DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 51/2021

Optimal monetary policy using reinforcement learning

Natascha Hinterlang
(Deutsche Bundesbank)

Alina Tänzer
(Goethe University Frankfurt)

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–95729–861–4
ISSN 2749–2958

Non-technical summary

Research Question

Monetary policy is often described by a simple interest rate reaction function, responding to inflation and the output gap. While this approach is meant to represent the actual interest rate setting behaviour of a central bank, the question arises to what extent the policy function is optimal with respect to fulfilling given inflation and output gap targets.

Contribution

This paper contributes to the discussion by introducing a machine learning based approach called reinforcement learning (RL) to compute optimal interest rate reaction functions. The method allows to incorporate restrictions like the zero lower bound, nonlinear economy structures as well as uncertainty about these. In a first step, we use quarterly U.S. data from 1987-2007 to estimate model transition equations for inflation and the output gap, relying either on a linear structural vector autoregressive (SVAR) model or on a nonlinear artificial neural network (ANN) representation. In the second step, we apply RL to compute optimal reaction functions that are also specified by (nonlinear) ANNs. In doing so, we assume that the estimated model equations remain valid, irrespective of changes in the reaction function.

Results

Concerning the model equations, we find that the ANN specification is able to capture nonlinearities present in the data, improving the fit compared to the SVAR model. The results over the sample from 1978-2007 show that all RL optimized reaction functions outperform other common policy functions as well as the actual observed interest rate. In particular, the nonlinear RL reaction functions stand out positively, as measured by the assumed central bank loss function, penalizing deviations of inflation and the output gap from their targets. A model comparison exercise further indicates robustness of the linear RL reaction functions.

Nichttechnische Zusammenfassung

Fragestellung

Geldpolitik wird oft durch eine einfache Reaktionsfunktion des Zinses auf Inflation und Produktionslücke dargestellt. Während es dabei um eine Beschreibung der tatsächlichen Zinssetzung einer Zentralbank geht, stellt sich die Frage nach der Optimalität: Welche Reaktionsfunktion ist am besten geeignet, um gegebene Ziele für Inflation und Produktionslücke zu erreichen?

Beitrag

Wir stellen einen neuen Ansatz aus dem Bereich des maschinellen Lernens zur Berechnung optimaler Zinsreaktionsfunktionen vor. Die Methode des sogenannten „verstärkenden Lernens“ (reinforcement learning - RL) erlaubt es, Restriktionen wie eine Zinsuntergrenze, nichtlineare ökonomische Zusammenhänge sowie Unsicherheit über diese Zusammenhänge einzubeziehen. Anhand vierteljährlicher US-Daten von 1978-2007 schätzen wir im ersten Schritt Modellgleichungen für Inflation und Produktionslücke. Dazu verwenden wir entweder ein lineares strukturelles vektorautoregressives (SVAR) Modell oder nichtlineare künstliche neuronale Netze (artificial neural networks - ANN). Im zweiten Schritt ermitteln wir anhand von RL optimale Reaktionsfunktionen, die ebenfalls die Form eines (nichtlinearen) ANN haben. Dabei nehmen wir an, dass die zuvor geschätzten Modellgleichungen auch bei Änderungen der Reaktionsfunktion ihre Gültigkeit behalten.

Ergebnisse

Die Ergebnisse zeigen, dass ANN vorhandene Nichtlinearitäten und somit die Daten besser abbilden können als das lineare SVAR Modell. Im Schätzzeitraum von 1978-2007 schneiden alle RL optimierten Reaktionsfunktionen besser ab als übliche Reaktionsfunktionen, und auch als die tatsächlich beobachteten Zinsen. Gemessen an der für die Zentralbank angenommenen Verlustfunktion, die Abweichungen von den Zielwerten für Inflation und Produktionslücke bestraft, stechen insbesondere die nichtlinearen RL Reaktionsfunktionen positiv heraus. Ein Modellvergleich deutet zudem auf Robustheit der linearen RL Reaktionsfunktionen hin.

Deutsche Bundesbank Discussion Paper No 51/2021

Optimal Monetary Policy Using Reinforcement Learning*

Natascha Hinterlang† & Alina Tänzer‡

Deutsche Bundesbank & Goethe University

December 3, 2021

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. The method is generally flexible enough to incorporate restrictions like the zero lower bound, nonlinear economy structures or asymmetric preferences. We use quarterly U.S. data from 1987:Q3-2007:Q2 to estimate (nonlinear) model 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 43 %. A DSGE model comparison exercise further indicates robustness of the optimized rules.

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

*We would like to thank Klaus Adam, Ajit Desai, Refet Gürkaynak, Uwe Hassler, Josef Hollmayr, Andreas Joseph, Sören Karau, Malte Knüppel, Philipp Lieberknecht, Anika Martin, Emmanuel Mönch, Frank Schorfheide, Nikolai Stähler, Harald Uhlig, Volker Wieland, participants at the 23rd INFER Annual Conference and the International Conference on Economic Modeling and Data Science 2021 for helpful comments and discussions.

†Deutsche Bundesbank, DG Economics, Public Finance Division, Wilhelm-Epstein-Strasse 14, 60431 Frankfurt am Main, Germany, Natascha.Hinterlang@bundesbank.de. The views expressed in this paper are those of the authors; they do not necessarily reflect the views of the Deutsche Bundesbank.

‡Chair of Monetary Economics, Goethe University Frankfurt, alina.taenzer@hof.uni-frankfurt.de.

1 Introduction

A simple linear rule can describe actual monetary policy decisions quite well, as shown 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 contributes a machine learning based approach for optimal monetary policy. Specifically, in the first step, we estimate macroeconomic transition equations by artificial neural networks (ANNs), which allows to capture non-specified nonlinearities due to their universal approximator property. In the second step, we model monetary policy as a reinforcement learning (RL) problem, where the central bank learns its optimal reaction function for the nominal interest rate by interacting with the economic environment, which evolves according to the transition equations. While the central bank observes the current state of the economy, it does not know these transition equations.

Such an approach has three advantages. First, due to its highly flexible form, RL can be used to address monetary policy optimization problems under various settings and restrictions. In particular, it allows to take nonlinearities like the zero lower bound (ZLB), convex Phillips curves or asymmetric preferences of the central bank into account. Second, the employed algorithm is model-free, i.e. it does not require complete knowledge of the model equations. Rather, learning occurs from past experiences and through exploration, thereby mitigating the problem of model uncertainty. Third, deep RL, i.e. the combination with multiple layers of artificial neural networks, does not suffer from the curse of dimensionality. It is generally possible to include many state or control variables in the analysis, which might be interesting if one thinks of a larger information set or a larger scope of control variables of the central bank.

Using quarterly U.S. data from 1987:Q3-2007:Q2, we first estimate transition equations for inflation and the output gap, that serve as constraints of the monetary policy optimization problem. We distinguish two cases: a linear economy, estimated by a structural vector autoregression (SVAR), and a nonlinear economy, approximated by artificial neural networks (ANNs). As it turns out, the latter can capture nonlinearities present in the data, improving the data fit by 35 % compared to the SVAR representation.

In the second step, these estimated relations are assumed to be given. The central bank is then provided with a reward function reflecting the dual mandate, which is maximized using RL. The generated policy function is represented by either a linear or a nonlinear ANN, suggesting an optimal interest rate in response to the observed economic stance.

By means of a 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 Fed's monetary policy report (MPR). Relying on the nonlinear model transition equations, a nonlinear reaction function performs best in the counterfactual. While the optimized linear rules reduce the central bank's loss, by over 35 %, the nonlinear reaction functions even boosts the improvement to over 43 %. These quantitative results depend of course 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).

Hence, in order 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 are also robust, generally providing greater stability measured by unconditional variances compared to the common policy rules. Moreover, we find that policy rules including lags of the input variables stabilize inflation and output better than the ones without lags in the DSGE context.

Our paper is related to different streams of the literature. In general, it adds to the literature on optimal monetary policy reaction functions. We would like to emphasize at this point that 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 in which 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 is closely connected to papers that deviate from the L-Q framework by assuming asymmetric preferences, a nonlinear aggregate supply curve or considering the ZLB (e.g. Orphanides and Wieland (2000), Schaling (2004), Dolado et al. (2004, 2005), Adam and Billi (2006)). We also relate to the expanding literature on monetary policy rules vs. discretion following Taylor (1993) (see also Nikolsko-Rzhevskyy et al. (2018) 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)). Concerning robustness analyses using

a comparative DSGE model approach, we further rely on Wieland et al. (2012, 2016).

The one commonality of essentially all papers on optimal monetary policy rules is that the underlying methods are rooted in optimal control theory.[1] It dates back to the 1950s and generally describes a problem of designing a controller in a dynamical system over time such that an objective function is optimized. The key concepts of Dynamic Programming (DP), like the Bellman equation based on Bellman (1957a,b), value function and policy iteration (see Howard (1960)) also constitute the basis of RL theory and algorithms. Instead of going into detail here, we refer the reader to Sutton and Barto (2018) for elaborations on the history of RL and the connection to DP. While not being selective in general, one difference between RL and DP is that the former does not require complete knowledge of the dynamical system. Further, while traditional DP suffers from the curse of dimensionality since the computational requirement increases exponentially with the number of state variables, the impact of dimensionality can be reduced with RL methods that approximate the value function by ANNs. This combination of ANNs and DP dates back to Bertsekas and Tsitsiklis (1996). The particular algorithm of this category that we apply is called deep deterministic policy gradient (DDPG) and was developed by Lillicrap et al. (2015). See also Botvinick et al. (2019) for a survey on the development and general applications of deep RL. There are several papers considering 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, Chen et al. (2021) consider RL also as a method for replacing the assumption of rational expectations within a monetary model. They allow households to learn their optimal policies over time using deep RL and show that the model is solvable this way. 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 historical counterfactuals under different policy rules and evaluate the results of the RL optimized rules. It also includes the DSGE model comparison exercise and a discussion. Section 4 concludes.

[1] See also Hawkins et al. (2015) on the relationship between monetary policy rules and industrial proportional-integral-derivative (PID) controllers.

2 Reinforcement Learning and Data

In this section, we describe the methodology of our machine learning based approach to compute optimal monetary policy, as well as the data we use. Specifically, we employ a branch of machine learning called *reinforcement learning* (RL).[2]

RL refers to learning from the explorative interaction with an environment. There is no need to provide pairs of input and correct output variables, as learning happens through a reward function. The goal of reinforcement learning is to find an optimal policy function that describes a reaction of an agent given observations.

