also Nikolsko-Rzhevskyy et al., 2021 and Cochrane et al., 2019 for more recent results). The issue of model and parameter uncertainty in the context of optimal monetary policy has been tackled by many authors using Bayesian and robust control-related methods along the lines of Hansen and Sargent (2001) (see, e.g., Wieland, 2000, Tetlow and Von zur Muehlen, 2001, Levin et al., 2003). Regarding robustness analyses using a comparative DSGE model approach, we further rely on Wieland et al. (2012, 2016).

Regarding the methodological part, we apply the deep deterministic policy gradient (DDPG) algorithm by Lillicrap et al. (2015). See also Botvinick et al. (2019) for a survey on the development and general applications of deep RL. Several papers consider RL in the areas of operations research, game theory, and (public) finance (see also Charpentier et al., 2021 for a recent survey). For example, Castro et al. (2021) use RL to approximate banks' optimal liquidity provision in a given payment system, while Zheng et al. (2020) rely on RL to compute optimal tax policies that trade off equality and productivity. Moreover, Shi (2021) and Chen et al. (2021) consider RL as a method for replacing the assumption of rational expectations in structural models. They allow households to learn their optimal policies over time using deep RL. However, to the best of our knowledge, this is the first paper applying (deep) RL in the context of optimal monetary policy.

The remainder of the paper is organized as follows. Section 2 describes the reinforcement learning methodology and the data we use. In Section 3, we present estimation results of the economy, the optimized policy rules, and counterfactual analyses. Section 4 discusses the results, and Section 5 concludes.

2. Methodology and Data

Consider a central bank with a dual mandate of price stability and maximum employment, akin to the Federal Reserve. The primary policy instrument is the nominal interest rate. However, the economic environment in which the central bank operates is fraught with uncertainties. Recent examples include the COVID-19 crisis, characterized by an unknown mix of supply and demand shocks in real-time, and the energy crisis following the Russian war of aggression. Common forecasting models failed to predict the true developments, particularly in inflation. This led central banks to transition from a forward guidance regime to a more *data-driven* approach, emphasizing the need for monetary policy decisions to wait for the most recent incoming data.[2] The reinforcement learning (RL) method captures this step-by-step explorative behavior under uncertainty about the true underlying model, making it a suitable tool for analyzing optimal monetary policy. This section describes the specific RL structure, the choice of hyperparameters, and the data.

2.1. Structure

The general idea of RL is illustrated in Figure 1. The *agent* (central bank) interacts with an unknown environment E, receiving a vector of *observations* x_t (containing inflation π_t and output gap y_t, and possibly lags thereof), taking an *action* i_t (nominal interest rate setting), and receiving a *reward* signal r_t, depending on the deviation of observations from targeted values. Given observations and reward, the agent evaluates past behavior and adapts its action. Through this iterative process over multiple discrete time steps T, an optimal *policy* (monetary policy reaction function) given the environment and reward signal is learned. The following sections describe all elements of the RL scheme in more detail.

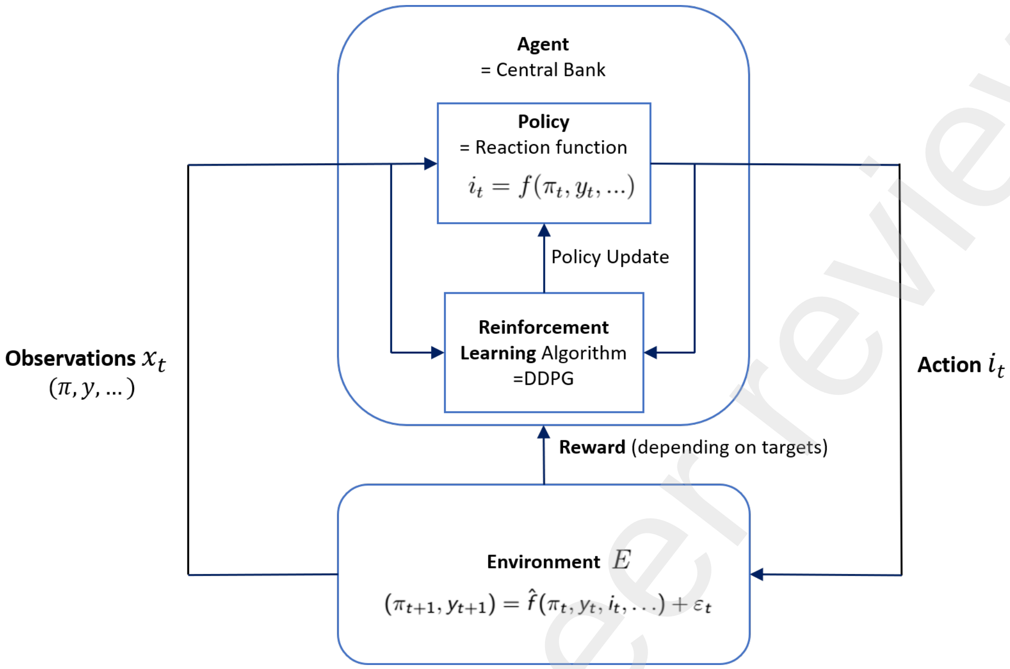
2.1.1. The Environment

As shown in Figure 1, RL requires the provision of an environment E, which determines the next observations in response to the agent's actions. In our application, the environment represents the economy excluding the central banking part. We approximate this part of the economy by a two-equation system including the variables inflation, π_t, and the output gap, y_t. Regarding the variable choice for the whole economy, we consider the basic three-equation New Keynesian model (NKM) of Rotemberg and Woodford (1997). The NKM consists of an aggregate demand equation (dynamic investment-saving (IS) curve), an aggregate supply equation (New Keynesian Phillips curve), and a central bank reaction function.

As stated above, we aim for a rather small-scale setup. The reason for this specific setup is twofold. First, inflation and output gap are the key variables central banks are concerned with. Second, by restricting the observation set to these variables, we can

[2] See, e.g., recent press release `https://www.federalreserve.gov/monetarypolicy/files/monetary20240320a1.pdf`.

Figure 1: Reinforcement Learning Scheme



Agent
= Central Bank
Policy
= Reaction function
$i_t = f(\pi_t, y_t, \ldots)$
Policy Update
Reinforcement Learning Algorithm =DDPG
Observations x_t
$(\pi, y, \ldots)$
Action i_t
Reward (depending on targets)
Environment E
$(\pi_{t+1}, y_{t+1}) = \hat{f}(\pi_t, y_t, i_t, \ldots) + \varepsilon_t$

directly compare the RL-optimized policy rules to well-established policy rules like the Taylor rule. It also simplifies economic interpretability. Hence, our approach should serve as a proof of concept, which can be extended to a much larger information set of the central bank in future research, as the approach is prone to the curse of dimensionality.

Since the task of the environment is to provide the next period's states based on the current economic state and current policy behavior, it can also be interpreted as a forecasting setup. The general form of the economy transition equations is given by:

$$y_t = \hat{f}^y(y_{t-1}, y_{t-2}, \pi_t, \pi_{t-1}, \pi_{t-2}, i_t, i_{t-1}, i_{t-2}) + e_t^y \tag{1}$$
$$\pi_t = \hat{f}^\pi(y_t, y_{t-1}, y_{t-2}, \pi_{t-1}, \pi_{t-2}, i_t, i_{t-1}, i_{t-2}) + e_t^\pi, \tag{2}$$

where the output gap, y_t, and inflation, π_t, depend on lagged and contemporaneous values of themselves as well as on the nominal interest rate, i_t, and lags thereof. Regarding the specific functional form of the equations, we consider two scenarios. First, we assume a standard linear model structure, i.e., Equations (1) and (2) collapse to the well-known reduced form representation of a vector autoregression (VAR). Second, we use artificial neural networks (ANNs) to approximate the economy, allowing for non-linear relationships among variables while being agnostic about the specific functional forms.

Linear Economy. When estimating the linear economy, we use the general form given above, with $\hat{f}^m$ representing a simple linear function of the respective inputs collected in vector s_t^m:

$$\hat{f}^m = C^m + \alpha^{m\prime} s_t^m, \tag{3}$$

where C^m and α^m, $m \in \pi, y$, represent the constant and the vector of coefficients, respectively. For the historical counterfactual analysis later on, we transform the reduced form to a structural (SVAR) representation. Following, e.g., Rotemberg and Woodford (1997), we directly estimate the recursive SVAR equation-by-equation using OLS, assuming that demand pressures affect inflation contemporaneously as in, e.g., Orphanides (2003) or Orphanides and Wieland (2000). This recursive structure implies that the output gap reacts to inflation only with a lag of one period, while inflation depends on the current level of the output gap. Moreover, while there is no direct effect of the nominal interest rate upon inflation and the output gap, the central bank reacts to the current levels of both, as will be shown in the following section on the policy function. Aiming for a parsimonious model, we start with an SVAR(2) specification and drop all second lags that are insignificant at the 10% significance level, which yields the following input vectors:

$$\begin{aligned} s_t^y &= (y_{t-1}, \pi_{t-1}, i_{t-1}, i_{t-2}) && (4) \\ s_t^\pi &= (y_t, y_{t-1}, \pi_{t-1}, \pi_{t-2}, i_{t-1}). && (5) \end{aligned}$$

By restricting our SVAR in this way, we aim for a parsimonious model structure driven by statistical evidence. The information criteria (BIC and AIC) favor the restricted version given in Equations (6) and (7) over the SVAR(2) specification:

$$\begin{aligned} y_t &= C^y + a_{y,1}^y\, y_{t-1} + a_{\pi,1}^y\, \pi_{t-1} + a_{i,1}^y\, i_{t-1} + a_{i,2}^y\, i_{t-2} + \varepsilon_t^y && (6) \\ \pi_t &= C^\pi + a_{y,0}^\pi\, y_t + a_{y,1}^\pi\, y_{t-1} + a_{y,2}^\pi\, y_{t-2} + a_{\pi,1}^\pi\, \pi_{t-1} + a_{\pi,2}^\pi\, \pi_{t-2} + a_{i,1}^\pi\, i_{t-1} + \varepsilon_t^\pi. && (7) \end{aligned}$$

Nonlinear Economy. There is a growing literature on possible nonlinear relationships within the economy and the consequent policy implications. While some consider a convex Phillips or IS curve and the effect on optimal monetary policy (see, e.g., Schaling (2004), Dolado et al. (2004, 2005), and Tambakis (2009)), recent studies try to explain a flattening of the Phillips curve after the global financial crisis (see, e.g., Watson (2014), Coibion and Gorodnichenko (2015), Ball and Mazumder (2019)). Hence, we also consider a nonlinear economy but are agnostic about its specific functional forms. To do so, we estimate the transition equations using ANNs in a *supervised* manner, i.e., given actual values from the data, a training algorithm learns the respective relationship between the variables by periodically updating the network's parameters.

The ANN representation for $\hat{f}^m$, $m \in \pi, y$, is given by:

$$\hat{f}^m = b_0^m + \sum_{j=1}^{h} \nu_j^m G\,(\omega_j^{m'} s_t^m + b_j^m), \tag{8}$$

which applies a nonlinear transformation to the input state s_t^m. The parameters collected in the vectors ω_j, ν_j, $j = 1, ..., h$ and b_i, $i = 0, \ldots, h$, are known as weights and biases to be estimated. We keep the previous recursive structure fixed and assume that y_t and π_t are unknown functions of the same variables as in the linear economy to ensure a fair comparison. Regarding the network structure, we opted for a simple feed-forward network since its performance is satisfactory. In principle, more advanced network architectures like recurrent neural networks or transformers are also integrable.

The number of hidden units h represents a hyperparameter that we determine by dividing the sample into a training and validation set,[3] where we use 15% of the observations for validation. Looping over one to ten hidden units using 30 different random initial weights yields 2 and 8 hidden units for the output gap and inflation equation, respectively. After fixing the optimal number of hidden units, we take the set of initial weights and biases that corresponds to the lowest mean squared error.

We wish to make clear at the outset that in our RL setup, the estimated transition equations are taken as given, i.e., changes in the reaction function do not affect parameters of the transition equations as would be the case with models including rational expectations. Hence, the Lucas critique applies. However, given that only the parameters of the policy function are optimized during RL, one would expect smaller adjustments in expectations than under a complete shift in the monetary policy regime, i.e., in target values. Moreover, there is a growing literature on optimal monetary policy under nonrational private agents (e.g., Orphanides and Wieland, 2000, Gaspar et al., 2006, Reis, 2009, Benchimol and Bounader, 2023, Benchimol, 2024). In our RL case, even the central bank does not know the underlying transition equations. Assuming private agents to be more rational than the central bank seems implausible to us. Alternatively, one may think of private agents adapting expectations only very slowly over time compared to the central bank. Of course, this is a critical assumption. However, as the degree of rationality is uncertain, the estimated transition equations can still serve as a useful benchmark environment to introduce the RL concept.

2.1.2. The Agent

The RL policy function is a mapping of observed economic states into actions. In RL nomenclature, we set up a *critic*, which is equivalent to an approximate value function representing the expected long-term reward of the present policy and thus drives

[3] Note that the validation set is also used to prevent the algorithm from overfitting the training data by introducing an early stopping mechanism. Training is stopped when the mean squared error of the validation set fails to improve or remains the same for six consecutive epochs.

the policy parameter updating. Further, we define an *actor*, representing the central bank policy function, describing the nominal interest rate setting behavior in response to observations from the environment. Concerning the critic, we use nonlinear neural networks to approximate the value function. The actor is given by a simple linear or nonlinear neural network depending on the structure of the economy. Through the training process, the functional parameters are updated to maximize the expected long-term reward (minimize the long-term loss).

The Policy. Defining the policy representation includes delimiting the observation and action spaces. While the details of the economic structure, i.e., parameters and functional forms of (6)-(7) or (8), are unknown during training, the agent observes certain state variables that serve as inputs to the policy function. We consider two different specifications concerning the dimension of the observation space. The first setup shown below in equation (9) is supposed to mirror the standard Taylor (1993) type monetary policy inputs, while the second setup (10) additionally contains one lag of each variable. By using these specifications, on the one hand, we aim for a fair comparison to standard Taylor-type rules. On the other hand, lags of inflation and output gap in the policy function are shown to produce robust stabilizing behavior (Hawkins et al., 2015) and therefore constitute our second choice.

$$x_t^1 = (y_t, \pi_t) \tag{9}$$
$$x_t^2 = (y_t, y_{t-1}, \pi_t, \pi_{t-1}) \tag{10}$$

The action space is one-dimensional and real-valued. We start our analysis employing linear neural network policy structures. This approach allows for a direct comparison with other common linear policy rules before turning to a more flexible, nonlinear functional form. The structural form of the resulting policy function P_t is given by:

$$P_t = i_t = \alpha_0 + \sum_{j=1}^{q} \delta_j G(\beta_j' x_t^z + \alpha_j) \tag{11}$$

with x_t^z, $z \in 1, 2$ being the vector of observations from (9) or (10), respectively, and $G(\cdot)$ being a monotonically bounded increasing transfer function. Equation (11) is the representation of a single-hidden-layer feed-forward neural network as used to approximate the economy in (8).[4] In the linear policy case, $q = 1 = \delta_j = 1$ and $G()$ collapses to the *purelin* transfer function, which simply maps the input value onto itself ($purelin(n) = n$). The response coefficients are then given by β_π^l and β_y^l, where $l \in 0, 1$ refers to contemporaneous and lagged variables, respectively. For the nonlinear case,

[4] We do not consider multiple hidden layers for our rather simple application. However, it is possible to use such deep neural networks within the RL framework in general.

we use the hyperbolic tangent sigmoid function as before. The parameters to be optimized by the RL algorithm are the weights β_j and δ_j, $j = 1, ..., q$ and the biases α_j, $j = 0, ..., q$, where q denotes the number of hidden units that has to be determined in advance as explained in the following section.

The Objective. To optimally adjust the policy coefficients, it is essential to determine the respective action value function, which serves as a performance measure for policy interventions and thus forms the basis for policy updates:

$$Q^P(x_t, i_t) = \mathbb{E}[R_t | x_t^z, i_t] \tag{12}$$

with

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r_t(x_i^z, i_i). \tag{13}$$

The function Q^P describes the expected return after taking action i_t, observing state x_t, and subsequently following policy P_t. This recursive relationship is grounded in the Bellman equation. The return R_t is defined as the sum of discounted future rewards, with a typical discount factor of $\gamma = 0.99$ (cf. Svensson (2020)). The *critic* approximates (12) using an artificial neural network (ANN), where θ^Q encompasses all connection weights and biases of the network. The set of observable variables x_t^z enters the critic via the observation path, while the control variable i_t is included through the action path. Both paths are concatenated into a common path, and its output is the expected long-term reward based on the observed state and action, i.e., equation (12).

In defining the objective of our agent, we adhere to the Federal Reserve's mandate. According to the Federal Open Market Committee's (FOMC) statement on longer-run goals and monetary policy strategy,[5] an inflation rate of 2% is "most consistent over the longer run with the Federal Reserve's statutory mandate." Additionally, the Federal Reserve aims to promote maximum employment, with its specific level allowed to vary over time. These two objectives are generally seen as complementary. Consequently, we rely on the standard quadratic reward function given by:

$$r_t(x_t^z, i_t) = -\omega_\pi \left(\pi_{t+1} - \pi^*\right)^2 - \omega_y \, y_{t+1}^2 \tag{14}$$

with equal[6] $\omega_\pi = \omega_y = 0.5$ and $\pi^* = 2\%$.[7] We would like to emphasize at this point, that it is generally possible to analyze optimal monetary policy under different

[5] See the FOMC's Longer Run Goals and Monetary Policy Strategy document on `https://www.federalreserve.gov/monetarypolicy.htm`

[6] Equal weights on inflation and the unemployment gap actually translate into a weight of 0.125 on the output gap using Okun's law. We still stick to 0.5, since Debortoli et al. (2019) show that an output gap weight similar to that of inflation improves social welfare in a DSGE model context.

[7] For computational reasons, the continuous reward function given by (14) is accompanied by a second part, which punishes deviations from targets that exceed 2 percentage points: $r_t^{\pi_p} = 10 \cdot r_t^{\pi}$ (if $r_t^{\pi} > 4$) and $r_t^{y_p} = 10 \cdot r_t^{y}$ (if $r_t^{y} > 4$), where the subscript p stands for penalty and r_t^{π} and r_t^{y} denote the squared

loss functions using RL, as well. Only recently, on August 27th 2020, the Fed actually switched to an average inflation target of 2 %. Future research could consider alternative loss functions reflecting average inflation targeting as stated in Svensson (2020).[8]

2.2. *Training Algorithm and Hyperparameters*

We employ the *Deep Deterministic Policy Gradient* (DDPG) algorithm, first presented by Lillicrap et al. (2015).[9] This algorithm builds on the *Deterministic Policy Gradient* algorithm by Silver et al. (2014) and combines the actor-critic approach with *Deep Q Networks* (see Mnih et al. (2013, 2015)). The result is a model-free, online, off-policy actor-critic algorithm using (deep) function approximators.[10] The goal of the learning algorithm is to find an optimal policy that maximizes the expected long-term reward.

Table 1 provides an overview of the DDPG algorithm's individual steps, while each part is explained more formally in Appendix A. Before entering such a training cycle, we must decide on the *critic* network structure. To find the best layer structure, we run the training cycle several times, looping over the number of hidden nodes (one to ten), holding the number constant across the hidden layers of the observation and action path for simplicity. For the linear *actor* version, no further choice is involved.[11] However, when optimizing the nonlinear policy version in (11), we also need to determine the number of hidden units of the actor. In this case, we loop over these *actor* nodes from one to ten, while fixing the critic nodes.

One training cycle consists of different steps. It starts with an initialization phase, and each cycle consists of $M = 500$ episodes in total.[12] Steps *f)* to *m)* are repeated until the agent either fulfills our defined stopping criteria, which is $1.8 < \pi_{t+1} < 2.2$ and $-0.2 < y_{t+1} < 0.2$, i.e., 0.2 percentage points absolute deviations from target values, or

deviations from targets. The penalty rewards are added to (14). This kind of *mixed reward signal* drives the system away from bad states while simultaneously promoting convergence.

[8] Note that the Fed does not specify an explicit averaging period. The goal is only stated as an average inflation of 2 % *over time*.

[9] There exist many different RL algorithms. The specific choice depends on the observation and action spaces, i.e., whether they are discrete or continuous, whether it is based on a value or an action-value function, and how the actor is modeled. We chose the DDPG because it is capable of handling continuous observation and action spaces and, in contrast to other algorithms, it returns one value for the action instead of probabilities of taking each action in the action space. Hence, the term *deterministic* in DDPG refers to the final policy function, which is not stochastic.

[10] The term *model-free* indicates that the environment is not known to the actor. Only a set of observable variables combined with the reward signals influence the action taken. An *online* algorithm interacts with the environment while learning (trial and error principle). *Off-policy* means that the policy function is updated relying on sampled experiences from previous policy functions in the iteration process. By contrast, *on-policy* learning means that it only uses experiences generated by the latest learned policy (behavioral policy).

[11] Remember that the linear version of (11) includes setting $q = 1$.

[12] This means that per critic and actor configuration, the algorithm runs 500 episodes, where each episode represents a different policy function (a different agent) and only one has to be chosen.

the episode stops automatically after a maximum of $T = 12$ quarters.[13]

Table 1: DDPG Algorithm

```
Initialization  a)  Randomly initialize critic network $Q(x, i|\theta^Q)$ and actor $P(x|\theta^P)$
                    with weights $\theta^Q$ and $\theta^P$
                b)  Initialize the target network $Q'$ and $P'$ with weights $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{P'} \leftarrow \theta^P$
                c)  Initialize experience replay buffer $B$
for $m = 1 : M$
                d)  Initialize a random process $\mathcal{N}$ for action exploration
                e)  Receive initial observation state $x_0^z$,    $(z = 1 \text{ or } 2)$
    for $t = 1 : T$
                f)  Select action $i_t = P(x_t|\theta^P) + \mathcal{N}_t$ according to the current policy
                    and exploration noise
                g)  Execute action $i_t$, observe reward $r_t$ and observe new state $x_{t+1}$
                h)  Store transition $(x_t, i_t, r_t, x_{t+1})$ in $B$
                i)  Sample a random minibatch of $N$ transitions $(x_j, i_j, r_j, x_{j+1})$ from $B$
                j)  Set $h_j = r(x_j, i_j) + \gamma Q'(x_{j+1}, P'(x_{j+1}|\theta^{P'})|\theta^{Q'})$
                k)  Update critic by minimizing the loss:
                        $L = \frac{1}{N} \sum_j (h_j - Q(x_j, i_j|\theta^Q))^2$
                l)  Update the actor policy using the sampled policy gradient:
                        $\bigtriangledown_{\theta^P} J \approx \frac{1}{N} \sum_j [\bigtriangledown_i Q(x, i|\theta^Q)|_{x=x_j, i=P(x_j)} \bigtriangledown_{\theta^P} P(i|\theta^P)|_{x_j}]$
                m)  Update the target networks:
                        $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$
                        $\theta^{P'} \leftarrow \tau\theta^P + (1-\tau)\theta^{P'}$
    end for
end for
```

Note: This scheme leans on Lillicrap et al. (2015) and is adapted to our variable and parameter specification. It describes a training cycle of the Deep Deterministic Policy Gradient algorithm. See Appendix A for details.

As indicated in step *g)*, the reward r_t is calculated according to our designed reward function (14) in each step t. To measure performance for one training episode, the episode reward is calculated as the sum of the rewards per step ($ER_m = \sum_t r_t$). During training, we save the best-performing agents according to pre-specified criteria (see Appendix A). Subsequently, we calculate the steady state of each saved agent, representing the long-term equilibrium of the economy, and derive the respective steady state reward according to (14). We then select the agent with the best steady state reward per set of critic nodes. Out of these ten results, we choose the optimal number of critic nodes (and hence the final optimal policy function) according to the same criteria.

[13] We also experimented with smaller bands around the target values, but it produced inferior results.

Tables A.1 and A.2 in the Appendix summarize all network structures and hyperparameters, including the chosen numbers of hidden nodes for the six different cases under investigation. For most of our cases, having one node in the critic network yields the best results. Our steady state reward approach further yields ten and eight nodes in the policy function with observation inputs x_t^1 and x_t^2, respectively.

2.3. Data

In our benchmark analysis, we use quarterly U.S. data from 1987:Q3 to 2023:Q2. The chosen starting point coincides with the appointment of Alan Greenspan as the Federal Reserve's chairman, marking the beginning of the era of (implicit) inflation (and unemployment) targeting (see Goodfriend (2004)). The terminating period was determined by the last available data.

Inflation π_t is measured by the GDP implicit price deflator as the percentage change from one year ago.[14] The output gap y_t is computed as the percentage deviation of actual GDP from its potential. For the latter, we use estimates produced by the U.S. Congressional Budget Office. The nominal interest rate i_t is given by the effective federal funds rate. For periods when the effective lower bound was binding, we use shadow rate estimates of Wu and Xia (2016).[15]

We are aware of the difficulties arising from using ex post revised instead of real-time data in a central bank's reaction function, as mentioned by, e.g., Orphanides (2001). However, we only use the complete data set to estimate the transition equations for inflation and output gap (see 2.1.1). The reaction function itself is not estimated but optimized. Hence, actual values only enter the reaction function in the RL algorithm through the initial observation state of an episode (see step *e)* of Table 1). During the subsequent learning steps, inflation and output gap data are simulated by our estimated economy, drawing random shocks. The central bank only observes the values of π and y, but does not know the nature of the shock.

3. Results

In this section, we begin by presenting the fit of the estimated economy representations. Subsequently, we compare the parameters of our reinforcement learning (RL)-based optimal monetary policy functions to those of other common reaction functions, before turning to historical counterfactual analyses.

[14] The Federal Reserve actually targets inflation measured by the personal consumption expenditure (PCE) index. However, the GDP deflator is closer to the inflation in macroeconomic models that we employ for analyzing the robustness of the optimized rules. For the same reason, we use the output gap instead of the targeted unemployment rate.

[15] Time series were downloaded from FRED and Federal Reserve Bank of Atlanta websites. We performed Augmented Dickey-Fuller (Dickey and Fuller (1979)) and KPSS tests (Kwiatkowski et al. (1992)) that indicated stationarity of the three series.

3.1. Economy Representations

Before presenting our results on RL-based optimal monetary policy, we illustrate the differences between our linear SVAR (6)-(7) and nonlinear (8) economy representations. Table B.1 in the Appendix summarizes the estimation results for the SVAR representation. The Durbin-Watson and Lagrange Multiplier statistics indicate that the error terms (representing the structural shocks) are serially uncorrelated. The drawback of ANNs is that estimated parameters are more difficult to interpret. However, the SVAR representation is restricted by its predetermined linear form and might consequently miss certain dynamics of the actual time series data.

Figure 2 compares the fit of the SVAR model (in red) and the ANN economy (in blue) for inflation and the output gap across different subperiods, specifically pre- and post-COVID-19 crisis. We compute the differences between the fitted and actual time series and plot the squared errors. Larger values can be interpreted directly as a worse fit. While the difference between the SVAR and ANN models is less pronounced for the output gap, the nonlinear model clearly yields a better fit for inflation.

Regarding timing, the results indicate that the ANN outperforms its linear counterpart, especially during crisis periods. For the recession in the early 1990s following the stock market crash in 1989, the years after the Great Financial Crisis (at least for inflation), and more recent years including COVID-19 and the onset of the Russian war of aggression, the ANN yields lower squared errors. While the ANN often cannot improve the fit at the peak, it better matches the subsequent adjustment processes.

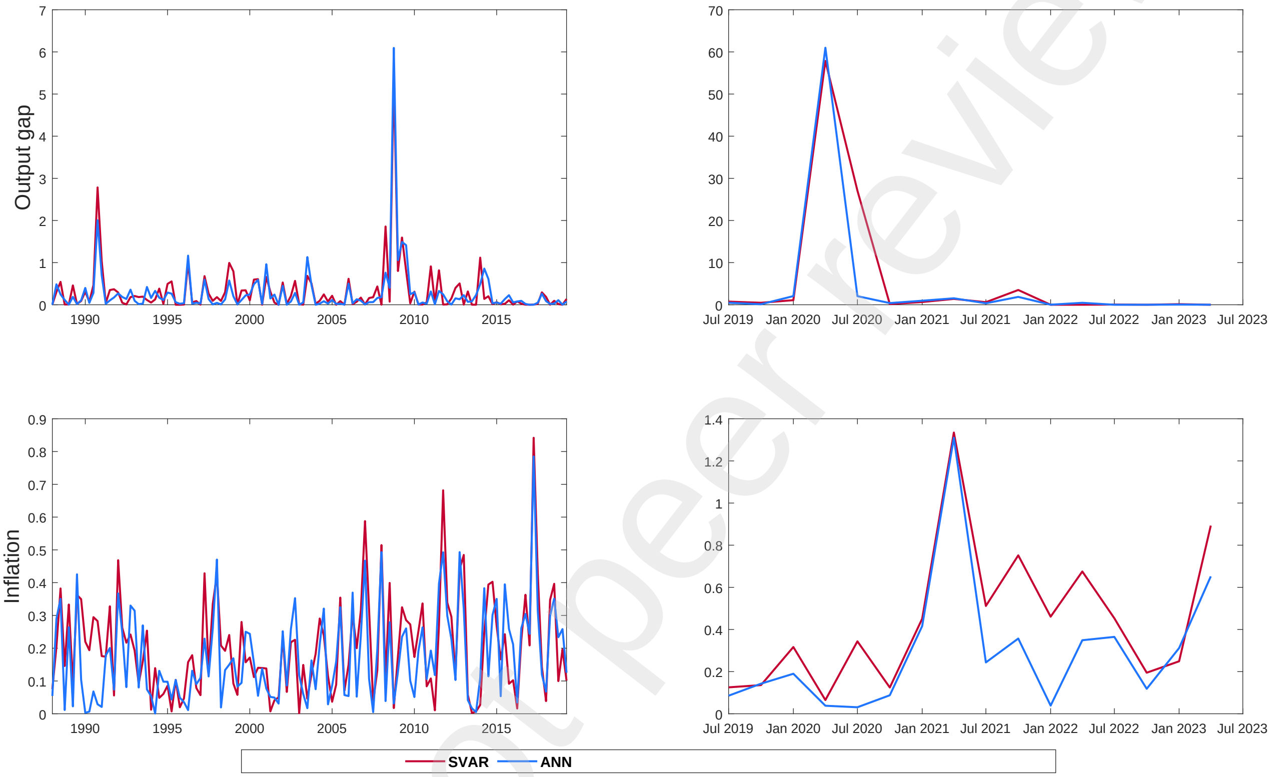
The superior performance becomes even clearer when comparing the mean squared errors (MSE) in Table 2. Using the ANN to approximate the economy, the overall fit of the output gap and inflation variables improves by 20% and 27%, respectively. Averaging over the MSE for the output gap and inflation, the ANN outperforms the SVAR by 21%. Table B.2 in the Appendix provides the respective results for the subperiods, emphasizing the superior performance of the ANN regarding inflation in more recent times. Since we use a validation set (see 2.1.1) to prevent the ANN from overfitting, the result indicates the presence of nonlinearities that cannot be captured by the SVAR model.[16]

Visualizing these nonlinearities is not straightforward. The marginal relationship between input and output variables in an ANN is not constant as in the linear case but depends on the levels of the input variables. Since inflation and the output gap are functions of five and four explanatory variables, respectively, all functional dependencies cannot be plotted. However, we can visualize parts of it using partial dependence (PD) plots. These plots show output predictions against a single or a pair of input variables by marginalizing out the effects of the remaining variables.[17]

[16] The ANN also outperforms the SVAR by a similar magnitude when focusing on the validation set only. Hence, the superiority of the ANN should not be due to overfitting the training sample.