2.1 Structure

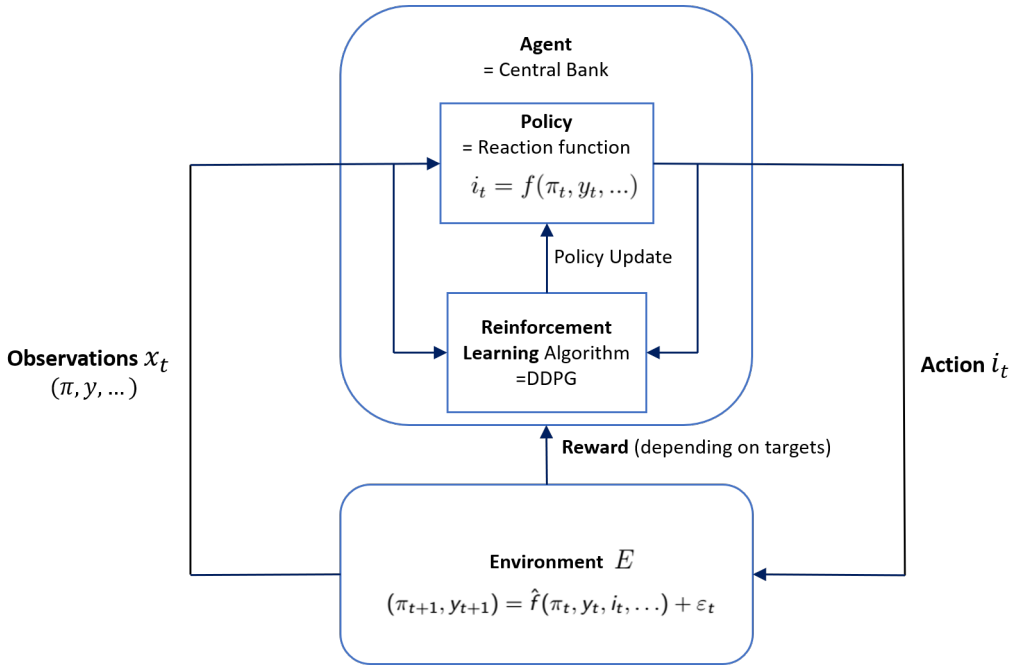
The general idea of RL is shown in Figure 1. The *agent* (central bank) interacts with an unknown environment E, receiving a vector of *observations* x_t (containing inflation π_t and output gap y_t and possibly lags thereof), takes an *action* i_t (nominal interest rate setting) and receives a *reward*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

2.1.1 The Environment

As shown in Figure 1, RL requires the provision of an environment E, that 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. Concerning the variable choice for the whole economy, we think of 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 (NK Phillips curve) and a central bank reaction function. Since the task of the environment is to provide

[2]The other two branches are unsupervised and supervised learning, which are used to find a hidden structure in unlabeled data or to approximate an unknown functional relationship in order to make predictions, respectively. We also make use of the latter when estimating the economy representation by an ANN.

Figure 1: Reinforcement Learning Scheme



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. With respect to the specific functional form of the equations, we consider two scenarios. First, we assume a standard linear model structure, i.e. Equation (1) and (2) collapse to the well-known reduced form representation of a vector autoregression (VAR). Second, we use artificial neural networks to approximate the economy, which allows for nonlinear relationships among variables while being agnostic about the specific functional forms.

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

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

where C^m and α^m, $m \in \{\pi, y\}$, represent the constant and the vector of coefficients, respectively. In order to be able to conduct the historical counterfactual analysis later on, we further need to transform the reduced form to its structural (SVAR) representation. Usually, the first step would be to estimate the reduced form VAR, followed by a shock identification procedure through a Cholesky decomposition. Instead, 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. 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:

$$s_t^y = (y_{t-1}, \pi_{t-1}, i_{t-1}, i_{t-2}) \tag{4}$$

$$s_t^\pi = (y_t, y_{t-1}, y_{t-2}, \pi_{t-1}, \pi_{t-2}, i_{t-1}). \tag{5}$$

By restricting our SVAR in that 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:

$$y_t = C^y + a_{y,1}^y\, y_{t-1} + a_{\pi,1}^y\, \pi_{t-1} + a_{i,1}^y\, i_{t-1} + a_{i,2}^y\, i_{t-2} + \varepsilon_t^y \tag{6}$$

$$\pi_t = C^\pi + a_{y,0}^\pi\, y_t + a_{y,1}^\pi\, y_{t-1} + a_{y,2}^\pi\, y_{t-2} + a_{\pi,1}^\pi\, \pi_{t-1} + a_{\pi,2}^\pi\, \pi_{t-2} + a_{i,1}^\pi\, i_{t-1} + \varepsilon_t^\pi. \tag{7}$$

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. In order to do so, we estimate the transition equations by 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 $\hat{f}^m$, $m \in \{y, \pi\}$ is given by

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

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 so-called 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.

As shown by Hornik et al. (1989), ANNs have the property of being universal approximators, i.e. they can approximate any function to an arbitrary degree. We use Matlab's nonlinear autoregressive neural network with exogenous inputs (NARX), which is a so-called single-hidden-layer neural network.[3] The structure of the ANN corresponding to (8) is illustrated in Figure 2 for $m = y$. Each network consists of an input layer for the explanatory variables. The hidden layer comprises hidden units (also called nodes or neurons), which represent activation functions.[4] The weighted inputs $\omega_j^{m'} s_t^m$ are summed up, a bias term is added and the sum is transferred by the activation function (G). We employ hyperbolic tangent sigmoid functions $\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$, which map on the interval $[-1, 1]$.[5] The sum of weighted outputs of each neuron plus another constant bias term finally yields the output of the last (linear) layer, which is the dependent variable of the left hand-side in equations (4). The structure of (5) looks analogously.

The number of hidden units h represents a hyperparameter that we determine by dividing the sample into a training and validation set[6], where we use the last 15 % of observations for validation, i.e. they are not used during training. We then loop over one to ten hidden units using 30 different random initial weights each and choose the number of hidden units with the lowest mean squared error in the validation set averaged over the 30 trials.[7] It turns out

[3] We use Matlab2019b version for all analyses.

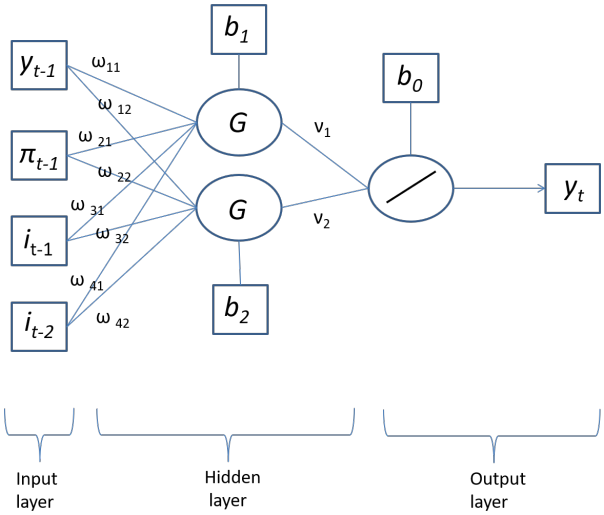
[4] To simplify the figure, only 2 nodes are included here, which does not reflect the actual chosen number.

[5] This activation function is usually used because of faster convergence rates (see e.g. LeCun et al. (2012)).

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

[7] This model selection strategy is similar to the one of Aras and Kocakoç (2016). The ANNs are initialized using the Nguyen-Widrow method.

Figure 2: Structure of a Single-hidden-layer Artificial Neural Network with Two Hidden Nodes



that the optimal number of hidden units of the ANN representing the output gap and inflation are 3 and 4, respectively. After fixing the optimal number of hidden units, we take the set of initial weights and biases that produces the lowest overall mean squared error. The networks are trained using the Levenberg-Marquardt algorithm (see Levenberg (1944) and Marquardt (1963)).

We wish to make clear at the outset that during 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 it would be the case with models including rational expectations. Orphanides and Wieland (2000), for example, follow a similar approach when analyzing optimal monetary policy under inflation zone targeting without the explicit modelling of expectations. Given that only the parameters of the policy function are optimized during RL, while the overall target remains the same, the effects on the transition equations' parameters should be less critical.[8] While expectations certainly play a role, the degree of rationality is uncertain and the estimated transition equations can still serve as a useful benchmark environment to introduce the RL concept.

[8]One could also think of private agents adapting expectations only very slowly over time.

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 the policy parameter updating. Further, we define an *actor*, representing the central bank policy function, describing the nominal interest rate setting behaviour in response to observations from the environment. Concerning the critic, we use nonlinear neural networks in order 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 in order 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 (4)-(5), 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 one side, we aim for a fair comparison to standard Taylor type rules. On the other side, 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 further add a zero lower bound (ZLB) restriction on the nominal interest rate (i.e. $i_t \in \mathbb{R}^+$), which is easily implemented in a neural network structure using a rectified linear unit output (ReLu) layer.[9] It performs a threshold operation such that every input of this layer less than zero is set to zero:

$$f(x) = \begin{cases} x, x \geqslant 0 \\ 0, x < 0 \end{cases} . \tag{11}$$

[9]ReLu also often replaces tanh as an activation function, see e.g. Nair and Hinton (2010).