[17] To produce these plots, we build a grid for inflation, output gap, and interest rate from 0:6, -5:3, and -3:7, respectively, and compute the corresponding outputs of the ANN. We then assume constant values

Figure 2: Economy Fit: Squared Errors



Note: This figure shows the squared errors between the actual time series and the fitted values of the SVAR (red) and ANN (blue) model for the output gap (top row) and inflation (bottom row). The two columns separates periods between pre and post COVID19 crisis (Pre Covid: 1987Q3-2019:Q1, Post Covid: 2019:Q2-2023:Q2).

Table 2: Economy Fit: Mean Squared Errors

Representation	MSE Output Gap	MSE Inflation	MSE Total
SVAR	0.935	0.090	0.512
ANN	0.748	0.065	0.407

Note: This table summarizes the mean squared errors of the linear SVAR (6)-(7) and ANN (8) economy representations for the variables output gap, inflation and the overall economy.

Figure 3 represents PD surface plots for inflation and the output gap. As expected, inflation increases with the output gap. For negative values of the nominal interest rate, the slope is quite steep, while it is flatter for values closer to the zero lower bound, supporting the view that the Phillips curve has flattened during these times. For nominal interest rate values larger than 1, inflation decreases with tighter monetary policy as expected. This channel is stronger the larger the output gap. Surprisingly, this is not the case for negative values of the nominal interest rate. Here, we find that inflation increases with the interest rate. Interpreting the shadow rate as a measure of quantitative easing (QE), this would imply that QE is followed by a decrease in inflation instead of an intended increase. One interpretation of this result could be that the central bank reveals its concerns about deflation by implementing QE, actually reducing inflation expectations of private agents. This corresponds to an output gap increasing with the nominal interest rates in negative territory. Larger nominal interest rates only contract output if the interest rate level is high and inflation is low. The output gap further decreases with inflation for low interest rate values, while the relation is flatter for higher interest rate values. Of course, the figure does not represent causal relations, and results for extreme (combinations of) values that are underrepresented in the sample should be interpreted with caution. We must keep in mind that the ANN represents an estimated relationship, driven by our sample observations, which may change over time. Still, the plotted relations are an interesting outcome of our ANN economy that a simple linear representation with a constant partial derivative could not produce.

3.2. Optimized Policy Parameters

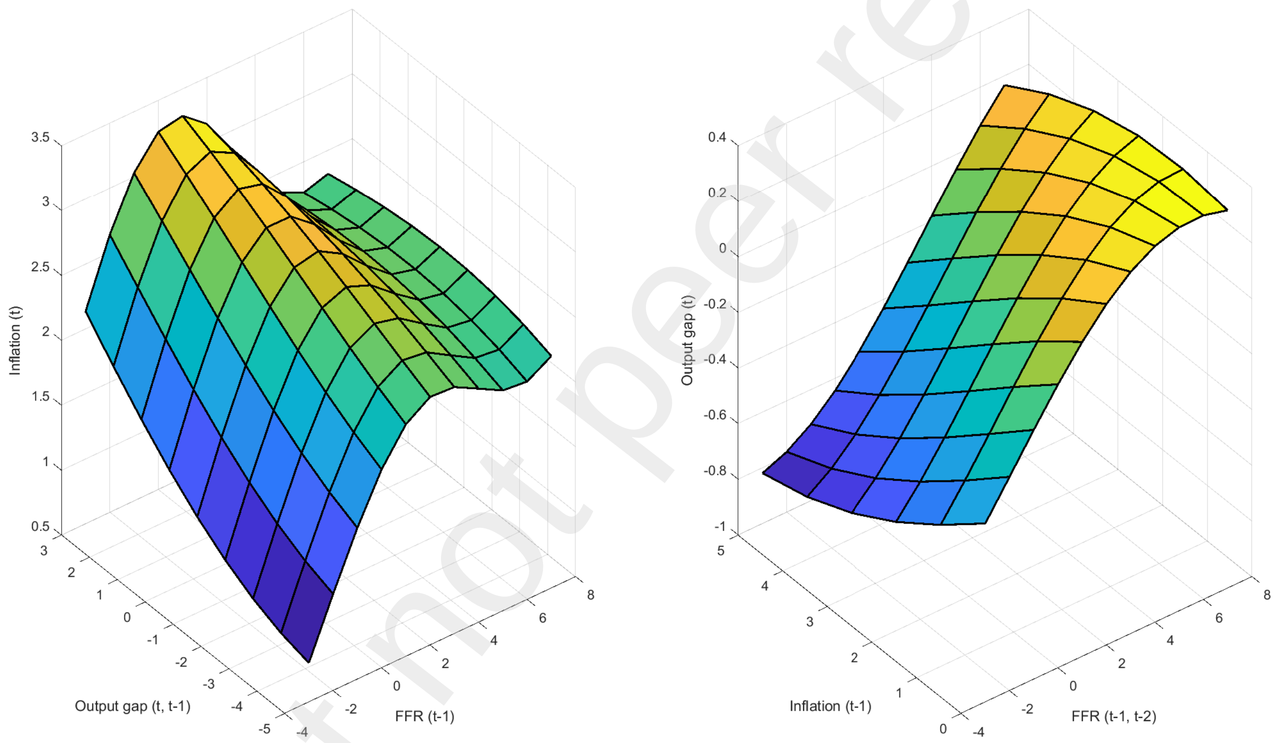
We compute six optimal monetary policy rules in total: two linear ones within the SVAR economy framework and four (two linear and two nonlinear) within the ANN economy. We denote the policies based on our approach by RL, where the subindex specifies whether it is based on the linear (SVAR) or the nonlinear economy (ANN), the input structure, and the functional form (linear vs. nonlinear) used. *No lag* indicates that observations x_t^1 serve as inputs, whereas *one lag* corresponds to x_t^2 (see (9)-(10)).

Linear Policies. Table 3 summarizes the optimized coefficients of the linear policy rules. For comparability, we also present the coefficients of the original Taylor (1993) rule (TR93) and the Balanced-approach (BA), both included in the Federal Reserve's Monetary Policy Report (MPR).[18] Additionally, we consider a rule termed the inflation tilting rule brought up by Nikolsko-Rzhevskyy et al. (2018, 2021) (NPP). The general form of

across lags and marginalize over the remaining variables to reduce the dimensions for the surface plot.

[18] The rules in the MPR actually contain the deviation of unemployment from its natural rate instead of the output gap by using the Okun's law relationship $y_t = 2\,(u_t - u_t^*)$. However, we stick to the version with the output gap. Moreover, we abstract from a time-varying r_t^* and assume a constant value of 2 % that enters the intercept. We also do not consider the price level targeting and the first-difference rules of the MPR. While the former reflects a different monetary policy strategy in general, the latter is not unambiguously defined since it translates previous deviations from the rule into a permanent part.

Figure 3: Partial Dependence Surface Plot - ANN Economy



Note: This figure shows the partial dependence of inflation (left), π_t, on last period's nominal interest rate (FFR), i_{t-1}, and on the output gap, assuming $y_t = y_{t-1}$ and marginalizing over $\pi_{t-1} = \pi_{t-2}$. On the right, it shows the partial dependence of the output gap, y_t, on last period's inflation, π_{t-1}, and nominal interest rate (FFR), assuming $i_{t-1} = i_{t-2}$ and marginalizing over y_{t-1}.

Table 3: Linear Policy Parameters

Policy	α_0	β_π^0	β_π^1	β_y^0	β_y^1
TR 93	1	1.5	-	0.5	-
NPP	0	2.0	-	0.5	-
BA	1	1.5	-	1	
$RL_{SVAR,\,no\,lag}$	1.11	1.60	-	0.39	-
$RL_{SVAR,\,one\,lag}$	1.03	0.78	0.78	0.22	0.22
$RL_{ANN,\,no\,lag}$	1.12	1.61	-	0.38	-
$RL_{ANN,\,one\,lag}$	1.11	0.83	0.84	0.14	0.14

Note: As introduced in the general policy function structure in equation (11), α_0 is the constant term, with $\alpha_0 = r^* - (\beta_\pi - 1)\pi^*$, and r^* denoting the long-run equilibrium real interest rate and π^* representing the inflation target of 2 %. β_π^l is the inflation and β_y^l the output gap coefficient with $l = 0$ indicating the contemporaneous period and $l = 1$ the first lag.

these rules is given by $i_t = r^* + \beta_\pi^0 (\pi_t - \pi^*) + \beta_y^0 y_t$, where r^* and π^* denote the long-run equilibrium real interest rate and the inflation target, respectively.

Interestingly, the RL-optimized linear policy parameters are robust across specifications. All of them have an inflation coefficient between 1.6 and 1.7 ($\beta_\pi = \beta_\pi^0 + \beta_\pi^1$). Hence, they respond more aggressively to deviations of inflation from its target value compared to TR93 and BA and less strongly than NPP. The output gap reaction coefficient falls in the range of 0.3-0.4 ($\beta_y = \beta_y^0 + \beta_y^1$), which is mildly lower than those of TR93 and NPP, and significantly lower than that of BA. Using the intercept relation $\alpha_0 = r^* - (\beta_\pi - 1)\pi^*$, one can further infer a value for the inflation target π^* or the equilibrium real interest rate r^* by holding one of the two constant. Assuming $\pi^* = 2$, the RL-implied r^* ranges between 2.1% and 2.5%, which is slightly higher than the 2% implied by the other rules.

Nonlinear Policies. As a final step, we allow the policy function to be of a nonlinear form as shown in (11) with $G(\cdot)$ being the hyperbolic tangent transfer function. The optimal agents with and without lags are denoted by $RL_{ANN,,one,lag,,nonlin}$ and $RL_{ANN,,no,lag,,nonlin}$, respectively. The coefficients of ANNs are no longer directly interpretable. However, we can still investigate the relationship between the input variables and the nominal interest rate implied by the ANN using partial dependence (PD) plots.

Figure 4 represents PD surface plots for $RL_{ANN,,no,lag,,nonlin}$ (left) and $RL_{ANN,,one,lag,,nonlin}$ (right). The former rule has only two inputs, π_t and y_t. Hence, the inputs' influence on the output variable can be illustrated by a three-dimensional surface plot without the need for marginalizing over other variables. For comparison, we add the optimized linear policy $RL_{ANN,,no,lag}$ to the plot. For the PD plot of $RL_{ANN,,one,lag,,nonlin}$, we hold values constant over the lags of inflation and output gap.[19] Our first observation is that even though we do not impose a zero lower bound constraint on the interest rate, the nonlinear RL rules rarely enter negative territory on our grid. The rule without lags prescribes negative values only for a combination of very low inflation and output gap values. The rule with lags completely stays in the positive domain. This corresponds to the fact that for most of the region, the nonlinear rules imply larger interest rate values than their linear counterparts. The federal funds rate (FFR) under the linear rules exceeds the nonlinear rules mainly for large inflation values. One reason could be that the nonlinear rules stabilize inflation closer to the target, such that they have never encountered larger inflation values during training.

Given that the output gap is closed, i.e., actual GDP equals potential such that $y_t = 0$, the FFR increases in both cases, reaching a value of about 7 at $\pi_t = \pi^* = 2$. As expected, the implied interest rate declines on average when the output gap falls, especially in negative territory and when combined with low inflation values. The profile is rather flat when the output gap is positive. Figure 5 shows PD line plots for inflation and output gap in each case by averaging over the respective remaining variable. Graphically, these lines represent average cross-sections of the nonlinear policies in Figure 4. On average, prescribed interest rates increase with inflation and output gap. However, results for both nonlinear rules suggest that it is optimal not to increase the interest rate if the output gap is sufficiently large. Recall, however, that the algorithm may have never encountered some of the extreme values. It can only provide locally optimal solutions.

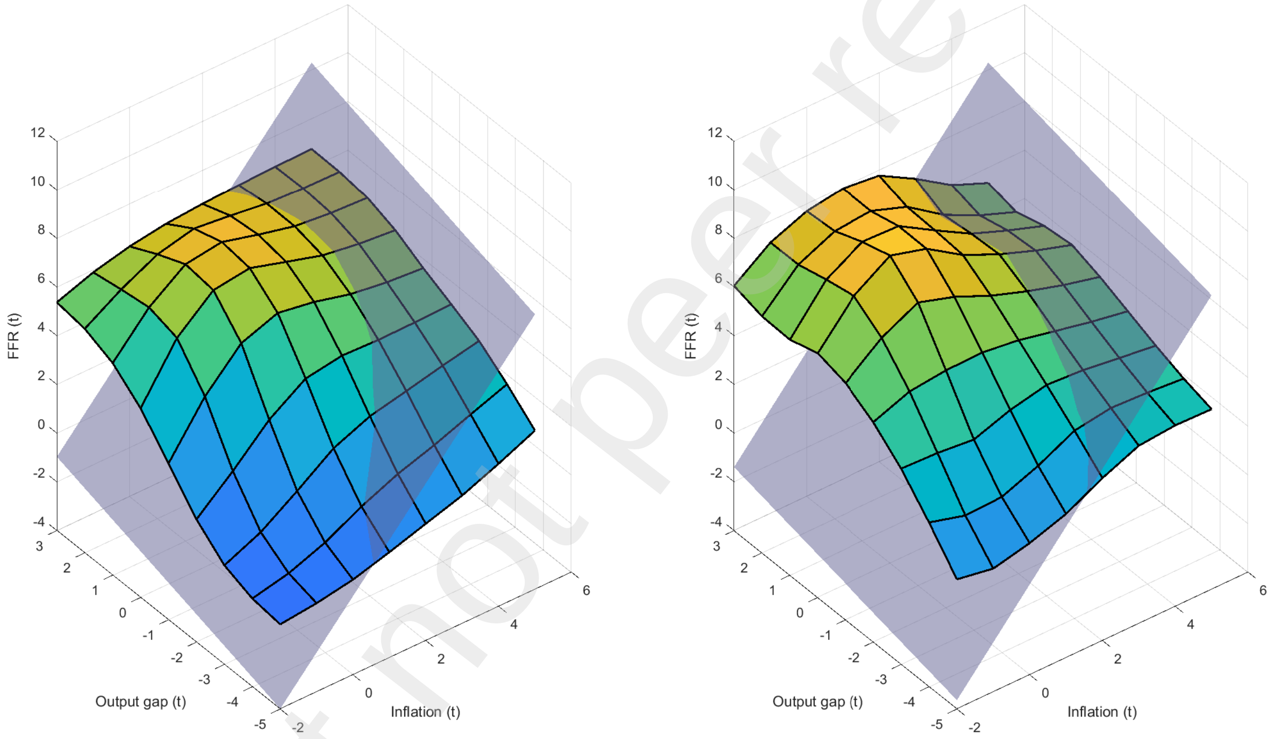
3.3. *Historical Counterfactuals*

To evaluate our RL-based reaction functions, we compare their performance with the actual interest rate setting behavior of the Federal Reserve and alternative rules. Often, different policy rules are compared using a static setup, where data on inflation and the output gap are simply plugged into each rule without considering any feedback mechanisms. However, such an analysis does not allow for conclusions on which policy is best suited to reach target values. Therefore, we first conduct counterfactual analyses that take into account the dynamics of inflation, the output gap, and the interest rate, as well as feedback effects (*Historical Counterfactual, HCF*). This analysis predicates: starting with the historical situation in 1987, what would have happened if the central bank had followed the respective policy rules? Results for a static exercise are then added to set up propositions about which central bank behavior would have been favorable in certain historical situations (*Static Counterfactual, SCF*).

Within the HCF, similar to Primiceri (2005), we use our estimated transition equations (6)-(7) or (8) and the structural shocks thereof $\varepsilon_{it}, i = 1, 2$ to simulate the economy under different reaction functions. Specifically, we replace the interest rate equation of the dynamical system with the respective (optimized) policy rules, while keeping the estimated equations for inflation and the output gap unchanged. We focus here on the

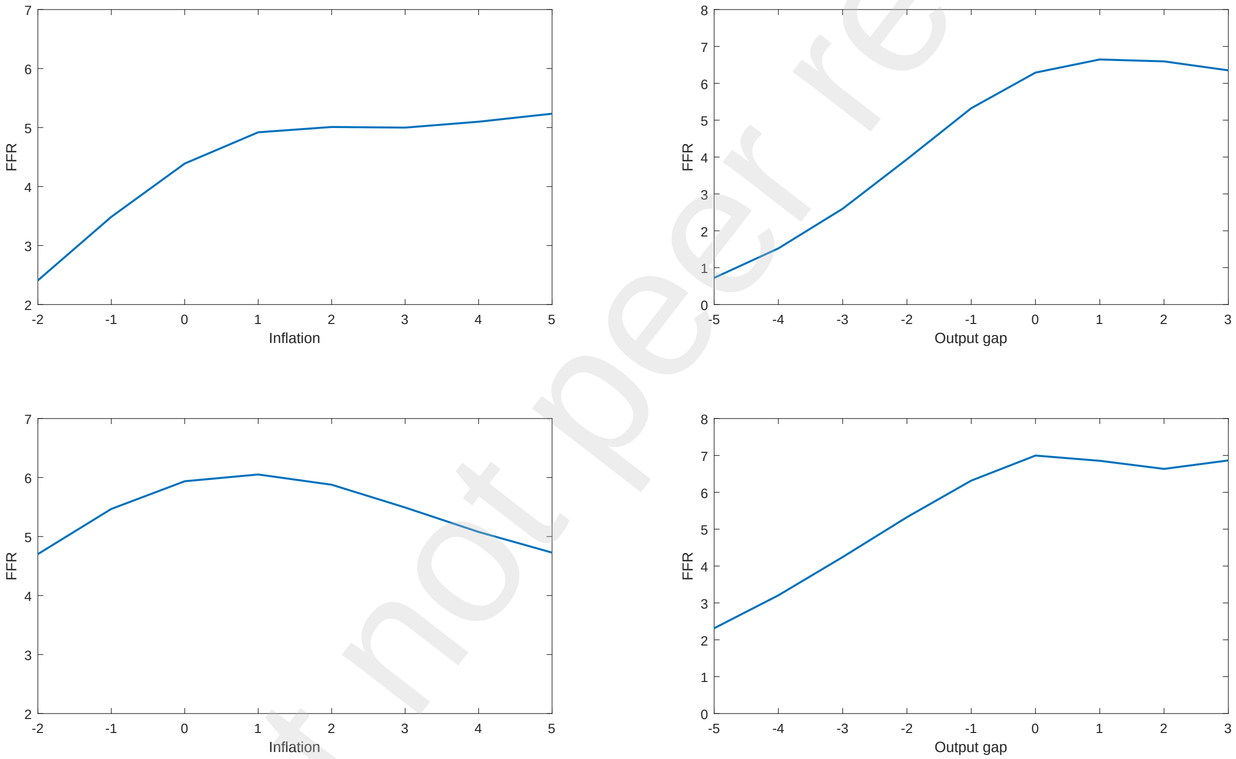
[19] Note that since inflation and the output gap are (auto)correlated, the true partial relationship can only be approximated. The PD plots further rely on the assumption that each input combination is equally likely.

Figure 4: Partial Dependence Surface Plot - Nonlinear vs. Linear RL Rules



Note: This figure shows the partial dependence of the nominal interest rate, i_t, (FFR) on inflation, π_t, and on the output gap, y_t, under $RL_{ANN,\,no\,lag,\,nonlin}$ (left). On the right, it shows the partial dependence of the FFR under $RL_{ANN,\,one\,lag,\,nonlin}$, on inflation and on the output gap, assuming $\pi_t = \pi_{t-1}$ and $y_t = y_{t-1}$. The transparent plains represent the corresponding linear counterparts $RL_{ANN,\,no\,lag}$ and $RL_{ANN,\,one\,lag}$, respectively.

Figure 5: Partial Dependence Line Plots - Nonlinear Rules



Note: This figure shows the partial dependence of the nominal interest rate (FFR) under $RL_{ANN, no\,lag, nonlin}$ (left column) on inflation, π_t, (upper left) and on the output gap, y_t, (upper right), marginalizing over the respective remaining variable. The right column shows the partial dependence of FFR under $RL_{ANN, one\,lag, nonlin}$ on inflation (bottom left) and on the output gap (bottom right) , assuming $\pi_t = \pi_{t-1}$ and $y_t = y_{t-1}$ and marginalizing over the respective remaining variable.

counterfactuals using the ANN economy (8) since this is our preferred specification. Results for the linear SVAR economy are relegated to Appendix C. The dynamic counterfactual simulation period lasts from 1987:Q3 to 2023:Q2.

Since we have to take the estimated structural parameters of the economy as given and unchanged, the Lucas (1976) critique applies, i.e., the behavior of rational and forward-looking private agents might be different when they take the change of policy into account. However, we assume that the central bank as well as the private sector exhibit limited rationality and hence believe that the effects of the Lucas critique are rather small, since we do not compare policies from totally different regimes. By contrast, we only change the policy within a period where inflation targeting was already practiced, using the same technique as in, e.g., Primiceri (2005) and Sims and Zha (2006). Hence, possible behavior modifications of private agents should be minor. Certainly, future research should experiment with the economy framework, possibly incorporating a role for rational expectations. Nevertheless, this section's exercise can still be interpreted as a proof of concept for the RL approach.

ANN Economy & Linear Policy. Figure 6 shows the counterfactual time series for the interest rate, inflation, and output gap, respectively, for the different linear reaction functions. All simulations are based on the estimated ANN economy in (8).

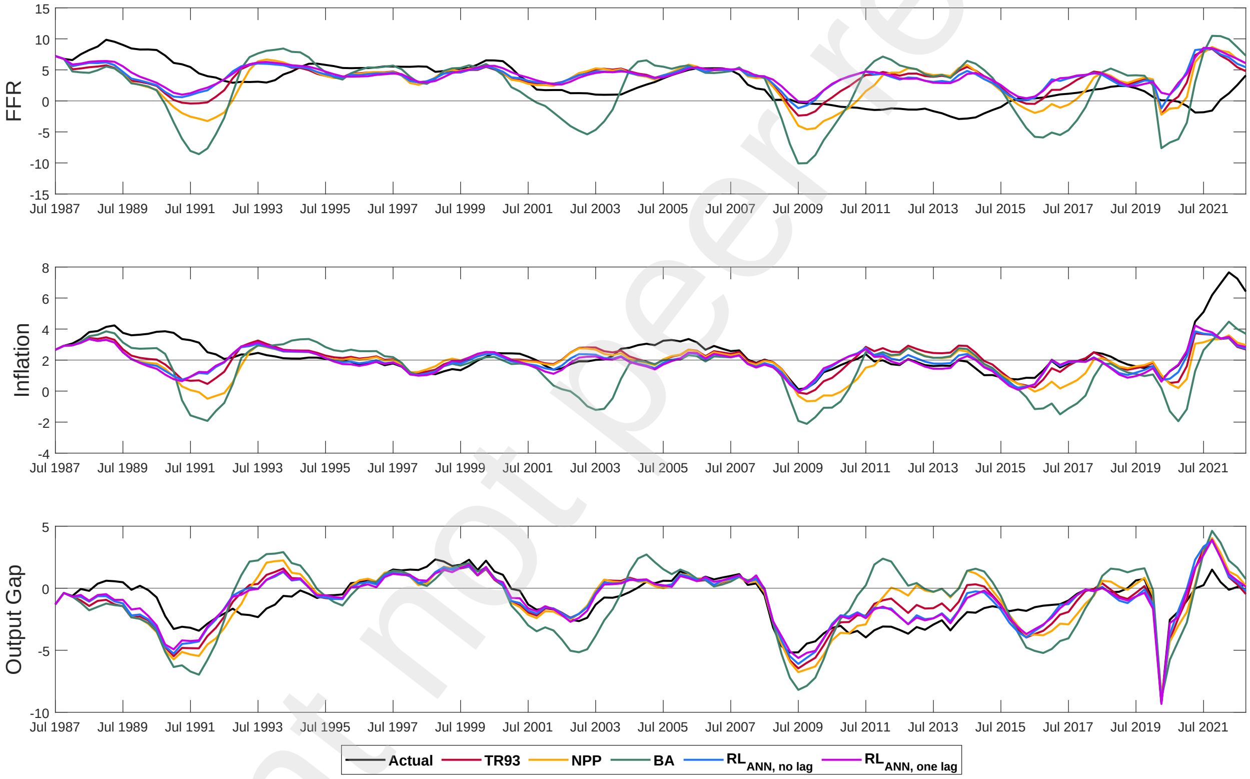
Starting with a situation where the interest rate is around 7%, inflation above target (3%), and the output gap below target (-1), the interest rate paths prescribed by the policy rules decline until the 1990s, which differs from the FFR set at that time. This corresponds with inflation and the output gap declining earlier than actually observed, bringing inflation closer to the target but creating undesirably low output gap values. The rule performing the worst around 1991, after the stock market crash, is the BA. Having a rather large weight on the output gap, the FFR drops drastically. In contrast, the NPP with a large inflation focus performs second worst here, reflecting that the optimal reaction coefficients lie between these two rules (see Table 3). Subsequently, all rules prescribe similar rates, which are close to the actual ones, corresponding to the great moderation narrative.

The dot-com bubble crisis in 2001 and the resulting recession again yield a drop in the output gap and inflation, leading to a pronounced FFR drop under the BA. In contrast to the previous crisis, however, the other rules maintain higher interest rate levels than the Fed actually did between 2001 and 2005. As a consequence, the economy is closer to target in that period following the RL rules, TR93, or the NPP, indicating a *too low for too long* issue as raised by, e.g., Taylor (2007).[20]

From 2007 onwards, the policy rule prescriptions start to differ much more from the actual FFR path. It is well-known that following the financial crisis in 2007, the Fed lowered the interest rate very quickly towards zero and started unconventional policy measures as the crisis endured, leading to a negative shadow rate between 2008 and

[20] Note, however, that the simulated counterfactual values of inflation and the output gap differ from the observed ones during that time due to the dynamic setup of this exercise.

Figure 6: Actual and Counterfactual Series (ANN Economy, Linear Policies)



Note: Starting with 1987:Q3, this figure shows FFR, inflation and output gap series from a dynamic counterfactual analysis of common rules (*TR93*: red, *NPP*: yellow, *BA*: green) and optimized linear rules ($RL_{ANN, no\,lag}$: blue, $RL_{ANN, one\,lag}$: purple) within the ANN economy. *Actual* refers to the historic time series (black).

2015. While the other policy rules mimic the sharp drop in the FFR initially, the interest rate values return to normal levels quite quickly (around 2011) given that inflation was close to target again. It seems as if the late and more moderate FFR reduction has a less negative signaling effect on the economy, keeping the inflation drop smaller and allowing it to stabilize around the target by 2011 while the negative output gap is constantly reduced.

In the following years prior to the COVID-19 crisis, the policy rules show more variation in the FFR. While the Fed was trying to maneuver out of the ZLB slowly, the other rules would have reduced the interest rates back from the previously higher levels around 2015. The start of the COVID-19 crisis in 2020, indicated by a sharp drop in the output gap and increasing inflation, causes all rules to increase their interest rates earlier than the Fed did. This delay could have caused inflation to rise much more than necessary (up to around 8%), whereas it stays below 5% following the other policy rules.

Table 4: Actual and Counterfactual Target Deviation and Loss (ANN Economy)

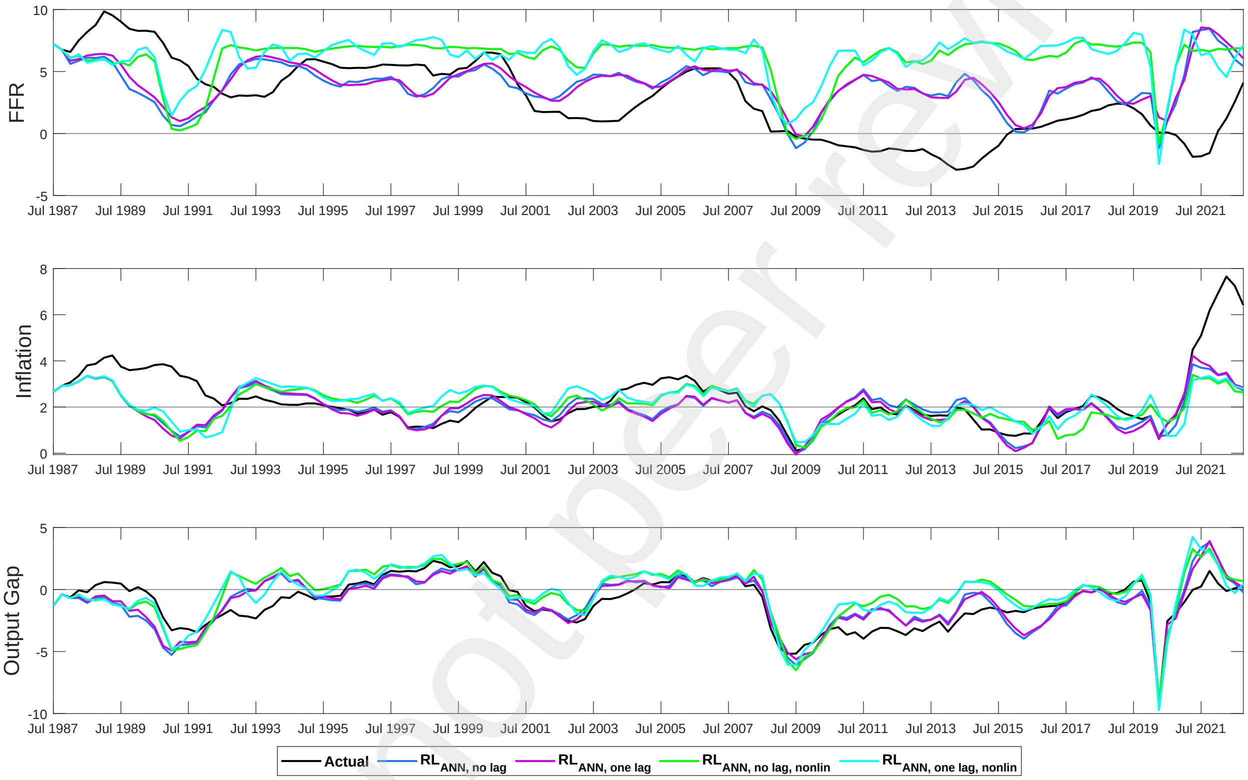
Policy	$\Delta^2(\pi^*, \pi_t)$	$\Delta^2(y^*, y_t)$	Loss
Actual	1.70	4.37	3.03
TR93	0.67	5.22	2.95
NPP	0.98	6.01	3.50
BA	2.72	9.08	5.90
$RL_{ANN,\,no\,lag}$	0.55	4.87	2.71
$RL_{ANN,\,one\,lag}$	0.62	4.56	2.59
$RL_{ANN,\,no\,lag,\,nonlin}$	**0.47**	4.10	2.29
$RL_{ANN,\,one\,lag,\,nonlin}$	0.52	**3.90**	**2.21**

Note: Δ^2 denotes the mean squared deviation of the respective variable from its target value ($\pi^* = 2$ and $y^* = 0$). The loss is calculated averaging over both: $Loss = 0.5 \cdot \Delta^2(\pi^*, \pi_t) + 0.5 \cdot \Delta^2(y^*, y_t)$.