This allows to take into account the nonlinear restriction through the ZLB during the learning process. The ZLB might change the optimal parameters of the reaction function as shown by e.g. Adam and Billi (2006). Apart from the ZLB restriction, we start our analysis employing linear neural network policy structures. This approach allows for a direct comparison to other common linear policy rules before turning to a more liberate, nonlinear functional form. The structural form of the resulting policy function P_t is given by:

$$P_t = i_t = \max\{f(x_t^z), 0\}, \qquad \text{where} \tag{12}$$

$$f(x_t^z) = \alpha_0 + \sum_{j=1}^{q} \delta_j G(\beta_j' x_t^z + \alpha_j) \tag{13}$$

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 (13) is the representation of a single-hidden-layer feed-forward neural network as we used it for approximating the economy in (4) and (5).[10] In the linear policy case, $q = 1 = \delta_j = 1$ and $G()$ collapses to the *purelin* transfer function, that simply maps the input value onto itself ($a = purelin(n) = n$). 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, 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 In order to adjust the policy coefficients in an optimal way, one needs to determine the respective action value function. It works as a measure of performance for policy interventions and thus constitutes the basis for policy updates:

$$Q^P(x_t, i_t) = \mathbb{E}[R_t | x_t^z, i_t] \tag{14}$$

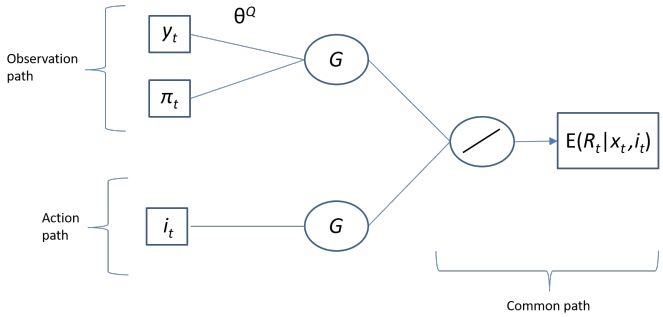
with

$$R_t = \sum_{i=t}^{T} \gamma^{i-t} r_t(x_i^z, i_i). \tag{15}$$

[10] 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.

The function Q^P describes the expected return after taking action i_t, observing state x_t, and following policy P_t thereafter. This recursive relationship is based on the Bellman equation. The return R_t itself, is defined as the sum of discounted future rewards with a typical discount factor of $\gamma = 0.99$ (cf. e.g. Svensson (2020)). The *critic* approximates (14) with an ANN as shown in Figure 3, where θ^Q comprises all connection weights and biases of the network that are omitted in this graph for simplicity. Via the observation path, the set of observable variables x_t^z ($z = 1$ in the Figure) enters the critic. Through the action path, the control variable i_t is included. Both paths are concatenated to the common path and its output is the expected long-term reward based on observed state and action, i.e. equation (14).

Figure 3: Structure of the Critic Neural Network Approximating the Value Function



We follow the Fed's mandate when defining the objective of our agent. According to the Federal Open Market Committee's (FOMC) statement on longer-run goals and monetary policy strategy[11], an inflation rate of 2 % is "most consistent over the longer run with the Federal Reserve's statutory mandate". Further, it tries to promote maximum employment, while its specific level is allowed to vary over time. The two objectives are seen as complementary in general. Hence, we rely on the standard quadratic reward function that is given by

$$r_t(x_t^z, i_t) = -\omega_\pi \left(\pi_{t+1} - \pi^*\right)^2 - \omega_y \, y_{t+1}^2 \tag{16}$$

with equal[12] $\omega_\pi = \omega_y = 0.5$ and $\pi^* = 2\%$.[13] We would like to emphasize at this point, that

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

[12] Equal weights on inflation and the unemployment gap actually translates 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 the one of inflation improves social welfare in a DSGE model context.

[13] For computational reasons, the continuous reward function given by (16) is accompanied by a second part, which punishes deviations from targets that exceed 2 percentage points: $r_t^{\pi_p} = 10 \cdot r_t^\pi \,(\text{if}\, r_t^\pi > 4)$ and $r_t^{y_p} = 10 \cdot r_t^y \,(\text{if}\, r_t^y > 4)$, where the subscript p stands for penalty and r_t^π and r_t^y denote the squared deviations from

it is generally possible to analyze optimal monetary policy under different 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).[14]

2.2 Training Algorithm and Hyperparameters

We employ a reinforcement learning algorithm first presented by Lillicrap et al. (2015), called *Deep Deterministic Policy Gradient* (DDPG) algorithm, which is implemented in Matlab 2019a (and later versions).[15] It 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.[16] 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 C.1. Before entering such a training cycle, we have to decide about the *critic* network structure. To find the best layer structure in the end, 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, there is no further choice involved.[17] But when we optimize the nonlinear policy version in (13), we also need to determine the number of hidden units of the actor. In this case, we loop over these *actor* nodes also from one to ten, while fixing the critic nodes.

One training cycle consists of different steps. It starts with an initialization phase and each

targets. The penalty rewards are added to (16). This kind of *mixed reward signal* drives the system away from bad states while simultaneously promoting convergence.

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

[15]There exist many different RL algorithms. Among other things, the specific choice depends on the observation and action spaces, i.e. whether they are discrete or continuous, if it is based on a value or an action-value function, and how the actor is modeled. We decided for the DDPG since it is capable to handle 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.

[16]The term *model-free* relates to the fact 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. In contrast, *on-policy* learning means that it only uses experiences generated by the latest learned policy (behavioural policy).

[17]Remember, that the linear version of (13) includes setting $q = 1$.

cycle consists of $M = 500$ episodes in total.[18] Steps *f)* to *m)* are repeated until the agent either fulfills our defined stopping criteria, which is $1.7 < \pi_{t+1} < 2.3$ and $-0.3 < y_{t+1} < 0.3$, i.e. 0.3 p.p. absolute deviations from target values, or the episode stops automatically after a maximum of $T = 50$ quarters.[19]

Table 1: DDPG Algorithm

Initialization	a)	Randomly initialize critic network $Q(x, i\|\theta^Q)$ and actor $P(x\|\theta^P)$ with weights θ^Q and θ^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 $x_0^z, \quad (z = 1 \text{ or } 2)$
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: $\nabla_{\theta^P} J \approx \frac{1}{N}\sum_j [\nabla_i Q(x, i\|\theta^Q)\|_{x=x_j, i=P(x_j)} \nabla_{\theta^P} P(i\|\theta^P)\|_{x_j}]$
	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 for details.

As indicated in g), the reward r_t is calculated according to our designed reward function (16) in each step t. To get a measure of performance of one training episode, the episode reward is calculated as the sum over the rewards per step ($ER_m = \sum_t r_t$). 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 (16), 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

[18] 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.

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

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 $1 < ES_m < 50$ to avoid, on 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 is reached, before the maximum number of steps per episode is reached. Afterwards, we calculate the steady state of each saved agent, which represent the long-term equilibrium of the economy and derive the respective steady state reward according to (16). 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.

Table 2 summarizes the chosen numbers of hidden nodes for the six different cases under investigation. Except for the version with the linear economy combined with policy inputs x_t^2, two nodes in the critic network yields the best results. Hence, in order to determine the number of nodes for the nonlinear policy case, we fix the number of nodes in the critic at two, loop over one to ten nodes in the actor and choose the best one according to the steady state reward as above. It turns out that with observation inputs x_t^1 and x_t^2, ten and eight nodes in the policy function produce the best results, respectively.[20]

Table 2: Chosen Numbers of Hidden Nodes

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

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.

[20] Since the former case touches the upper bound of the pre-defined loop, we also experimented with more than 10 nodes, which did not improve the results.

2.3 Data

In our benchmark analysis, we use quarterly U.S. data from 1987:Q3 to 2007:Q2. The period's starting point coincides with the appointment of Alan Greenspan as the Fed's chairman, while it stops before the financial crisis. We chose this time span since inflation (and unemployment) targeting was implicitly practiced (see Goodfriend (2004)). Periods following the financial crisis are excluded since these are characterized by unconventional monetary policy measures like large-scale asset purchases. While we do account for a ZLB constraint of the nominal interest rate, which was binding during the crisis, we do not consider additional instruments of the central bank nor their possible combinations. Hence, our optimized policy rules rather correspond to "normal" times taking into account the lower bound restriction. However, in Appendix C.2.3, we also show static interest rate prescriptions of the considered policy rules for the periods after the great financial crisis.

Inflation π_t is measured by the GDP implicit price deflator as the percentage change from one year ago.[21] The output gap y_t is computed as the percentage deviation of actual GDP from its potential. For the latter, we use estimates of the U.S. Congressional Budget Office. Whenever we plot the actual behavior of the central bank, we mean the effective federal funds rate.[22]

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 following learning steps, inflation and output gap data is 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 start with presenting the fit of the estimated economy representations. Afterwards, we compare parameters of our RL based optimal monetary policy functions to

[21] The Fed 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 analysing the robustness of the optimized rules. For the same reason, we use the output gap instead of the targeted unemployment rate.

[22] All time series were downloaded from the FRED website. We performed Augmented Dickey Fuller (Dickey and Fuller (1979)) and KPSS tests (Kwiatkowski et al. (1992)) that indicated stationarity of the three series.

those of other common reaction functions, before we turn to historical counterfactual analyses. In addition, we show robustness of our optimized policy rules with respect to model uncertainty by means of a DSGE model comparison exercise.

3.1 Economy Representations

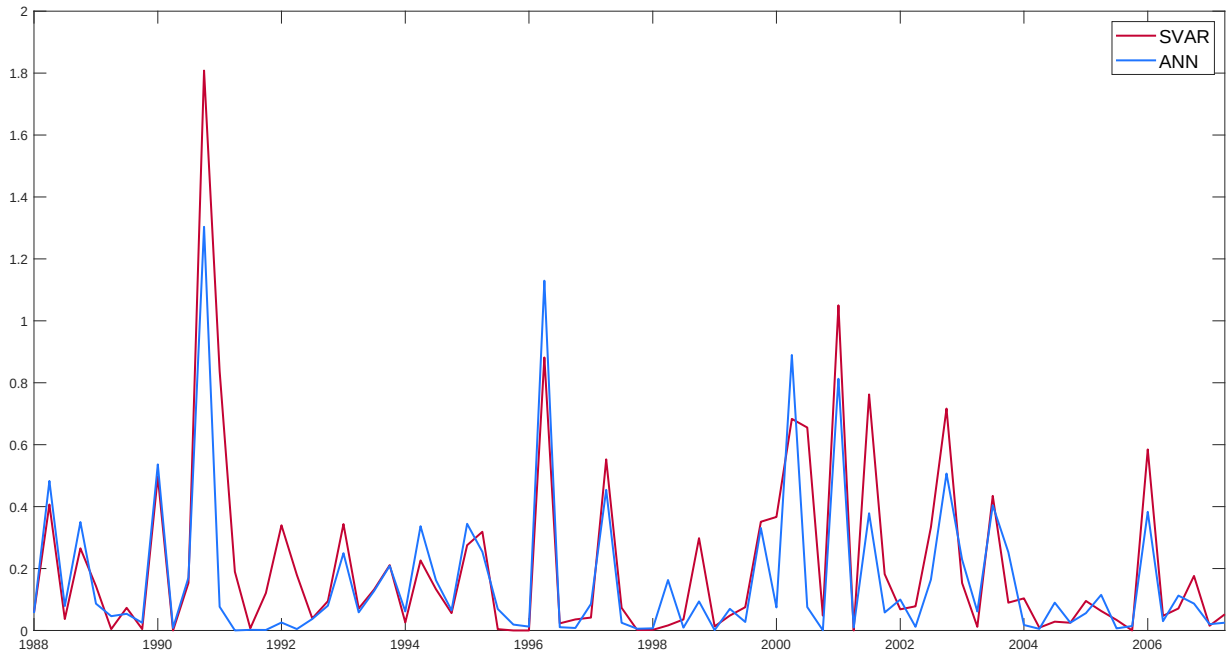
Before we present our results on RL based optimal monetary policy, we illustrate the differences between our linear SVAR ((6)-(7)) and nonlinear ((4)-(5)) economy representations. Table 9 in the Appendix summarizes the estimation results for the SVAR representation. The Durbin-Watson as well as the Lagrange 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. The SVAR representation, however, is restricted by its predetermined linear form and might consequently miss certain dynamics of the actual time series data.