Adding numbers to the descriptive analysis, Table 4 shows that $RL_{ANN,,no,lag}$ ($RL_{ANN,,one,lag}$) reduces the squared deviations of inflation from its target by a remarkable 68% (64%) compared to actual values. This is achieved by sacrificing some output gap stabilization, as sticking to the linear optimized rule $RL_{ANN,,one,lag}$, its mean squared deviation increases by 11% relative to actual data. $RL_{ANN,,no,lag}$ reduces the impairment with respect to the output gap to 4%. Looking at the other rules, the statistics prove the oversteering by the BA resulting in unacceptably high deviations, while TR93 and NPP improve upon the Fed in stabilizing inflation, whereas output gap values are too far off. These results show that the optimized RL rules are capable of reducing the total loss of the Fed policy by 11% and 15%, respectively, and thus yield the best results among the linear rules. The trade-off between inflation and output gap stabilization is better tackled by a larger (smaller) weight on inflation (output gap) (see also 3).

ANN Economy & Nonlinear Policy. In Figure 7, we add the *nonlinear* RL optimized policies within the ANN economy to the *linear* ones, conducting the same dynamic counterfactual.

Figure 7: Actual and Counterfactual Series (ANN Economy, RL Policies)



Note: Starting with 1987:Q3, this figure shows FFR, inflation and output gap series from a dynamic counterfactual analysis of linear ($RL_{ANN,\,no\,lag}$: blue, $RL_{ANN,\,one\,lag}$: purple) and nonlinear ($RL_{ANN,\,no\,lag,\,nonlin}$: green, $RL_{ANN,\,one\,lag,\,nonlin}$: petrol-blue) RL optimized rules within the ANN economy. *Actual* refers to the historic time series (black).

A first glance at the results reveals that overall, the nonlinear RL rules $RL_{ANN,,no,lag,,nonlin}$ and $RL_{ANN,,one,lag,,nonlin}$ keep the interest rate at higher levels (around 5-6%) than the linear RL rules. This is caused by, on average, slightly higher output gap and inflation values. Taking a closer look at the Gulf War period in 1989, the nonlinear RL rules increase the FFR moderately before starting a late but sharp drop. Referring back to the visualization of the policies in 4, the sharp decline in the FFR results from inflation close to zero combined with a large negative output gap. During the great moderation, the nonlinear rules prescribe larger interest rate values than the linear ones. There is a negligible reaction to the dot-com bubble crisis, but a large drop after the financial crisis and the COVID-19 crisis, even briefly turning into negative territory in the latter case. However, the interest rate under the nonlinear rules always quickly returns to its previous value, which especially stabilizes inflation at the end of the sample.

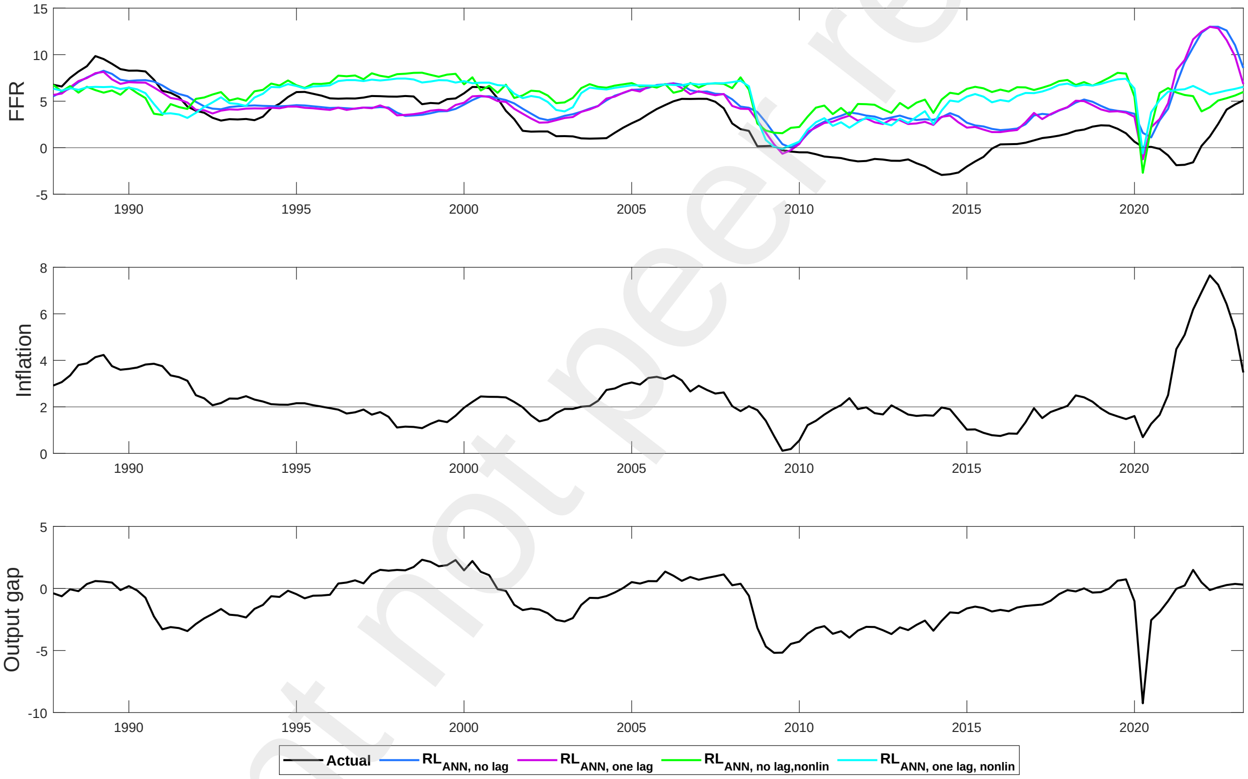
The overall result in Table 4 provides evidence that during the observed time span, the $RL_{ANN,,one,lag,,nonlin}$ ($RL_{ANN,,no,lag,,nonlin}$) performs best (and second best) as it is able to reduce loss compared to actual by 27% (24%). Both nonlinear rules are also able to improve upon output gap stabilization, reducing the squared deviations by 11% (6%) compared to actual, whereas inflation deviations improve even more compared to the linear rules, reducing the actual loss by 69% (72%). This result demonstrates that allowing the policy function to take on a nonlinear form has a positive impact on overall performance. By allowing the FFR to react differently to inflation and the output gap depending on the state of the economy, the trade-off between inflation and output gap stabilization can be mitigated to some extent.

3.4. *Static Counterfactuals*

Thus far, we have conducted dynamic counterfactual analyses, wherein inflation and output gap variables respond according to our estimated transition equations, resulting in observations that differ from actual values. To analyze the interest rate implications of optimized rules given realized values of inflation and output gap, we now turn to a static counterfactual setup. By inputting quarterly data on inflation and the output gap into each rule without considering any feedback mechanisms, we can directly compare the different interest rate prescriptions to the actual behavior of the Federal Reserve. This approach is frequently employed in the literature to compute measures of discretion (see, e.g., Nikolsko-Rzhevskyy et al. (2014, 2021) and Cochrane et al. (2019)). However, it does not provide information on which policy is preferable.

Figure 8 plots the counterfactuals of all RL rules based on the ANN economy, as well as the actual interest rate path (the counterfactuals based on the SVAR economy are in Appendix D). While the actual path falls between the linear and nonlinear optimized rules from the mid to late 1990s, all RL rules prescribe higher interest rates compared to the actual path prior to the Great Financial Crisis, supporting the *too low for too long* argument raised by, e.g., Taylor (2007). Moreover, none of the rules yield a persistently negative interest rate afterward. They only briefly touch the zero lower bound (ZLB) in 2009 and prescribe much higher rates until 2019. The nonlinear rules exceed the linear ones, reaching interest rate values close to 7% prior to the COVID-19 crisis, explained by a closed output gap and inflation close to target. The crisis leads to a significant drop in interest rates across all rules, down to below -2% due to the immense drop in output. The response of the RL is similar in value but leads the actual one by some quarters. As previously stated, we cannot draw conclusions on optimality from this figure. The high level of uncertainty surrounding COVID-19 perhaps speaks in favor of a smoother interest rate path than the RL rules recommend. At the end of the sample, interest rates skyrocket under the linear RL rules due to historically high inflation rates. Nonlinear RL rules suggest a more modest increase, reaching interest levels around 6

Figure 8: Actual and Static Counterfactual FFR Prescriptions under RL Rules



Note: This figure shows the static counterfactual of the optimized policy rules within the ANN economy. *Actual* refers to the FFR time series (black). Our ANN-based optimized linear policy rules are depicted in blue ($RL_{ANN,\,no\,lag}$) and purple ($RL_{ANN,\,one\,lag}$). Our ANN-based optimized nonlinear policy rules are depicted in green ($RL_{ANN,\,no\,lag,\,nonlin}$) and petrol-blue ($RL_{ANN,\,one\,lag,\,nonlin}$).

4. Discussion

Taking a step back, we aim to analyze the differences between the results of the conducted exercises, also laying out key assumptions on which our findings depend.

Starting with the representation of the economy, we have seen that ANNs can serve as a beneficial modeling tool. They allow us to capture nonlinearities while being agnostic about the specific functional form, bringing the model closer to the data compared to a standard SVAR, which is restricted to its linear structure. The nonlinear relationships between inflation and output gap can influence the effects of monetary policy and should therefore be taken into account when computing optimal interest rate reaction functions.

We wish to emphasize that the RL optimized policies depend on the assumed loss function and the underlying transition equations, as these are fixed during learning. The interaction between policy design and expectation formation is certainly important in practice and may alter our results. In the historical counterfactual exercise, we also take the estimated economy representations as given. Different models lead to different counterfactual paths, and quantitative results depend on the model chosen as well as the respective data used to estimate it. We do not claim that the ANN economy is the true model, nor do we intend to criticize the Fed's monetary policy. Rather, our aim is to provide a proof of concept of the RL algorithm within a data-driven economy representation and to contrast the policy rules in varying settings.

Relying on our estimated nonlinear ANN economy, reinforcement learning finds policies that perform well in such a world. The loss under the best-performing common rule, TR93, is 3% smaller than the actual one. Nevertheless, it performs worse compared to both of our optimized linear rules $RL_{ANN,,no,lag}$ and $RL_{ANN,,one,lag}$, where the latter reduces the actual loss by 15%. This is achieved by a smaller reaction coefficient on the output gap and a larger one on inflation compared to the TR93. The rules also perform reasonably well in a DSGE model comparison exercise, as illustrated in Appendix E.

Concerning linear versus nonlinear policy rules, the losses resulting from the historical counterfactual suggest that nonlinear optimized policies are even better economic stabilizers (Table 4). The best-performing rule, $RL_{ANN,,one,lag,,nonlin}$, reduces the loss by over 27%.

The success of the nonlinear RL rules is justified by their flexibility regarding the degree of responsiveness to their input variables depending on their levels. As the partial dependence plots (PDPs) suggest, there is a plateau around the target inflation and output gap, where not much action by the central bank is required. The reaction is stronger to the output gap when the latter is negative. The same holds true for inflation when it is below target.

In the static counterfactual, our RL optimized rules imply that the Fed set interest rates too low prior to the financial crisis. Moreover, our results speak against quantitative easing, as the prescribed rules stay in positive territory. Regarding the response to the COVID-19 crisis, the rules suggest that an earlier and potentially larger increase in the nominal interest rate would have been optimal compared to actual monetary policy.

This paper introduces a novel machine learning-based approach to deriving optimal monetary policy reaction functions in accordance with a central bank's preferences. Initially, we demonstrate how artificial neural networks (ANNs) can be employed as a modeling tool for transition equations, effectively capturing nonlinear interdependencies and thereby aligning the model more closely with empirical data compared to a standard structural vector autoregression (SVAR). Subsequently, we apply reinforcement learning (RL), a computational method for goal-directed learning from interaction and optimal control.

Reinforcement learning is sufficiently flexible to accommodate various types of nonlinearities. Additionally, it does not necessitate perfect knowledge of the environment, thereby mitigating the issue of model uncertainty. Consequently, RL retains many advantages associated with optimal control approaches while offering greater flexibility in the design of the optimality problem. This flexibility is particularly beneficial for future expansions involving an increased number of state variables, enabling the creation of more realistic and complex environments.

Utilizing a nonlinear ANN representation of the economy, our optimized linear rules are shown to reduce the central bank's loss by up to 15%, while the best-performing nonlinear rule achieves a reduction of over 27% compared to values implied by actual data. Linear RL rules enhance inflation stabilization, albeit with a slight trade-off in output gap stabilization. In contrast, the nonlinear RL rules are capable of reducing deviations in both inflation and output gap compared to all other rules and the actual central bank behavior. The static counterfactual analysis also supports the view that the Federal Reserve maintained interest rates at excessively low levels over many periods. Furthermore, the optimized rules suggest that an earlier and more pronounced increase in the nominal interest rate would have been optimal since the onset of COVID-19.

Our paper should be regarded as a proof of concept for the application of RL to monetary policy optimization problems, leaving ample scope for future research. The Federal Reserve announced a shift to average inflation targeting and plans to review its strategy approximately every five years. Future research could therefore consider loss functions that incorporate an average inflation target. Additionally, one could explore combining the advantages of economic modeling and RL by deviating from rational expectations and integrating a learning central bank with learning private agents within a dynamic stochastic general equilibrium (DSGE) model. While this paper focuses on reaction functions for the nominal interest rate, future research might also consider reaction functions for unconventional monetary policy measures such as asset purchases. The most evident step for future research is to incorporate more variables to broaden the representation of the economy as well as the controlled variables of the central bank. Unlike standard dynamic programming algorithms that suffer from the curse of dimensionality, increasing the amount of data is unproblematic with deep RL.

Based on the promising results of this paper, we recommend incorporating reinforcement learning into the toolkit of central bankers for determining optimal monetary policy reaction functions.

References

ADAM, K. AND R. M. BILLI (2006): "Optimal Monetary Policy under Commitment with a Zero Bound on Nominal Interest Rates," *Journal of Money, Credit and Banking*, 83, 1877–1905.

BALL, L. AND S. MAZUMDER (2019): "A Phillips Curve with Anchored Expectations and Short-Term Unemployment," *Journal of Money, Credit and Banking*, 51, 111–137.

BENCHIMOL, J. (2024): "Central Bank Losses, Monetary Policy Rules, and Limited Information," .

BENCHIMOL, J. AND L. BOUNADER (2023): "Optimal monetary policy under bounded rationality," *Journal of Financial Stability*, 67, 101151.

BERNANKE, B. S., M. GERTLER, AND S. GILCHRIST (1999): "The Financial Accelerator in a Quantitative Business Cycle Framework," *Handbook of Macroeconomics*, 1, 1341–1393.

BOTVINICK, M., S. RITTER, J. X. WANG, Z. KURTH-NELSON, C. BLUNDELL, AND D. HASSABIS (2019): "Reinforcement Learning, Fast and Slow," *Trends in Cognitive Sciences*, 23, 408–422.

CARABENCIOV, I., M. C. FREEDMAN, M. R. GARCIA-SALTOS, M. D. LAXTON, M. O. KAMENIK, AND M. P. MANCHEV (2013): "GPM6: The Global Projection Model with 6 Regions," IMF Working Paper 13-87, International Monetary Fund.

CASTRO, P. S., A. DESAI, H. DU, R. GARRATT, AND F. RIVADENEYRA (2021): "Estimating Policy Functions in Payment Systems using Reinforcement Learning," Staff Working Paper 2021-7, Bank of Canada.

CHARPENTIER, A., R. ELIE, AND C. REMLINGER (2021): "Reinforcement learning in economics and finance," *Computational Economics*, 1–38.

CHEN, M., A. JOSEPH, M. KUMHOF, X. PAN, R. SHI, AND X. ZHOU (2021): "Deep Reinforcement Learning in a Monetary Model," *arXiv preprint arXiv:2104.09368*.

CHRISTIANO, L. J., R. MOTTO, AND M. ROSTAGNO (2014): "Risk Shocks," *American Economic Review*, 104, 27–65.

COCHRANE, J. H., J. B. TAYLOR, AND V. WIELAND (2019): "Evaluating Rules in the Fed's Report and Measuring Discretion," in *Hoover Institution, Strategies for Monetary Policy: A Policy Conference*.

COGAN, J. F., T. CWIK, J. B. TAYLOR, AND V. WIELAND (2010): "New Keynesian versus Old Keynesian Government Spending Multipliers," *Journal of Economic Dynamics and Control*, 34, 281–295.

Coibion, O. and Y. Gorodnichenko (2015): "Is the Phillips Curve Alive and Well After All? Inflation Expectations and the Missing Disinflation," *American Economic Journal: Macroeconomics*, 7, 197–232.

Cúrdia, V. and M. Woodford (2009): "Credit Frictions and Optimal Monetary Policy," BIS Working Paper 278, Bank for International Settlements.

Debortoli, D., J. Kim, J. Lindé, and R. Nunes (2019): "Designing a Simple Loss Function for Central Banks: Does a Dual Mandate Make Sense?" *The Economic Journal*, 129, 2010–2038.

Del Negro, M., M. P. Giannoni, and F. Schorfheide (2015): "Inflation in the Great Recession and New Keynesian Models," *American Economic Journal: Macroeconomics*, 7, 168–96.

Dickey, D. A. and W. A. Fuller (1979): "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 74, 427–431.

Dolado, J. J., R. María-Dolores, and M. Naveira (2005): "Are Monetary-Policy Reaction Functions Asymmetric?: The Role of Nonlinearity in the Phillips Curve," *European Economic Review*, 49, 485–503.

Dolado, J. J., R. M.-D. Pedrero, and F. J. Ruge-Murcia (2004): "Nonlinear Monetary Policy Rules: Some New Evidence for the US," *Studies in Nonlinear Dynamics & Econometrics*, 8.

Fernández-Villaverde, J., P. Guerrón-Quintana, K. Kuester, and J. Rubio-Ramírez (2015): "Fiscal Volatility Shocks and Economic Activity," *American Economic Review*, 105, 3352–84.

Gaspar, V., F. Smets, and D. Vestin (2006): "Adaptive learning, persistence, and optimal monetary policy," *Journal of the European Economic Association*, 4, 376–385.

Gertler, M. and P. Karadi (2011): "A model of Unconventional Monetary Policy," *Journal of Monetary Economics*, 58, 17–34.

Glorot, X. and Y. Bengio (2010): "Understanding the Difficulty of Training Deep Feedforward Neural Networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 249–256.

Goodfriend, M. (2004): "Inflation Targeting in the United States?" in *The Inflation-Targeting Debate*, University of Chicago Press, 311–352.

Hansen, L. and T. J. Sargent (2001): "Robust Control and Model Uncertainty," *American Economic Review*, 91, 60–66.

HAWKINS, R. J., J. K. SPEAKES, AND D. E. HAMILTON (2015): "Monetary Policy and PID Control," *Journal of Economic Interaction and Coordination*, 10, 183–197.

IACOVIELLO, M. AND S. NERI (2010): "Housing Market Spillovers: Evidence from an Estimated DSGE Model," *American Economic Journal: Macroeconomics*, 2, 125–64.

KINGMA, D. P. AND J. BA (2014): "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*.

KWIATKOWSKI, D., P. C. B. PHILLIPS, P. SCHMIDT, AND Y. SHIN (1992): "Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159–178.

LEVIN, A., V. WIELAND, AND J. C. WILLIAMS (2003): "The Performance of Forecast-based Monetary Policy Rules under Model Uncertainty," *American Economic Review*, 93, 622–645.

LILLICRAP, T. P., J. J. HUNT, A. PRITZEL, N. M. O. HEESS, T. EREZ, Y. TASSA, D. SILVER, AND D. WIERSTRA (2015): "Continuous Control with Deep Reinforcement Learning," *Computing Research Repository (CoRR)*.

LUCAS, R. E. (1976): "Econometric Policy Evaluation: A Critique," in *Carnegie-Rochester Conference Series on Public Policy*, 19–46.

MNIH, V., K. KAVUKCUOGLU, D. SILVER, A. GRAVES, I. ANTONOGLOU, D. WIERSTRA, AND M. RIEDMILLER (2013): "Playing Atari with Deep Reinforcement Learning," *arXiv preprint arXiv:1312.5602*.

MNIH, V., K. KAVUKCUOGLU, D. SILVER, A. A. RUSU, J. VENESS, M. G. BELLEMARE, A. GRAVES, M. RIEDMILLER, A. K. FIDJELAND, G. OSTROVSKI, ET AL. (2015): "Human-level Control through Deep Reinforcement Learning," *Nature*, 518, 529–533.

NIKOLSKO-RZHEVSKYY, A., , D. H. PAPELL, AND R. PRODAN (2021): "Policy Rules and Economic Performance," *Journal of Macroeconomics*, 68, 103291.

NIKOLSKO-RZHEVSKYY, A., D. H. PAPELL, AND R. PRODAN (2014): "Deviations from Rules-Based Policy and Their Effects," *Journal of Economic Dynamics and Control*, 49, 4–17.

——— (2018): "Policy Rules and Economic Performance," Working paper.

ORPHANIDES, A. (2001): "Monetary Policy Rules Based on Real-Time Data," *American Economic Review*, 91, 964–985.

——— (2003): "Monetary Policy Evaluation with Noisy Information," *Journal of Monetary economics*, 50, 605–631.

ORPHANIDES, A. AND V. WIELAND (2000): "Inflation Zone Targeting," *European Economic Review*, 44, 1351–1387.

PRIMICERI, G. E. (2005): "Time Varying Structural Vector Autoregressions and Monetary Policy," *The Review of Economic Studies*, 72, 821–852.

REIS, R. (2009): "Optimal monetary policy rules in an estimated sticky-information model," *American Economic Journal: Macroeconomics*, 1, 1–28.

ROTEMBERG, J. J. AND M. WOODFORD (1997): "An Optimization-based Econometric Framework for the Evaluation of Monetary Policy," *NBER Macroeconomics Annual*, 12, 297–346.

RUDEBUSCH, G. AND L. E. SVENSSON (1999): "Policy Rules for Inflation Targeting," in *Monetary Policy Rules*, University of Chicago Press, 203–262.

SARGENT, T., N. WILLIAMS, AND T. ZHA (2006): "Shocks and government beliefs: The rise and fall of American inflation," *American Economic Review*, 96, 1193–1224.

SCHALING, E. (2004): "The Nonlinear Phillips Curve and Inflation Forecast Targeting: Symmetric versus Asymmetric Monetary Policy Rules," *Journal of Money, Credit and Banking*, 36, 361–386.

SHI, R. A. (2021): "Learning from Zero: How to Make Consumption-Saving Decisions in a Stochastic Environment with an AI Algorithm," CESifo Working Paper 9255, CESifo.

SILVER, D., G. LEVER, N. HEESS, T. DEGRIS, D. WIERSTRA, AND M. RIEDMILLER (2014): "Deterministic Policy Gradient Algorithms," in *International Conference on Machine Learning*.

SIMS, C. A. AND T. ZHA (2006): "Were there Regime Switches in US Monetary Policy?" *American Economic Review*, 96, 54–81.

SMETS, F. AND R. WOUTERS (2007): "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach," *American Economic Review*, 97, 586–606.

SVENSSON, L. (1997): "Inflation Forecast Targeting: Implementing and Monitoring Inflation Targets," *European Economic Review*, 41, 1111–1146.

SVENSSON, L. E. (2020): "Monetary Policy Strategies for the Federal Reserve," *International Journal of Central Banking*, 16, 133–193.

TAMBAKIS, D. N. (2009): "Optimal Monetary Policy with a Convex Phillips Curve," *The BE Journal of Macroeconomics*, 9.

TAYLOR, J. B. (1993): "Discretion versus Policy Rules in Practice," in *Carnegie-Rochester Conference Series on Public Policy*, vol. 39, 195–214.

——— (2007): "The Explanatory Power of Monetary Policy Rules," *Business Economics*, 42, 8–15.

TAYLOR, J. B. AND J. C. WILLIAMS (2010): "Simple and robust rules for monetary policy," in *Handbook of Monetary Economics*, ed. by B. M. Friedman and M. Woodford, Elsevier, vol. 3, 829–859.

TETLOW, R. J. AND P. VON ZUR MUEHLEN (2001): "Robust Monetary Policy with Misspecified Models: Does Model Uncertainty Always Call for Attenuated Policy?" *Journal of Economic Dynamics and Control*, 25, 911–949.

WATSON, M. W. (2014): "Inflation Persistence, the NAIRU, and the Great Recession," *American Economic Review*, 104, 31–36.

WIELAND, V. (2000): "Monetary Policy, Parameter Uncertainty and Optimal Learning," *Journal of Monetary Economics*, 46, 199–228.

WIELAND, V., E. AFANASYEVA, M. KUETE, AND J. YOO (2016): "New Methods for Macro-financial Model Comparison and Policy Analysis," in *Handbook of Macroeconomics*, Elsevier, vol. 2, 1241–1319.

WIELAND, V., T. CWIK, G. J. MÜLLER, S. SCHMIDT, AND M. WOLTERS (2012): "A New Comparative Approach to Macroeconomic Modeling and Policy Analysis," *Journal of Economic Behavior & Organization*, 83, 523–541.

WOODFORD, M. (2001): "The Taylor Rule and Optimal Monetary Policy," *American Economic Review*, 91, 232–237.

WU, J. C. AND F. D. XIA (2016): "Measuring the macroeconomic impact of monetary policy at the zero lower bound," *Journal of Money, Credit and Banking*, 48, 253–291.

ZHENG, S., A. TROTT, S. SRINIVASA, N. NAIK, M. GRUESBECK, D. C. PARKES, AND R. SOCHER (2020): "The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies," *arXiv preprint arXiv: 2004.13332*.

Appendix (intended to be published as supplementary material)

Appendix A: DDPG Algorithm and Hyperparameters

Corresponding to Table 1, we would like to explain the DDPG algorithm in more detail:

a) First, the actor parameters $\theta^P = [\beta_j, \delta_j, \alpha_i]$ with $j = 1, \dots, q$ and $i = 0, \dots, q$ are initialized. Linear rules are initialized using the Taylor rule parameters. Nonlinear rules' parameters are initially set to one. The critic parameters θ^Q are randomly initialized using the *glorot* function of Glorot and Bengio (2010).[21] Biases are again set to zero initially. Hereby, the actor and critic networks $P(x|\theta^P)$ and $Q(x, i|\theta^Q)$ are initialized.

b) As the stability of the Q-learning process is increased by using 'soft' target updates instead of directly changing the calculated weights, copies of the actor $(P'(x|\theta^{P'}))$ and the critic $Q'(x, i|\theta^{Q'}))$ are generated in order to calculate the target values. Their parameters are initialized using initial θ^P and θ^Q.

c) The replay buffer B is also initialized in order to store the experiences of the agent in a later step.

d) A major challenge of learning in continuous action spaces is exploration (Lillicrap et al., 2015). The action taken at each time step t is therefore subject to some noise, which encourages exploration of the actor and can be suited to the environment. The underlying noise model $\mathcal{N}$ is an Ornstein-Uhlenbeck process with mean zero. To encourage exploration, it is common to set the variance between 1 % and 10 % of the action range, which is 18 in our case. Hence, we choose a variance of 1, while the variance decay rate stays at default.

e) To keep the analysis close to reality, we initialize the observational states $x_0^z, z \in 1, 2$ by randomly drawing pairs π_0 and y_0 from our data series. This approach can be interpreted as challenging the algorithm with different economic situations from our data set as a starting point for training. As further lags are required to compute the next state using our economy representations, we also initialize these from the data.

f) The action is computed based on the current policy function parameters, inputs plus a random noise:

$$i_t = P(x_t|\theta_t^P) + \mathcal{N}_t. \tag{A.1}$$

g) The previously chosen action and the state enter the environmental equations, i.e. our linear (6) and (7) or ANN economy (8). The next observations $x_{t+1} = (\pi_{t+1}, y_{t+1})$ can be calculated. Note that this simulation includes random shocks, with mean zero and variance equal to that of the estimated shocks.

[21] The *glorot* initializer independently samples from a uniform distribution with zero mean and variance $2/(InputSize + OutputSize)$. In our case, the denominator depends on the number of critic nodes.

h) The data tuple (x_t, i_t, r_t, x_{t+1}) is then stored in the replay buffer B.

i) As information mass can become a problem in such a continuous setting, the algorithm learns on mini-batches drawn from the replay buffer. This buffer contains only a certain amount of samples and drops the oldest when it is full. A minibatch of size N is sampled uniformly from the buffer B and is used to update actor and critic at every time step. We use the default values of 10000 and $N = 64$ for the experience buffer and mini-batch sizes, respectively.

j) For each sample in the minibatch, h_j is calculated according to

$$h_j = r(x_j, i_j) + \gamma Q'(x_{j+1}, P'(x_{j+1}|\theta^{P'})|\theta^{Q'}). \tag{A.2}$$

It is composed of the reward in j plus the discounted future reward, presuming adherence to the present target actor and critic networks. The discount factor is set to $\gamma = 0.99$ (default).

k) When calculating the squared deviations of $(h_j - Q(x_j, i_j|\theta^Q))$, one evaluates the performance of critic parameters θ^Q versus the target critic parameters $\theta^{Q'}$.

$$L = \frac{1}{N} \sum_j (h_j - Q(x_j, i_j|\theta^Q))^2 \tag{A.3}$$

By minimizing this loss function L, also called Bellman residuum, the critic parameters are updated. The speed of parameter adjustment is given by the learn rate, which is set to 0.000025.

l) The policy gradient, i.e. the gradient of the policy's performance with respect to the coefficients θ^P, is calculated using the chain rule:

$$\nabla_{\theta^P} J \approx \frac{1}{N} \sum_j [\nabla_i Q(x, i|\theta^Q)|_{x=x_j, i=P(x_j)} \nabla_{\theta^P} P(i|\theta^P)|_{x=x_j}], \tag{A.4}$$

where J denotes the expected cumulative discounted reward from the initial state. Differentiating the critic with respect to the nominal interest rate i and multiplying the derivative of the policy function with respect to the policy parameters yields the policy gradient. This gradient determines the update of the coefficients. By taking small steps at each iteration in the direction of the negative gradient of the loss, the loss function is minimized and the parameters are optimized. The applied optimizer is the *adaptive moment estimation* or *adam* (see Kingma and Ba (2014)). The learn rate is identical to the critic learn rate.

m) Both target network (actor and critic) weights are adjusted through a slow tracking of the actual networks' parameters: $\theta' \leftarrow \tau\theta + (1-\tau)\theta', \tau < 1$. This means that the target values are constrained to change slowly, greatly improving the stability of learning (Lillicrap et al., 2015). In our case, $\tau_\pi = \tau_y = 0.001$ are set to default values.