Figures 4 and 5 compare the fit of the linear economy model (in red) and the ANN economy (in blue) for inflation and output gap, respectively. Specifically, we compute the differences between the fitted and the actual time series and plot the squared errors. Larger values directly can be interpreted as a worse fit. While considering the output gap, the difference between SVAR and ANN model is less pronounced, the nonlinear model clearly yields a better fit for inflation. Concerning the timing, the results indicate that the ANN outperforms its linear opponent especially during crisis periods. For the recession in the early 1990s after the stock market crash in 1989 and the recession in the early 2000s, the ANN yields lower squared errors. The superior performance gets even clearer when comparing the mean squared errors (MSE) of Table 3. Using the ANN to approximate the economy, the overall fit of the output gap and inflation variables can be improved by 22 % and 48 %, respectively. Regarding the total economy, i.e. averaging over the MSE for the output gap and inflation, the ANN outperforms the SVAR by 25 %. 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.[23]

Visualizing these nonlinearities is not an easy task. The marginal relationship between input and output variables in an ANN are not constant as in the linear case, but depend on the levels of the input variables. Since inflation and output gap are functions of six and four

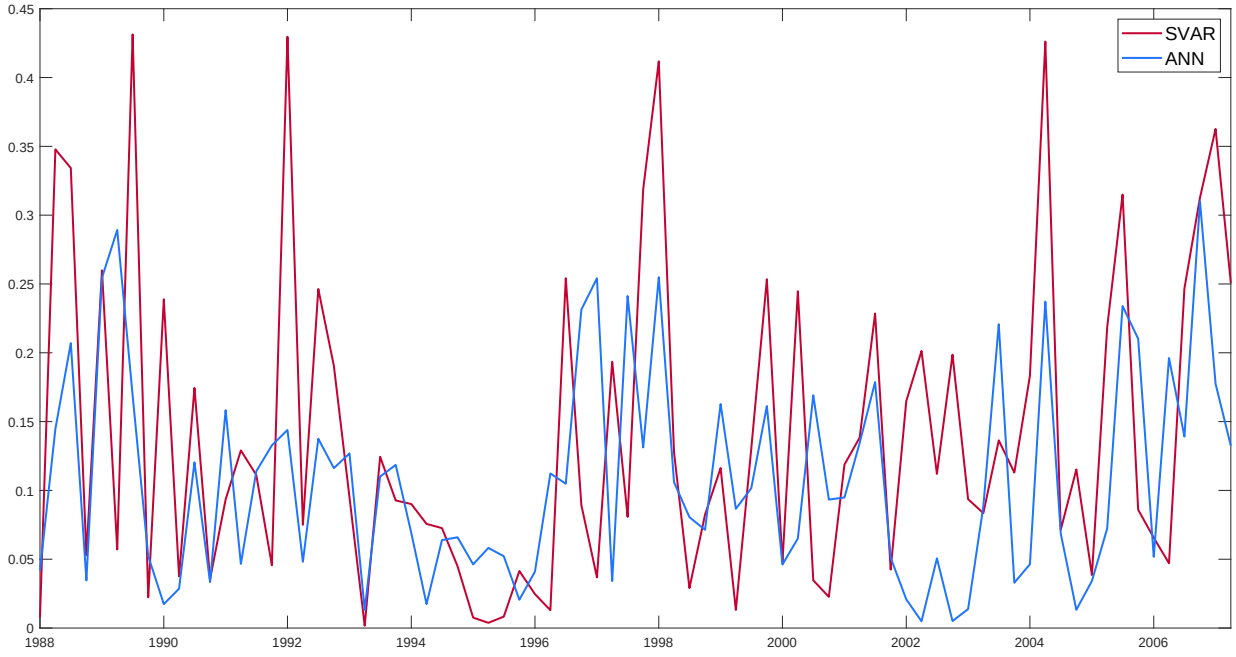
[23]The ANN also outperforms the SVAR by a similar magnitude when focusing on the validation set only. Hence, the superiority of the ANN should not be due to an overfitting of the training sample.

Figure 4: Output Gap Fit: Squared Errors



Note: This figure shows the squared errors between the actual output gap time series (1987:Q3 to 2007:Q2) and the fitted values of the SVAR and ANN model.

Figure 5: Inflation Fit: Squared Errors



Note: This figure shows the squared errors between the inflation time series (1987:Q3 to 2007:Q2) and the fitted values of the SVAR and ANN model.

Table 3: Economy Fit: Mean Squared Errors

Representation	MSE Output Gap	MSE Inflation	MSE Total
SVAR	0.211	0.033	0.122
ANN	0.165	0.017	0.091

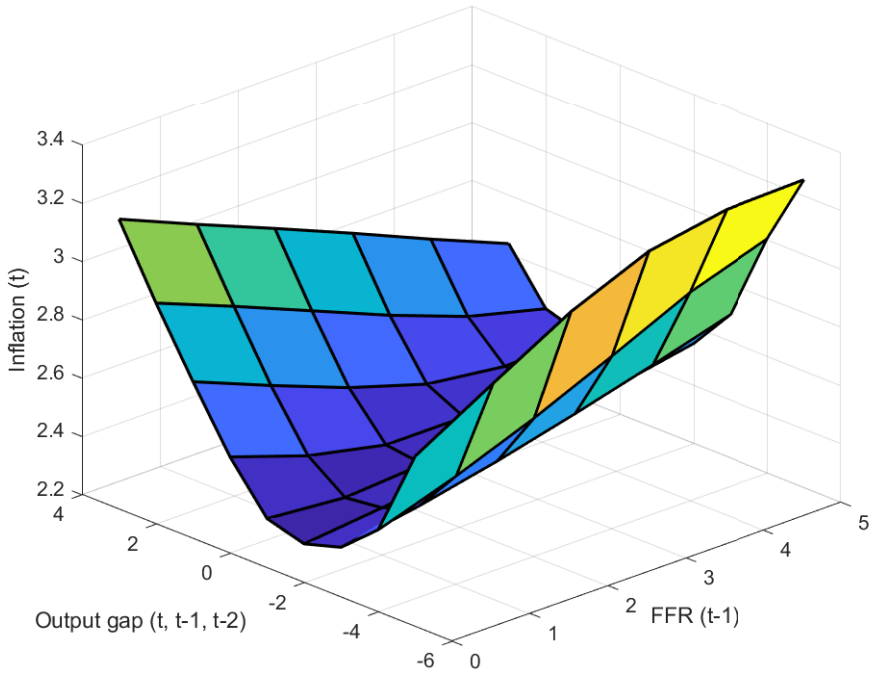
Note: This table summarizes the mean squared errors of the linear SVAR ((6)-(7)) and ANN ((4)-(5) economy representations for the variables output gap, inflation and the overall economy.

explanatory variables, respectively (cf. (4)-(5)), all functional dependencies cannot be plotted. Still, we can visualize parts of it by using partial dependence (PD) plots. These plot output predictions against a single or a pair of input variables by marginalizing out the effects of the (potential) remaining variables.[24] Figures 6 and 7 represent PD surface plots for inflation and output gap, respectively. For positive output gaps, the relations concerning inflation are as expected: inflation decreases with the nominal interest rate, and increases with the output gap. Surprisingly, this turns around for negative output gaps. There, inflation increases with the nominal interest rate, and decreases when the output gap increases. This somehow symmetric response of inflation with respect to the output gap might be explained by our sample period. While inflation and output gap move in the same direction at the beginning and the end of our sample, during the 1990s, we see a closing of the output gap (starting from a value of -4 following the stock market crash), while inflation decreases, among other things due to a cheaper supply of computer technology. Remember that following a demand shock, inflation and output gap move in the same direction, while they move in opposite directions after a supply shock. We have to keep in mind that the ANN represents an estimated relationship, driven by our sample observations, that may change over time. Still, it is an interesting outcome of our ANN economy, that a simple linear representation with a constant partial derivative could not produce.

Turning to the PD surface plot of the output gap, we observe less nonlinearities, corresponding to the smaller improvement in fit (see Table 3). As expected, the output gap decreases with the nominal interest rate, independent of the level of inflation. The relationship between inflation and the output gap, however, depends on the level of inflation. As long as inflation is below its target, output gap increases with inflation, while it decreases with inflation above the 2 % target.

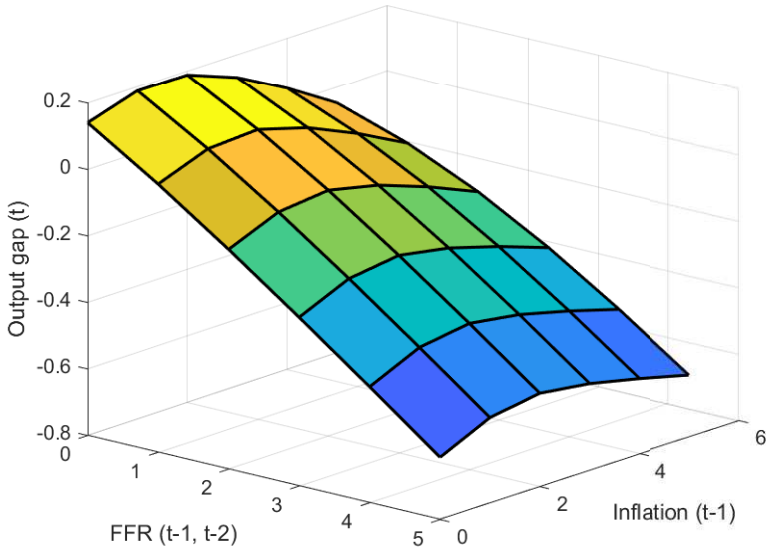
[24] In order to produce these plots, we build a grid for inflation, output gap and interest rate from 0:6, -5:3 and 0:5, respectively, and compute the corresponding outputs of the ANN. We then assume constant values across lags and marginalize over respective remaining variables to reduce the dimensions for the surface plot.