Agent Selection Criteria. During the RL routine, we save the best agents. Specifically, we save all agents during training that fulfill the following criteria. First of all, the episode reward (ER_m) divided by the episode steps (ES_m) has to be larger than -4, i.e. $ER_m/ES_m > -4$. According to our reward definition in (14), this is equivalent to an average (per step) inflation and output gap deviation from target of two percentage points and corresponds to the *inner* part of the reward function that is not further punished. We choose this criteria to be rather loose as we do not want to discriminate agents that start from worse initial states, i.e. from states further away from the targets. Moreover, we require $1 < ES_m < 13$ to avoid, on the one hand, choosing an agent that reached the target coincidentally after one step because of a close to target initial state. On the other hand, the episode shall be terminated, i.e. the episode stopping criteria are reached, before the maximum number of steps per episode is reached.

Table A.1: Chosen Numbers of Hidden Nodes

Economy	Policy Structure	Policy Inputs	Critic Nodes	Actor Nodes
SVAR	Linear	(y_t, π_t)	1	1
SVAR	Linear	$(y_t, y_{t-1}, \pi_t, \pi_{t-1})$	1	1
ANN	Linear	(y_t, π_t)	1	1
ANN	Linear	$(y_t, y_{t-1}, \pi_t, \pi_{t-1})$	2	1
ANN	Nonlinear	(y_t, π_t)	1	8
ANN	Nonlinear	$(y_t, y_{t-1}, \pi_t, \pi_{t-1})$	1	10

Note: This table summarizes the chosen number of neurons for the neural networks representing the critic and the policy function. The decision rules for the optimal numbers are described in the main text.

Table A.2: Network Structure and Hyperparameters

	Network Structure		Hyperparameters
Critic	Observation Path	imageInputLayer	
		fullyConnectedLayer	Nodes: n
		tanhLayer	
		fullyConnectedLayer	Nodes: n
	Action Path	imageInputLayer	
		fullyConnectedLayer	Nodes: n
		tanhLayer	
		fullyConnectedLayer	Nodes: n
	Common Path	concatenationLayer	
		fullyConnectedLayer	Nodes: 1
	General		Initializer: Glorot Learn Rate: 0.000025 Gradient Threshold: 1 Optimizer: adam
Actor	Linear Version	imageInputLayer	
		fullyConnectedLayer	Nodes: 1
	Nonlin. Version	imageInputLayer	
		fullyConnectedLayer	Nodes: n
		tanhLayer	
		fullyConnectedLayer	Nodes: 1
	General		Initializer: He Learn Rate: 0.000025 Gradient Threshold: 1 Optimizer: adam

Note: This is an overview of the critic and actor network structures (c.f. 2.1.2) applied within the DDPG algorithm. The number of nodes (n) is varied over different training cycles and the resulting optimal number of nodes is shown in Table A.1. The gradient threshold is the threshold value for the gradient of step l) in the algorithm. If the gradient exceeds this value, it is clipped. This limits the parameter change in a training iteration. For more information regarding the adam optimizer, see Kingma and Ba (2014). Hyperparameters not mentioned here are kept at Matlab's default values and settings.

Appendix B: Estimation Results

Table B.1: Estimation Results of Restricted SVAR

Parameters	Estimates	p-Values
Output gap		
C^y	-0.2751	0.2333
$a^y_{y,1}$	0.7683	0.0000
$a^y_{\pi,1}$	-0.0569	0.4336
$a^y_{i,1}$	0.1902	0.0000
$a^y_{i,2}$	0.1809	0.0014
$\bar{R}^2$	0.7248	
MSE	0.9686	
DW	2.1473	
$LM(1)$	1.3539	0.1720
Inflation		
C^π	0.2168	0.0031
$a^\pi_{y,0}$	0.1202	0.0000
$a^\pi_{y,1}$	-0.1033	0.0000
$a^\pi_{\pi,1}$	1.5504	0.0000
$a^\pi_{\pi,2}$	-0.6344	0.0076
$a^\pi_{i,1}$	-0.0041	0.7427
$\bar{R}^2$	0.9416	
MSE	0.0940	
DW	2.2447	
$LM(1)$	1.3539	0.1720

Note: This table shows the estimation results of our linear economy represented by a restricted recursive SVAR(2).

Table B.2: Economy Fit: Mean Squared Errors

Representation	MSE Output Gap		MSE Inflation		MSE Total	
	Pre Covid	Post Covid	Pre Covid	Post Covid	Pre Covid	Post Covid
SVAR	0.309	5.861	0.202	0.443	0.256	3.152
ANN	0.277	4.460	0.176	0.296	0.226	2.378

Note: This table summarizes the mean squared errors of the linear SVAR ((6)-(7)) and ANN ((8)) economy representations for the variables output gap, inflation and the overall economy for the two different subperiods (Pre Covid: 1987Q3-2019:Q1, Post Covid: 2019:Q2-2023:Q2).

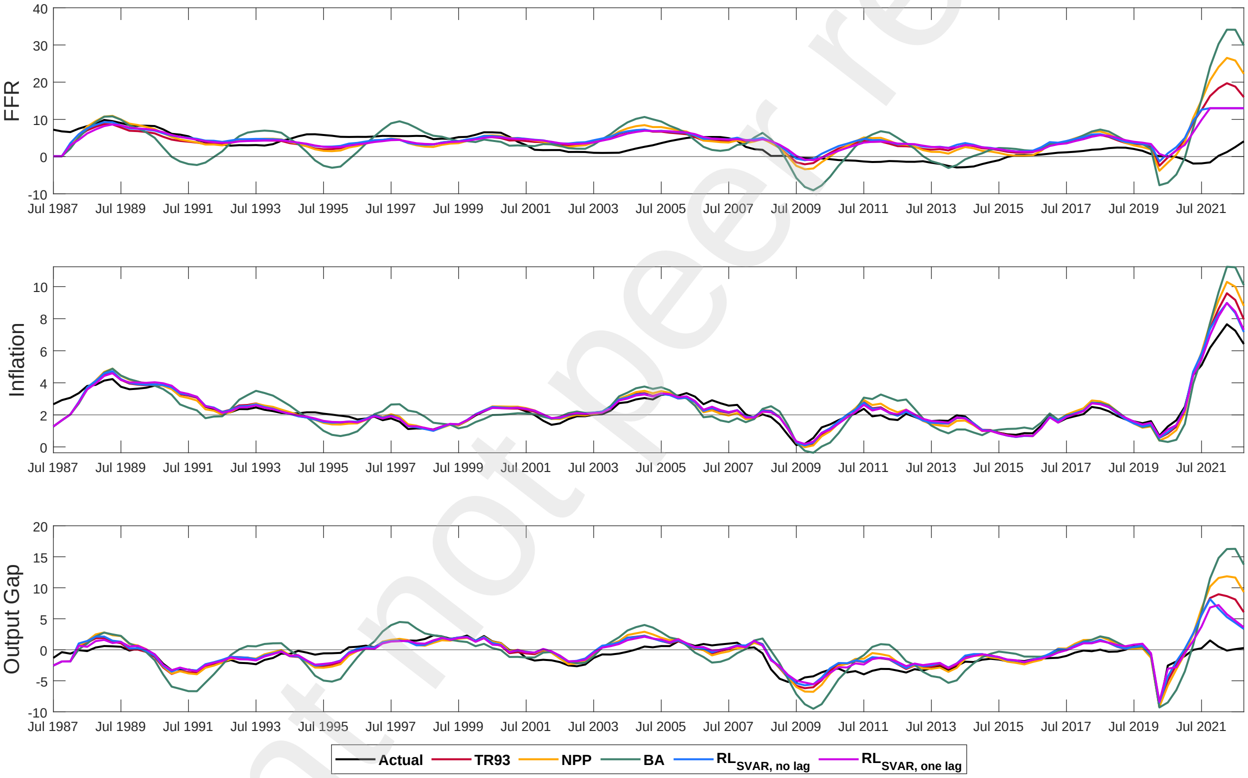
Appendix C: Historical Counterfactual - SVAR Economy & Linear Policy

This section presents the counterfactual simulation results using the estimated linear SVAR economy in (6)-(7). Figure C.1 shows the simulated time series for the interest rate, inflation and the output gap, respectively, under the different reaction functions.

First, it is remarkable that the interest rate level (top panel) with both optimized rules is, on average, larger than the actual federal funds rate and interest rate prescriptions of other rules. This, presumably, is caused by the combination of a relatively large constant term and larger coefficients on the input variables (see Table 3). Further, $RL_{SVAR,\,no\,lag}$ yields a smoother interest rate series than $RL_{SVAR,\,one\,lag}$, which might be explained by the latter having coefficients with different signs on the output gap (β_y^0 and β_y^1). At the beginning of the sample, TR93, NPP and BA produce interest rates slightly below, while our rules yield rates above the actual one. In relative terms, all rules share the drop in the interest rate after 1989, which mirrors the recession following the stock market crash. Between 1993 and 1999, the actual interest rate increases and draws closely to our optimized policies. Subsequently, the interest rate drops across all rules following the dot com bubble crisis. After that, the common rules change from running below to above actual. With respect to magnitude and time, however, they lag behind the optimized policies' interest rate increase. By explicitly including a ZLB during RL, it seems as if the policy rules increase the scope of monetary policy action by raising interest rates before the financial crisis. Through this behavior, our optimized policies support and affirm the *too low for too long* argument put forward, for example, by Taylor (2007).

What is actually of greater interest to us are the counterfactual inflation and output gap series, because these can be used to evaluate the performance of the RL reaction function. Looking at inflation, the paths under $RL_{SVAR,\,no\,lag}$ and $RL_{SVAR,\,one\,lag}$ are very similar and both produce values smaller than in the data. Between 1987 and 1991 and after 2003, the induced inflation is closer to the target of 2 %, while from 1995 to 2000, the Fed's actual behavior produced better inflation values. We find similar results for the output gap. The values induced by the optimized policies lie above the common rules and the data. Mostly, this yields values closer to the target of zero.

Figure C.1: Actual and Counterfactual Series (SVAR Economy)



Note: Starting with 1987:Q3, this figure shows FFR, inflation and output gap series from a dynamic counterfactual analysis of common rules (*TR93*: red, *NPP*: yellow, *BA*: green) and optimized linear rules ($RL_{SVAR,\,no\,lag}$: blue, $RL_{SVAR,\,one\,lag}$: purple) within the SVAR economy. *Actual* refers to the historic time series (black).

In order to draw a final conclusion about the performance of the RL policy functions, we look at the squared deviations of the counterfactual series from the respective targets and the resulting overall central bank loss, i.e. the reward defined in (14) multiplied by (-1). Table C.1 summarizes the results.

Table C.1: Actual and Counterfactual Target Deviation and Loss (SVAR Economy)

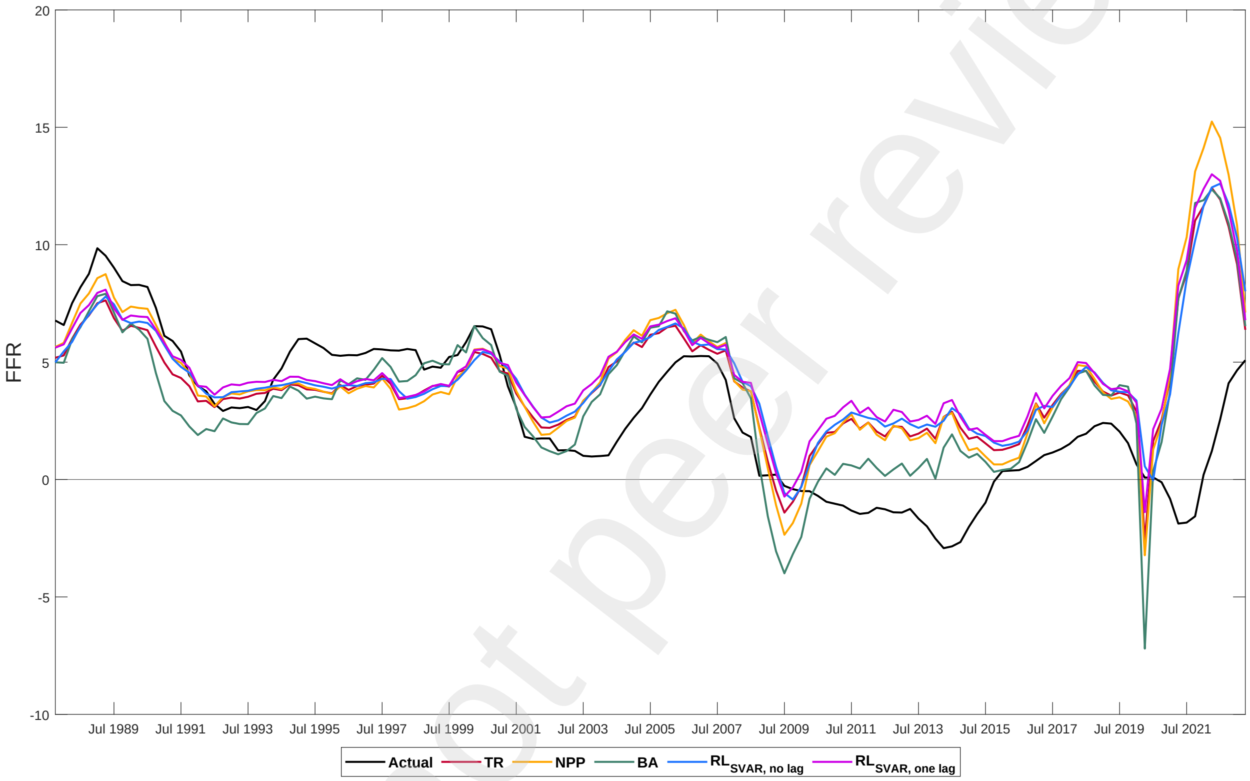
Policy	$\Delta^2(\pi^*, \pi_t)$	$\Delta^2(y^*, y_t)$	Loss
TR93	2.44	6.87	4.66
NPP	2.85	9.69	6.27
BA	3.44	17.46	10.45
$RL_{SVAR, no\, lag}$	2.22	5.28	3.75
$RL_{SVAR, one\, lag}$	2.14	5.02	3.58

Note: Δ^2 denotes the mean squared deviation of the respective variable from its target value ($\pi^* = 2$ and $y^* = 0$). The loss is calculated averaging over both: $Loss = 0.5 \cdot \Delta^2(\pi^*, \pi_t) + 0.5 \cdot \Delta^2(y^*, y_t)$.