Figure 6: Partial Dependence Surface Plot - ANN Economy, Inflation



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

Figure 7: Partial Dependence Surface Plot - ANN Economy, Output Gap



Note: This figure 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}.

3.2 Optimized Policy Parameters

In total, we compute six optimal monetary policy rules: two linear ones within the SVAR economy framework and four (two linear and two nonlinear ones) within the ANN economy.[25] We denote the policies based on our approach by RL, where the subindex specifies if it is based on the linear (SVAR) or the nonlinear economy (ANN), which input structure and which functional form (linear vs. nonlinear) is used. *No lag* means observations x_t^1 serve as inputs, whereas *one lag* corresponds to x_t^2 (see (9)-(10)).

Linear Policies Table 4 summarizes the optimized coefficients of the linear policy rules. For the sake of comparability, we also show the coefficients of the original Taylor (1993) rule (TR93) and the so-called Balanced-approach (BA), which are both included in the Fed's MPR.[26] Further, we consider the so-called inflation tilting rule brought up by Nikolsko-Rzhevskyy et al. (2018) (NPP). The general form of 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. In our SVAR economy, the best simple structured agent $RL_{SVAR,\,no\,lag}$, i.e. using (9) as inputs, has an inflation coefficient of 2.54 and is thus best comparable to the NPP rule. $RL_{SVAR,\,no\,lag}$ consequently responds more aggressively to deviations of inflation from its target value compared to TR93 and BA. In contrast, the output gap coefficient β_y^0 is slightly smaller with 0.42 than all three common policy rules and shows a smaller compensating tendency of GDP deviations from potential. We find, that the constant term α_0 is quite similar across TR93, BA and $RL_{SVAR,\,no\,lag}$. Using the intercept relation $\alpha_0 = r^* - (\beta_\pi - 1)\pi^*$, one can back out 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 $r^* = 4.2\,\%$ is considerably larger than the assumed value of $2\,\%$ by Taylor (1993) and also larger than the ones implied by BA and NPP.

Within the linear economy framework, we further optimize a second policy rule with lagged inputs as given in (10), $RL_{SVAR,\,one\,lag}$. With $\beta_\pi = \beta_\pi^0 + \beta_\pi^1 = 2.99$, its sensibility to inflation deviations is even larger than with $RL_{SVAR,\,no\,lag}$. This rule also reacts stronger to output gap

[25] Note, that linear and nonlinear here refers to the area outside the ZLB. Strictly speaking, all rules are nonlinear due to the ZLB constraint.

[26] 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.

Table 4: 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	
$RL_{SVAR,\,no\,lag}$	1.14	2.54	-	0.42	-
$RL_{SVAR,\,one\,lag}$	0.25	2.30	0.69	1.75	-1.14
$RL_{ANN,\,no\,lag}$	0.86	1.35	-	0.78	-
$RL_{ANN,\,one\,lag}$	0.24	0.66	1.05	0.15	0.79

Note: As introduced in the general policy function structure in equation (13), α_0 is the constant term, with $\alpha_0 = r^* - (\beta_\pi - 1)$, π^* 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.

deviations than the aforementioned rules ($\beta_y^0 + \beta_y^1 = 0.6$). The much smaller constant of 0.25 compared to our simple policy version combined with the larger inflation response coefficient yields the same estimate of $r^* = 4.2\,\%$ as before.

The subsequent policy rules are optimized based on the environment represented by the ANNs of (8). We again start with the simple structure containing no lagged input variables ($RL_{ANN,\,no\,lag}$). While $\beta_\pi^0 = 1.35$ comprises a more subtle reaction to inflation variations, which is even smaller than in TR93 and BA, the output gap is strongly reacted to ($\beta_y^0 = 0.78$). A constant of 0.86 translates into $r^* = 1.56\,\%$ which is much smaller than the one found under the linear economy. The best agent after adding lags to the policy rule ($RL_{ANN,\,one\,lag}$) yields inflation response coefficients that are much more moderate than under $RL_{SVAR,\,one\,lag}$. They sum up to 1.71, which falls in-between TR93/ BA and NPP. However, the sensibility to the output gap is increased, which is mainly driven by the coefficient on lagged output gap. With 0.94 in sum, $RL_{ANN,\,no\,lag}$ shows the strongest reaction to GDP deviations from potential. The implied equilibrium real interest rate amounts to 1.6 % in this case. We do not want to dig deeper into the r^* discussion at this point as our focus is on the performance of the rules optimized through RL. However, it is remarkable that changing the environment from linear to nonlinear leads to a much smaller implied equilibrium real interest rate. Figures 13 and 14 in the Appendix further show for which inflation and output gap combinations the ZLB binds according to our RL optimized rules.

Nonlinear Policies As a last step, we allow the policy function to be of a nonlinear form as shown in (13) 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 by PD plots.

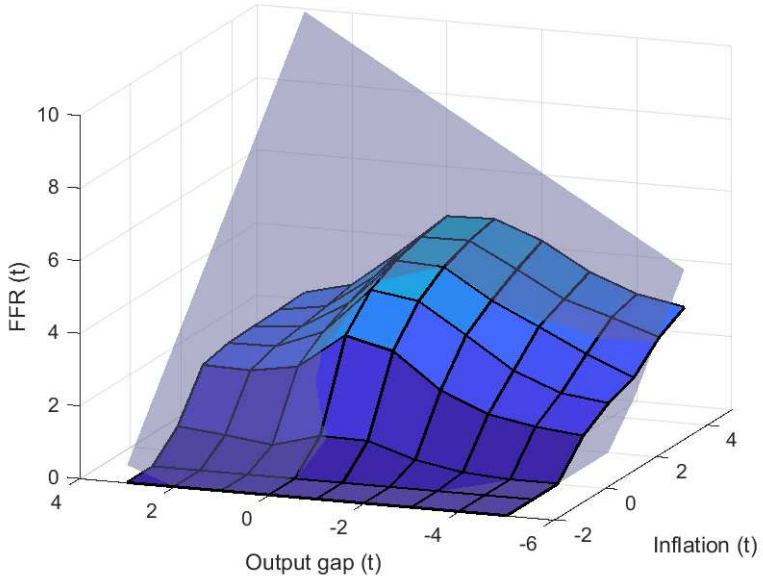
Figure 8 represents a PD surface plot for $RL_{ANN,\,no\,lag,\,nonlin}$. This optimized rule has only two inputs, π_t and y_t. Hence, the inputs' influence on the output variable can be illustrated by a three-dimensional surface plot without the need for marginalizing over other variables. For comparison, we add the optimized linear policy $RL_{ANN,\,no\,lag}$ to the plot. It shows that $RL_{ANN,\,no\,lag,\,nonlin}$ prescribes a zero nominal interest rate as long as inflation falls below zero, irrespective of the output gap. Given the output gap is closed, i.e. actual GDP equals potential such that $y_t = 0$, the federal funds rate (FFR) increases with inflation reaching a value of 3.9 at $\pi_t = \pi^* = 2$. As expected, the implied interest rate declines when the output gap falls into negative territory. Surprisingly, however, the FFR under $RL_{ANN,\,no\,lag,\,nonlin}$ is also smaller for positive output gaps combined with inflation values above its target.[27] Figure 15 in the Appendix shows PD line plots for inflation and output gap each by averaging over the respective remaining variable. Graphically, these lines represent average cross sections of the nonlinear policy in Figure 8. It underlines the almost symmetric relationship between prescribed FFR and the output gap around zero. Contrary, under the optimized linear policy $RL_{ANN,\,no\,lag}$ (the transparent plain) the FFR is less responsive to inflation at lower inflation values, and more responsive for larger inflation values compared to the nonlinear rule.

Figure 9 illustrates how i_t relates to π_t and y_t under $RL_{ANN,\,one\,lag,\,nonlin}$ by holding values constant over the lags of inflation and output gap.[28] As before, we plot this rule against its linear counterpart. The FFR under $RL_{ANN,\,one\,lag,\,nonlin}$ generally increases with inflation (except for very negative output gap values) and output gap, reaching a plateau with FFR values of around 5.5 for positive output gaps combined with inflation values around 4. Figure 16 in the Appendix further shows the PD line plots for inflation and output gap individually, which also illustrates the irresponsiveness of the rule with respect to the output gap for $y_t > 2$. The comparison with the respective linear optimized policy $RL_{ANN,\,one\,lag}$ reveals that the FFR

[27] This might be influenced by the fact that inflation and output gap have maximum values of 4.2 and 2.4 in our data sample, respectively, and the optimization algorithm may have never encountered such value combinations.

[28] 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 8: Partial Dependence Surface Plot - $RL_{ANN,\,no\,lag,\,nonlin}$ vs. $RL_{ANN,\,no\,lag}$



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 $RL_{ANN,\,no\,lag,\,nonlin}$. The transparent plain represents the corresponding linear counterpart $RL_{ANN,\,no\,lag}$.

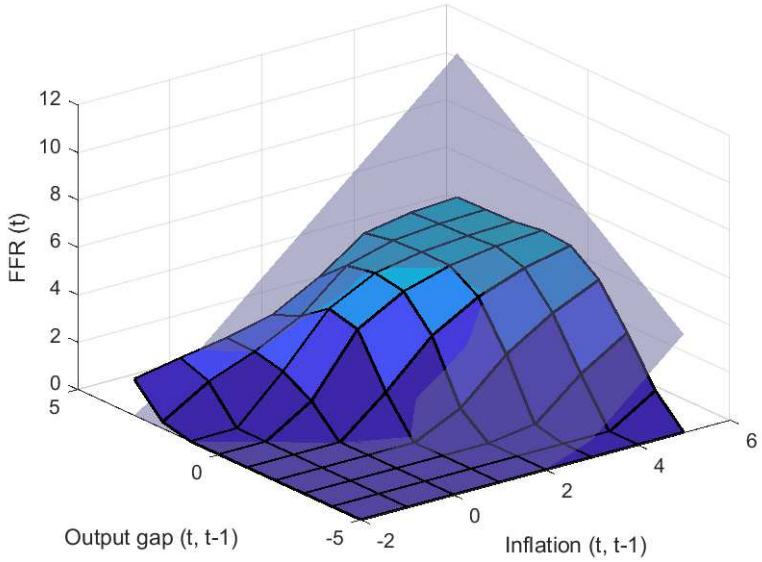
is held at zero for similar values of inflation and output gap. Compared to the linear rule, at output gap values around zero, the nonlinear rule prescribes a sharper increase of FFR for inflation values between 0 and 3, whereas the response is less pronounced for larger inflation values. The plateau at high inflation and output gap values can per definition not be captured by the linear rule.

Overall, results for both nonlinear rules suggest that it is optimal not to increase the interest rate when the output gap increases if the output gap is sufficiently positive large. Given a closed output gap, both nonlinear rules prescribe larger FFR values for inflation values between 0 and 3 and lower ones for $\pi_t > 3$ compared to their linear counterparts. Still, it might also be the case that the optimized rules stabilize the economy in a successful way, such that these large deviations from targets are not encountered (during simulation) at all.

3.3 Historical Counterfactuals

In order to evaluate our RL based reaction functions, we need to compare their performance with the actual interest rate setting behaviour of the Fed and alternative rules. Often, different policy rules are compared by using a static setup, which means that data on inflation and the output gap is simply plugged into each rule without considering any feedback mechanisms.

Figure 9: Partial Dependence Surface Plot - $RL_{ANN,\,one\,lag,\,nonlin}$ vs. $RL_{ANN,\,one\,lag}$



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, assuming $\pi_t = \pi_{t-1}$ and $y_t = y_{t-1}$, under $RL_{ANN,\,one\,lag,\,nonlin}$. The transparent plain represents the corresponding linear counterpart $RL_{ANN,\,one\,lag}$

Results of this exercise can be found in the Appendix for the sake of completeness. However, one cannot draw a conclusion on which policy is best suited to reach target values from such an analysis.

Therefore, we conduct counterfactual analyses that take the dynamics of inflation, output gap and interest rate as well as feedback effects into account. Similar to Primiceri (2005), we use our estimated SVAR (6)-(7) and the structural shocks thereof $\varepsilon_{it}, i = 1, 2$ to simulate the economy under different reaction functions. Specifically, we exchange the third equation of the dynamical system by the respective (optimized) policy rules, while keeping (6)-(7) unchanged. Equivalently, we compute counterfactuals using the ANN economy (8). The dynamic counterfactual simulation period lasts from 1987:Q3 to 2007:Q2. Since we have to take the estimated structural parameters of the economy in (6)-(7) and (8) as given and unchanged, the Lucas (1976) critique applies, i.e. the behaviour of rational and forward-looking private agents might be different when they take the change of policy into account. However, we are convinced that the effects of the Lucas critique are rather small, since we do not compare policies from totally different regimes. In contrast, we only change the policy within a period, where inflation targeting was already practiced, using the same technique as in the e.g. Primiceri (2005) and

Sims and Zha (2006). Hence, possible behaviour modifications of the private agents should be minor.[29] 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 work of the RL approach.

SVAR Economy & Linear Policy We start with the economy in linear form ((6)- (7)).[30] Figure 10, shows the simulated counterfactual time series for the interest rate, inflation and the output gap, respectively, under the different reaction functions.

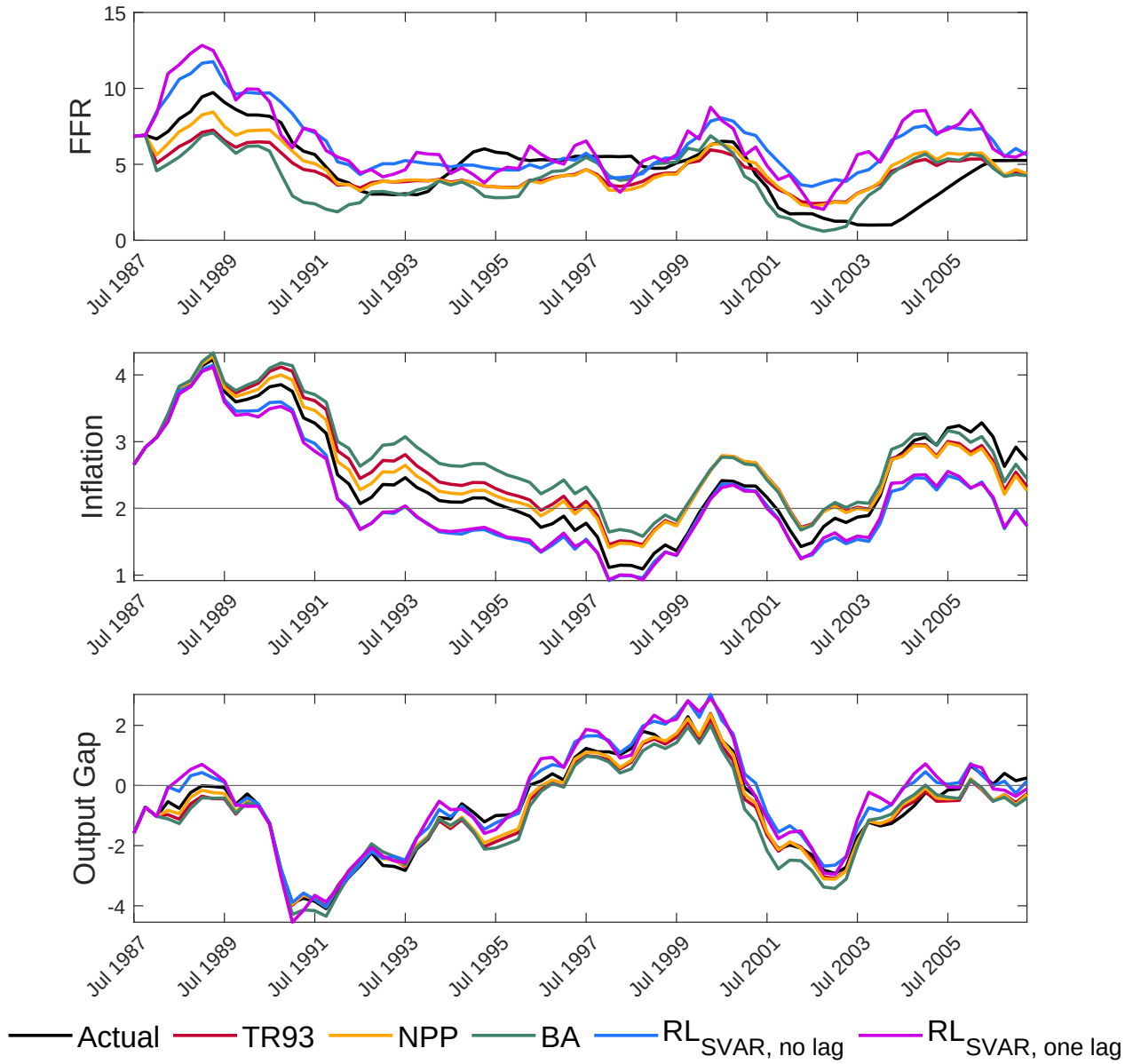
First, it is remarkable that the interest rate level (top panel) with both optimized rules is on average larger than the actual federal funds rate and interest rate prescriptions by other rules. This, presumably, is caused by the combination of a relatively large constant term and larger coefficients on the input variables (see Table 4). Further, $RL_{SVAR,\,no\,lag}$ yields a smoother interest rate series than $RL_{SVAR,\,one\,lag}$, which might be explained by coefficients with different signs on the output gap (β_y^0 and β_y^1) of the latter. At the beginning of the sample, TR93, NPP and BA produce interest rates slightly below, while our rules lie 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 dot com bubble crisis causes the interest rate to drop and produces similar reductions across all rules. 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 claimed for example by Taylor (2007).

What is actually more of interest to us are the counterfactual inflation and output gap series, because by these we can evaluate the performance of the RL reaction function. Concerning inflation, 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 actual Fed behavior produced better inflation

[29] Alternatively, one could also state that we implicitly assume that private agents adapt their expectations much more slowly than the central bank.

[30] Please note that SVAR economy in the following denotes the recursively estimated system given in (6)- (7).

Figure 10: 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).

values. We find similar results for the output gap. The values induced by the optimized policies lie above the common rules and the data. In most of the times, this yields values closer to the target of zero.

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 (16) multiplied by (-1). Table 5 summarizes the results.

Table 5: Actual and Counterfactual Target Deviation and Loss (SVAR Economy)

Policy	$\Delta^2(\pi^*, \pi_t)$	$\Delta^2(y^*, y_t)$	Loss
Actual	0.84	2.98	1.91
TR93	0.91	3.01	1.96
NPP	0.83	3.03	1.93
BA	1.04	3.37	2.21
$RL_{SVAR,\,no\,lag}$	0.66	**2.95**	**1.80**
$RL_{SVAR,\,one\,lag}$	**0.63**	3.11	1.87