We find that the optimized RL rules yield smaller squared deviations from target than these observed in the data. The other rules perform worse than the actual course, only the NPP rule yields a slightly lower squared deviation of inflation from target. In terms of the output gap, the $RL_{SVAR, no\, lag}$ rule slightly decreases the volatility around the target. The other rules perform worse. Calculating the loss, i.e. averaging over the squared deviations from target values, allows us to rank policy functions. Within the SVAR economy framework, the simple optimized rule $RL_{SVAR, no\, lag}$ performs best with a loss of 1.80, followed by $RL_{SVAR, one\, lag}$. Both rules yield an improvement compared to the actual monetary policy ($Loss = 1.91$). All other common policy rules are inferior.

Appendix D: Static Counterfactual

Figure D.1: FFR and Prescriptions from Common and RL Rules based on SVAR Economy



Note: This figure shows the static counterfactual of several common and the optimized policy rules within the linear economy. *Actual* refers to the FFR time series (black). In red we show the results of the Taylor (1993) rule (*TR93*), in yellow we see the inflation tilting rule (*NPP*) and the balanced approach (*BA*) is shown in green. Our optimized policy rules are depicted in blue ($RL_{SVAR, no\,lag}$) and purple ($RL_{SVAR, one\,lag}$).

Appendix E: Model Comparison

E.0.1. Model Comparison

Optimal monetary policy is always related to an environment, that,in the best case, closely reflects real-world relationships. So far, we have provided a linear SVAR environment and an ANN environment, with the latter producing the superior data fit. Hence, we postulate that the optimized policy rules based on the ANN economy are the better suited choice. Nevertheless, our ANN environment is just one economic *model*, and an optimal policy rule is required to be robust with respect to model uncertainty.

We therefore conduct a model comparison analysis, evaluating each policy rule's performance over 11 macroeconomic DSGE models.[22] The DSGE framework also allows for a true counterfactual analysis as the Lucas (1976) critique does not apply here.

E.0.1.1. The Models

We want our set of models to be manifold with respect to size and specific features like financial frictions. Therefore, we include models developed prior to and after the financial crisis. For this analysis we make use of the Macroeconomic Model Data Base (MMB) by Wieland et al. (2012, 2016).

Pre-crisis Models. First, we use a simple linear Keynesian model with backward-looking dynamics (Rudebusch and Svensson (1999)), which is compatible with our empirical setup. As the authors show, this estimated model explains U.S. data on inflation and GDP quite well. We refer to this model as *RS99*. Next, we consider a small forward-looking New Keynesian model (Levin et al. (2003)) and call it *LWW03*. Given their structure with three equations and the same variables as in our setup, they are well comparable to our economy specification. We further include the medium-scale DSGE model by Smets and Wouters (2007) (SW07) with a larger number of equations, variables and shocks. It can therefore better explain variation in key macroeconomic variables and data dynamics.

Post-crisis Models. Finally, we also add several medium and large-scale DSGE models that were developed after the financial crisis and which are more complex, including e.g. financial frictions. In total, we test out policy rules in 8 post-crisis models. The post-financial-crisis model by Cúrdia and Woodford (2009) (CW09) contains financial frictions and allows for a spread between savers and borrowers. The second large DSGE model by Iacoviello and Neri (2010) (IN10) focuses on the housing market and its spillovers to the rest of the economy. *IN10* contains financial frictions in the household sector and multiple shocks. The model by Cogan et al. (2010) (CCTW10) includes rule-of-thumb consumers and the fiscal sector allows for the analysis of fiscal multipliers. Gertler and Karadi (2011) (*GK11*) introduce a detailed banking sector with financial intermediaries facing endogenously determined balance sheet constraints. *GK11* further contains unconventional monetary policy measures such as governmental asset purchases by the central bank (quantitative easing). Next, we include the model by Christiano et al. (2014) (*CMR14*), which builds on *SW07* and adds a financial accelerator mechanism as in Bernanke et al. (1999). There is idiosyncratic uncertainty in the return on capital of individual entrepreneurs. *CMR14* identifies capital risk shocks to be the main driver of business cycles. Del Negro et al. (2015) (*DNGS15*) also build on *SW07*, adding financial frictions as in Bernanke et al. (1999) and a time-varying infla-

[22] We would like to stress at this point that optimal in our case does not refer to a social planner's optimal policy in terms of maximizing welfare. Rather the focus is on fulfilling pre-determined target values and providing stability. Results therefore crucially depend on the given loss function.

tion target.[23] We add another model emphasizing fiscal policy by Fernández-Villaverde et al. (2015) (*FGKR15*), which includes government expenditure and various taxes as instruments. Finally, we include a large multi-country model, which is used by the International Monetary Fund (Carabenciov et al., 2013). *IMF13* consists of six small country models integrated into a single global market. Special features are for example an unemployment sector, different exchange rates and varying lending options. Financial spillovers between regions are also considered. A more detailed description of all considered models can be found on the MMB web page.[24]

E.0.1.2. The Policy Rules

Common & Optimized Rules. The general form of the policy function added to the models looks as follows:

$$\widehat{i}_t = \beta_\pi^0\,\widehat{\pi}_t + \beta_\pi^1\,\widehat{\pi}_{t-1} + \beta_y^0\,\widehat{y}_t + \beta_y^1\,\widehat{y}_{t-1}. \tag{E.1}$$

Note, that this is a log-linearized version of the original policy function, i.e. the variables represent now log-deviations from their steady states (denoted by hats). Obviously, the constant term in this equation vanishes through the log-linearization around the steady state. We now represent the same common policy rules as in the previous analyses: TR93, NPP and BA. Additionally, we evaluate the RL optimized linear policy versions in the model context.[25] The respective coefficients for (E.1) are given in Table 3.

Optimal Simple Rules. In addition to the aforementioned policy rules, we also calculate optimal simple rules (OSR) for each model used in the comparison.[26] We consider two structural forms which lean on the structure of our RL optimized policy rules (a standard Taylor type rule and one including lags):

$$\widehat{i}_t = \varphi_\pi^0\,\widehat{\pi}_t + \varphi_y^0\,\widehat{y}_t \tag{E.2}$$

$$\widehat{i}_t = \varphi_\pi^0\,\widehat{\pi}_t + \varphi_\pi^1\,\widehat{\pi}_{t-1} + \varphi_y^0\,\widehat{y}_t + \varphi_y^1\,\widehat{y}_{t-1}. \tag{E.3}$$

By solving

$$\min_{\varphi}\; Var(\widehat{\pi}_t) + Var(\widehat{y}_t) + Var(\Delta\widehat{i}_t) \tag{E.4}$$

[23] The time-varying inflation target vanishes for our analysis as we exchange the monetary policy rule with our rules.

[24] See `http://www.macromodelbase.com`

[25] Unfortunately, although producing the best results in the dynamic counterfactual, our nonlinear policy rules cannot be evaluated in the DSGE model context. Due to their nonlinear structure, they would require nonlinear model equations and higher order approximation. Hence, we exclude the nonlinear policy functions from the model comparison exercise. For the same reason, we exclude the ZLB restriction of our rules.

[26] Except for RS99 and CMR14 since it is not possible for these.

subject to (E.2) or (E.3), we find the optimal response coefficients φ for each model.[27] The resulting coefficients are given in Tables E.2 and E.3 in the Appendix. We also compute the mean and the median over the models' OSR coefficients. Since some models require extraordinary large coefficient values, we consider the median to be the better summary statistic over the models. Comparing these to the common rules, the weight on the output gap is quite large, while the inflation coefficient is comparable to the NPP rule.

Table E.1 summarizes the results.

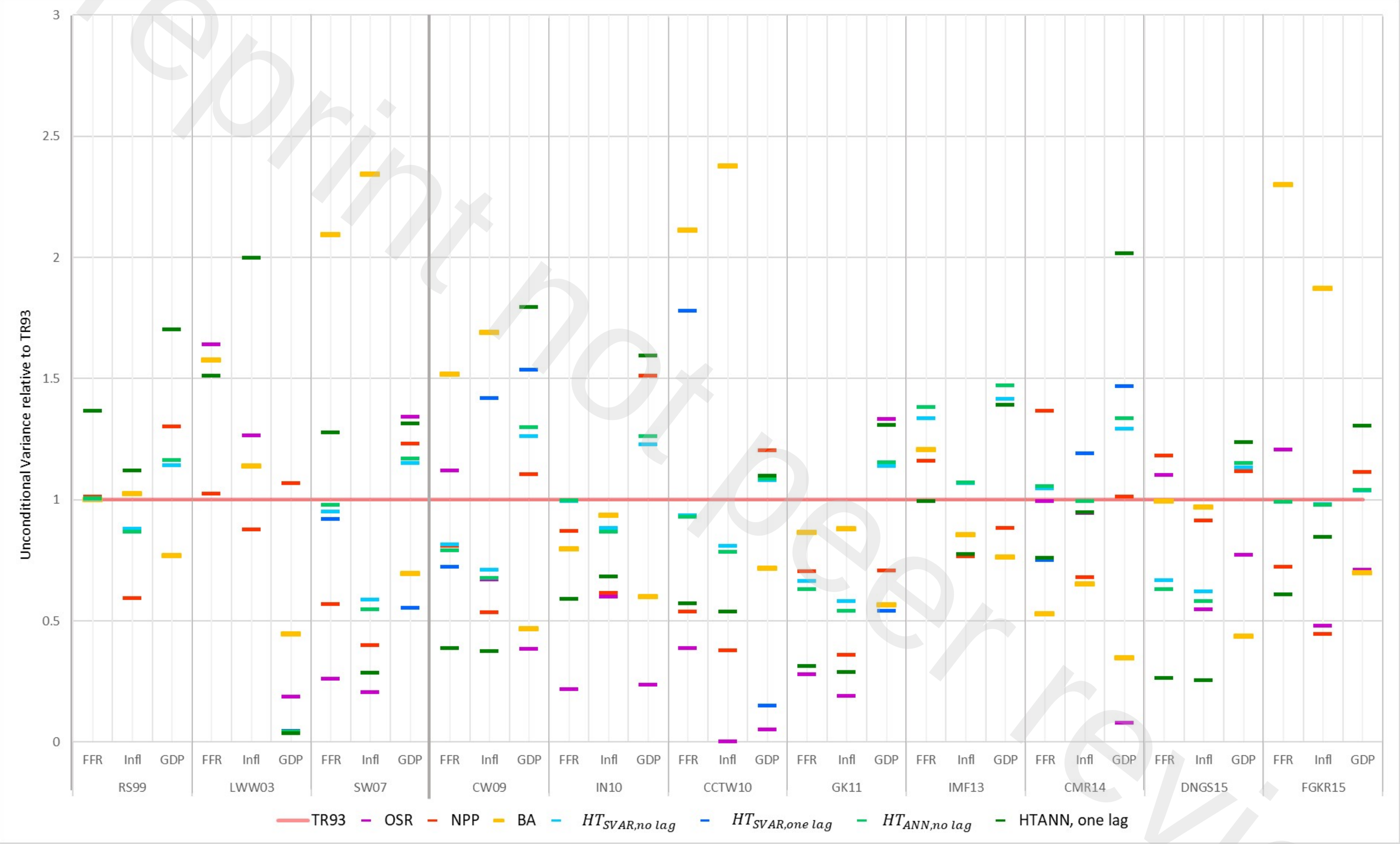
Table E.1: Unconditional Variances Relative to TR93

	All Models					Pre-crisis M.		Post-crisis M.	
Policy	$Var(\widehat{i}_t)$	$Var(\widehat{\pi}_t)$	$Var(\widehat{y}_t)$	$L_{\pi,y,i}$	$L_{\pi,y}$	$L_{\pi,y,i}$	$L_{\pi,y}$	$L_{\pi,y,i}$	$L_{\pi,y}$
$OSR_{no\,lag}$	0.89	0.61	0.72	0.74	0.66	1.00	0.90	0.69	0.63
$OSR_{one\,lag}$	0.80	0.54	0.57	0.64	0.55	0.79	0.67	0.56	0.49
NPP	0.90	0.60	1.11	0.87	0.85	0.90	0.91	0.86	0.83
BA	1.36	1.34	0.59	1.10	0.96	1.23	1.07	1.05	0.93
$RL_{SVAR,\,no\,lag}$	0.17	0.18	1.08	1.13	1.56	1.44	0.99	0.88	1.01
$RL_{SVAR,\,one\,lag}$	4.31	3.77	0.71	2.93	2.24	4.18	2.08	2.31	2.33
$RL_{ANN,\,no\,lag}$	1.15	1.15	0.11	1.14	1.13	1.53	1.42	0.99	1.02
$RL_{ANN,\,one\,lag}$	0.78	0.74	1.34	0.95	1.04	1.18	1.07	0.87	1.03

Note: We calculate the unconditional variances for the nominal interest rate, inflation and the output gap in each model, divide these values by the TR93 values in the respective model and then average over all models. These results are given in columns 1 to 3 ($Var(\widehat{i}_t)$, $Var(\widehat{\pi}_t)$, $Var(\widehat{y}_t)$). $L_{\pi,y,i}$ denotes the relative loss as an average over all three relative unconditional variances, while $L_{\pi,y}$ only averages over inflation and output gap variances. Columns 6-9 report the relative losses for pre- and post-crisis models separately. Results for $RL_{SVAR,\,one\,lag}$ are based on a smaller subset of models due to Blanchard Kahn issues.

[27] As coefficients on inflation and output gap become unreasonably large otherwise, we include the variance of interest rate changes in the objective. Moreover, we consider equal weights on each variance.

Figure E.1: Model Comparison



Note: This figure shows the results of our model comparison exercise in detail. For all included models, the resulting unconditional variances (relative to TR93) of the nominal interest rate (FFR), inflation (Infl) and the output gap (GDP) are given under each policy rule. While the optimal simple rule per model (shown in purple) naturally has very good results in all of the models, we can also see that our optimized rules (shown in green and blue) perform very well. This proves robustness with respect to model uncertainty.

Table E.2: OSR Parameters: No Lag Policy

Model	φ_π^0	φ_y^0
LWW03	1.25	1.99
SW07	2.03	0.24
CW09	6.87	4.63
IN10	3.39	8.06
CCTW10	2.24	0.29
GK11	10.92	6.18
IMF13	2.02	1.00
DNGS15	1.40	2.67
FGKR15	1.68	0.47
Average	3.53	2.84
Median	2.03	1.99

Note: This table shows the policy coefficients resulting from the OSR analysis, minimizing unconditional variances of inflation, output gap and interest rate changes. The policy at hand is obviously the one with two inputs and no lagged variables. We also show the calculated mean and median of the coefficients over the models.

Table E.3: OSR Parameters: Policy with Lags

Model	φ_π^0	φ_π^1	φ_y^0	φ_y^1
LWW03	0.93	0.73	1.31	1.03
SW07	1.23	1.31	0.83	-0.29
CW09	3.42	3.19	2.16	1.68
IN10	3.03	2.05	4.60	3.76
CCTW10	1.57	1.38	0.74	-0.07
GK11	11.46	11.04	0.83	6.01
IMF13	1.57	1.42	0.75	0.56
DNGS15	1.00	0.78	1.87	1.76
FGKR15	0.50	1.68	0.45	0.51
Average	2.75	2.62	1.51	1.66
Median	1.57	1.42	0.83	1.03

Note: This table shows the policy coefficients resulting from the OSR analysis, minimizing unconditional variances of inflation, output gap and interest rate changes. The policy at hand is obviously the one with four inputs and lagged variables. We also show the calculated mean and median of the coefficients over the models.