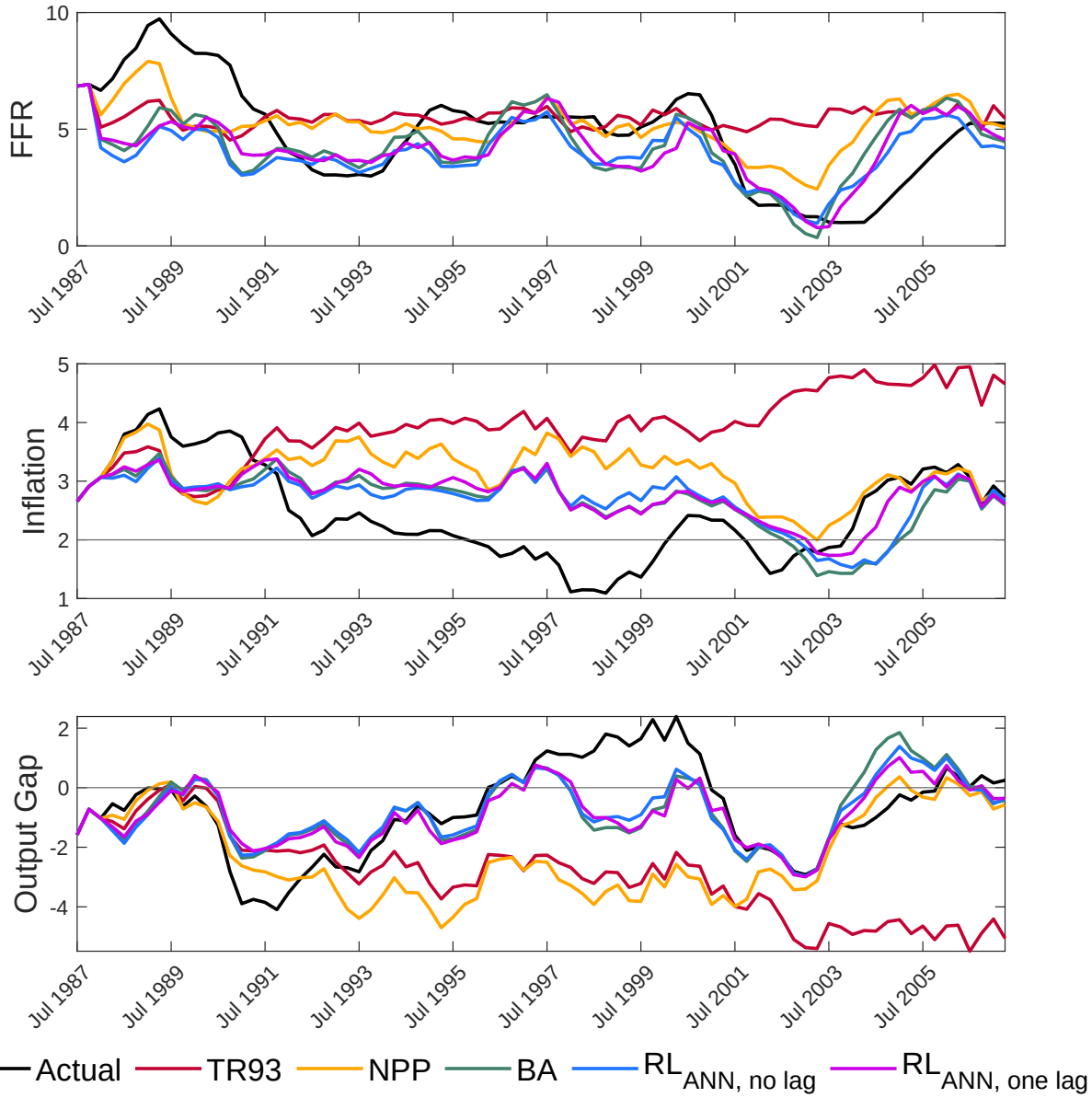
Note: Δ^2 denotes the mean squared deviation of the respective variable from its target value ($\pi^* = 2$ and $y^* = 0$). The loss is calculated averaging over both: $Loss = 0.5 \cdot \Delta^2(\pi^*, \pi_t) + 0.5 \cdot \Delta^2(y^*, y_t)$.

We find that the optimized RL rules yields 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. Concerning 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.

ANN Economy & Linear Policy For the same dynamic counterfactual analysis based on the nonlinear ANN economy, the results are summarized in Figure 11.

Focusing on the common policy rules, we find nominal interest rate series differing significantly from those of the SVAR economy (Fig. 10). The interest rate path prescribed by TR93 is rather constant, which leads to (and is a result of) diverging inflation and output gap values in opposite directions. Comparing the different paths under TR93 and BA, and having the

Figure 11: 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).

nonlinear economy structures in mind, helps to explain this rather counterintuitive result. Both rules only differ in the response to the output gap, with BA having a twice as large coefficient compared to TR93 (cf. Table 4). The interest rates paths start to differ in 1990, where inflation amounts to around 3 and the output gap is at -2 under both rules. The BA therefore prescribes a lower value than the TR93 due to the larger output gap coefficient. Figures 17 and 18 in the Appendix try to illustrate the following dynamics of inflation and output gap, respectively, induced by the different interest rate values. Due to the nonlinearities present in the estimated inflation equation (cf. Figure 6), inflation increases with larger interest rates when the output gap is negative. In the following period, the output gap decreases due to larger inflation above target (cf. Figure 7). Taken together, the larger interest rate under TR93 leads to a jump on an upwards path of inflation in Figure 17 and a declining negative output gap (Fig. 18). Instead, under the lower interest rate prescribed by BA, inflation moves to a more stable plateau, since also the output gap closes over time.

The inflation tilting rule (NPP) interest rate dynamics look a bit more similar to the actual FFR data, especially during the peak in 1989 and the trough in 2003. Compared to TR93, it produces an improved inflation performance, but still ranks below the other rules. Output gap values are fine between 1987 and 1990 and also after 2003, but in-between NPP performs even worse than TR93. Overall, the results of TR93 and NPP indicate that their output gap coefficients are too small, leading to inferior dynamics of both output gap and inflation within the ANN economy.[31]

Contrary, BA seems to be more robust with respect to different economies. It prescribes an interest rate pattern very similar to both policy rules optimized in the ANN economy ($RL_{ANN,\,no\,lag}$ and $RL_{ANN,\,one\,lag}$). Before 1991, all three policies show lower interest rates, which stem from lower inflation values during this time. Between 1991 and 2003, interest rate paths are pretty close, before BA, $RL_{ANN,\,no\,lag}$ and $RL_{ANN,\,one\,lag}$ precede and exceed the data after 2003. In terms of inflation, all three policies closer to the target until 1991, the opposite happening between 1991-1997 and 1999-2003. After 2003, our policies produce a similar inflation series shape, but our policies lag behind the actual data, falling closer to the target. The induced output gap time series of BA and our policies follow a similar pattern, as well. While it lies mostly above actual until 1994, it falls below until 2003 and exceeds the

[31] Since the results of the counterfactual exercise also depend on the time span and the particular starting point, we also evaluated the different rules using different subsets. We find that the diverging behaviour under TR93 is less pronounced when we start after 1990. However, it still leads to inferior results and the overall ranking of the policy rules stays unchanged. The respective detailed results are available upon request.

data again thereafter, implying less deviations from target. There is a large difference between the actual output gap and the implied one by the rules between 1997 and 2000, where actual output gap rises, while the counterfactuals decrease.

Table 6: Actual and Counterfactual Target Deviation and Loss (ANN Economy)

Policy	$\Delta^2(\pi^*, \pi_t)$	$\Delta^2(y^*, y_t)$	Loss
Actual	0.84	2.98	1.91
TR93	4.07	10.66	7.36
NPP	1.50	7.33	4.41
BA	0.69	1.89	1.29
$RL_{ANN,\,no\,lag}$	**0.68**	1.69	1.18
$RL_{ANN,\,one\,lag}$	0.75	1.74	1.24
$RL_{ANN,\,no\,lag,\,nonlin}$	0.74	**1.35**	**1.04**
$RL_{ANN,\,one\,lag,\,nonlin}$	0.81	**1.35**	1.08

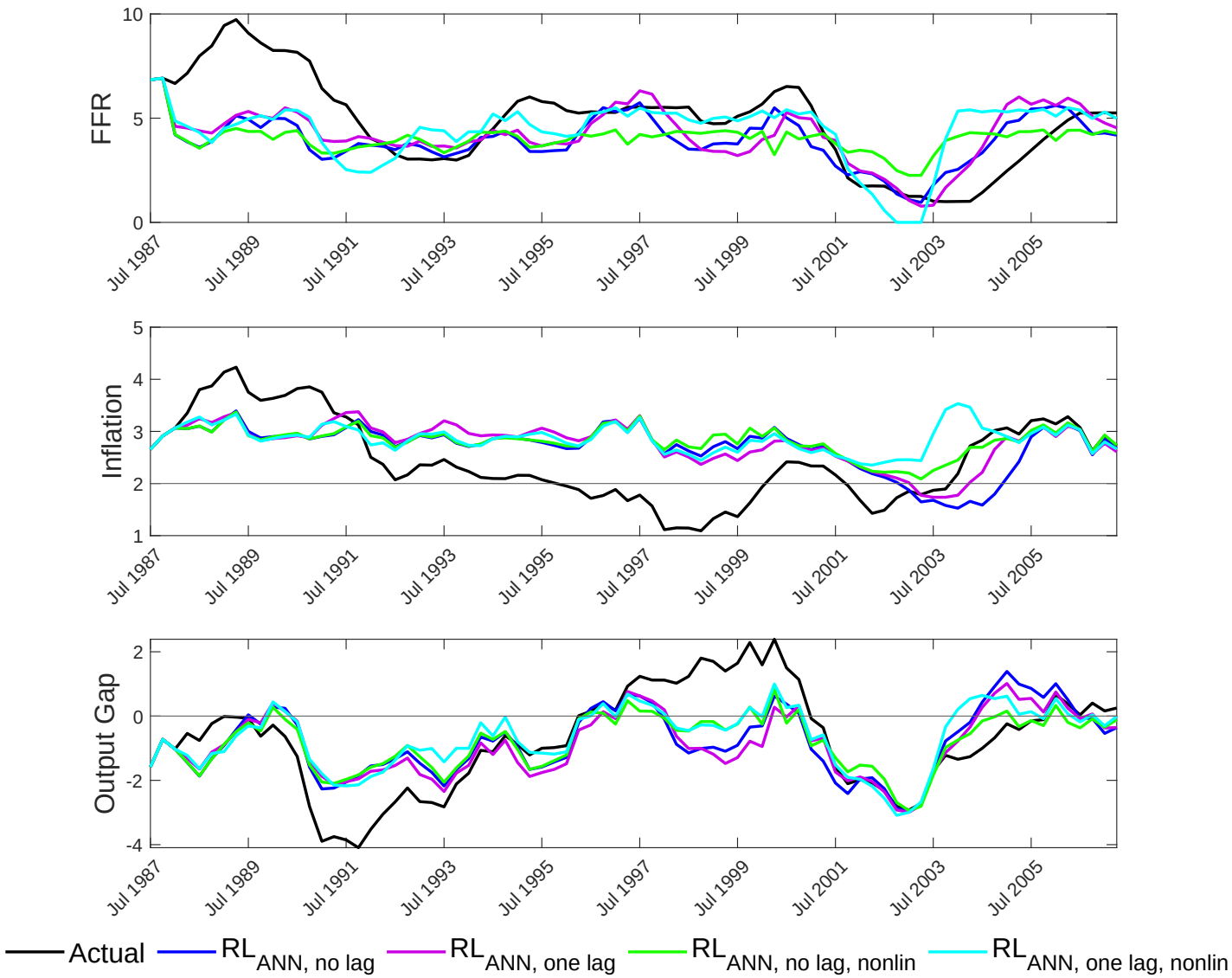
Note: Δ^2 denotes the mean squared deviation of the respective variable from its target value ($\pi^* = 2$ and $y^* = 0$). The loss is calculated averaging over both: $Loss = 0.5 \cdot \Delta^2(\pi^*, \pi_t) + 0.5 \cdot \Delta^2(y^*, y_t)$.

Adding numbers to the descriptive analysis, Table 6 shows that $RL_{ANN,\,no\,lag}$ ($RL_{ANN,\,one\,lag}$) reduces the squared deviations of inflation from its target by 19 % (10.7 %) compared to actual values. Even more remarkable is the reduction of output gap deviations from its target. Sticking to the linear optimized rule $RL_{ANN,\,one\,lag}$, the mean squared deviation amounts to 1.74, a 41.6 % relative to actual data. With $RL_{ANN,\,no\,lag}$, this outstanding result can even be improved to 43.3 %. This transfers to the loss, where the RL optimized policy with only two inputs yields the best result among the linear rules, shortly followed by the 4-inputs RL policy. Among the common policy rules, BA shows by far the best results, also outperforming the data but still ranking behind our RL rules.

ANN Economy & Nonlinear Policy In Figure 12, we contrast the differences between *linear* and *nonlinear* RL optimized policies within the ANN economy, conducting the same dynamic counterfactual.

First of all, $RL_{ANN,\,one\,lag,\,nonlin}$ yields the largest fluctuations in the interest rate, even hitting the ZLB between 2002 and 2003. The counterfactual inflation paths look quite similar across rules, except for the period between 2003 and 2005. While inflation under the nonlinear RL rules lies above the actual series, the linear RL yields values lower and closer to the target during that time. However, the nonlinear rules yield counterfactual output gap series closer

Figure 12: 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).

to the target in contrast to their linear counterparts and the actual series during the mid and end of the 1990s and at the very end of the sample. Table 6 summarizes the results. We see that both nonlinear RL rules provide slightly worse inflation stabilization than our optimized linear policy $RL_{ANN,\,no\,lag}$ and the BA. Contrary, the nonlinear rules are much better in output gap stabilization. Here, both reach a value of 1.35 which corresponds to an improvement with respect to the data of 54.7 %. Overall, $RL_{ANN,\,no\,lag,\,nonlin}$ wins the race with a loss of 1.04, which is equivalent to a 45.5 % improvement with respect to the loss implied by the actual data ($RL_{ANN,\,one\,lag,\,nonlin}$ produces a loss reduced by 43.5 %). Compared to the best performing optimized linear policy $RL_{ANN,\,no\,lag}$, the best nonlinear policy lowers the loss by further 11 %.

3.4 Model Comparison

Optimal monetary policy is always related to an environment, which is close to real-world interrelations in the best case. So far, we provided a linear SVAR environment and an ANN environment, with the latter producing the superior data fit. Hence, we postulate the optimized policy rules based on the ANN economy to be the more 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.[32] The DSGE framework also allows for a true counterfactual analysis as the Lucas (1976) critique does not apply here.

3.4.1 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 and post 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

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

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 inflation target.[33] 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.[34]

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

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

3.4.2 The Policy Rules

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

$$\widehat{i}_t = \beta_\pi^0 \widehat{\pi}_t + \beta_\pi^1 \widehat{\pi}_{t-1} + \beta_y^0 \widehat{y}_t + \beta_y^1 \widehat{y}_{t-1}. \tag{17}$$

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 consider 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.[35] The respective coefficients for (17) are given in Table 4.

Optimal Simple Rules In addition to the aforementioned policy rules, we also calculate optimal simple rules (OSR) for each model used in the comparison.[36] 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):

$$\widehat{i}_t = \varphi_\pi^0 \widehat{\pi}_t + \varphi_y^0 \widehat{y}_t \tag{18}$$

$$\widehat{i}_t = \varphi_\pi^0 \widehat{\pi}_t + \varphi_\pi^1 \widehat{\pi}_{t-1} + \varphi_y^0 \widehat{y}_t + \varphi_y^1 \widehat{y}_{t-1}. \tag{19}$$

By solving

$$\min_{\varphi} Var(\widehat{\pi}_t) + Var(\widehat{y}_t) + Var(\Delta\widehat{i}_t) \tag{20}$$

subject to (18) or (19), we find the optimal response coefficients φ for each model.[37] The resulting coefficients are given in Tables 10 and 11 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

[35] 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.

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

[37] 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.

models. Comparing these to the common rules, the weight on the output gap is quite large in $OSR_{no\,lag}$ (corresponding to (18)) with 1.99, while the inflation coefficient is comparable to the NPP rule. Also $OSR_{one\,lag}$ (referring to (19)) has pretty large weights, which are closest to $RL_{ANN,\,one\,lag}$.

3.4.3 Performance across Models

Following the principle of model comparison, common variables are consistently defined and related to model-specific variables. The annualized quarterly interest rate is denoted by i_t, π_t corresponds to the year-on-year inflation rate and y_t is equivalent to the quarterly output gap, defined as the deviation of actual output from the level of output that would have been realized under flexible prices. Further, the models' specific monetary policy rules are exchanged by the rules mentioned before, one at a time. We assume the absence of monetary policy shocks. By computing the stationary rational expectations solution of each model, we get the unconditional distribution of the endogenous variables and hence also the unconditional second moments (see also e.g. Levin et al. (2003) and Taylor and Wieland (2012) following the same approach). The result is driven by the model parameters, the covariance matrix of the structural shocks of that model and, most importantly, the policy rule. The size of model-specific shocks has a significant impact on the unconditional variance. Differences stemming from this fact are not of interest to us. Hence, following Cochrane et al. (2019), we compare the relative performance with respect to TR93.[38] The smaller the unconditional variances, the more stability does the respective policy rule provide.

A robust policy rule is supposed to perform well across all models, which are considered relevant for policy evaluation. Figure 22 in the Appendix shows the relative (to TR93) unconditional variances of i_t (FFR), π_t (Infl) and y_t (GDP) for each model in detail. Averaging over the results per model enables us to evaluate the overall performance. Table 7 summarizes the results. On the one hand, we take an average over all models. On the other hand, we distinguish between models created before the financial crisis and models established after 2008. We do so, as post-crisis models seem to be more realistic as they incorporate financial frictions, more details in the banking sector, housing markets and other aspects which gained importance through the financial crisis.

From a central banking perspective, the joint unconditional variances reflect the loss. Thus,

[38] An unconditional variance value larger than one indicates worse performance than TR93, while a value smaller than one means better performance.

we also average over $Var(\widehat{i}_t)$, $Var(\widehat{\pi}_t)$ and $Var(\widehat{y}_t)$, which constitutes our performance measure including interest rate variations ($L_{\pi,y,i}$). Additionally, we compute $L_{\pi,y}$, which reflects the average over $Var(\widehat{\pi}_t)$ and $Var(\widehat{y}_t)$ only. Since the reward function in our RL algorithm also focuses on inflation and output gap deviations, this measure is closer to the RL objective.

The first two rows in Table 7 show the results for the optimal simple rules.[39] As expected, both versions perform better than TR93, where $OSR_{one\,lag}$ yields a smaller loss than the simple 2-inputs version $OSR_{no\,lag}$. Both, the unconditional variance of output gap and inflation are reduced substantially by adding lags to the policy. As expected, the inflation tilting rule (NPP) achieves better results than TR93 with respect to $Var(\widehat{\pi}_t)$. Contrary, BA tilts towards output gap stabilization, with a twice as large response coefficient on y_t compared to TR93 (see Table 4). This translates into a relative unconditional output gap variance of 0.59. Considering now the RL optimized policy versions with two inputs, we find that $RL_{SVAR,\,no\,lag}$ yields even more inflation stability than NPP. $RL_{ANN,\,no\,lag}$ performs worst concerning inflation stabilization, which can be explained by an inflation coefficient that is even smaller than the one of BA. Its weight on the output gap is larger than in the reference policy, producing the second best relative output gap variance of 0.68. Analyzing the RL optimized 4-input policies, we find that $RL_{SVAR\,one\,lag}$ yields the lowest inflation variance, which is due to the large cumulative inflation coefficient of about three. Also the output gap coefficients are in sum larger than with TR93, but the negative coefficient on the lagged input seems to worsen output gap stability relative to TR93. $RL_{ANN,\,one\,lag}$ with cumulative inflation and output gap coefficients of 1.71 and 0.94 yields relative variances of 0.86 and 0.76, respectively. Besides the OSR that are not feasible in practice, it is the only rule that beats TR93 in both, inflation and output gap stabilization. Considering all models and focusing on the variances of π_t and y_t, $RL_{SVAR\,one\,lag}$ produces the loss closest to the optimal simple rules with $L_{\pi,y} = 0.79$. Taking interest rate variations into account, $RL_{ANN,\,one\,lag}$ performs best. It also yields the second best performance with respect to $L_{\pi,y}$. Due to the increased complexity of the post-crisis models, we assume that results might differ when we look at the two subgroups of models. Within the pre-crisis models, $RL_{SVAR\,one\,lag}$ still constitutes the best stabilizer for inflation and output gap. However, the NPP rule achieves very similar results and outperforms our rule taking also interest rate volatility into account. Focusing on post-crisis models, $RL_{ANN,\,one\,lag}$ shows the best stabilizing properties with remarkable relative losses of $L_{\pi,y} = 0.74$ and $L_{\pi,y,i} = 0.76$.

[39] Note, that $OSR_{no\,lag}$ $OSR_{one\,lag}$ comprise different rules each since each model has its own optimal simple rule.

Table 7: 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
$RL_{SVAR,\,no\,lag}$	0.95	0.43	1.34	0.91	0.88	0.97	1.01	0.88	0.84
$RL_{SVAR,\,one\,lag}$	1.32	0.39	1.20	0.97	**0.79**	1.12	**0.91**	0.91	0.75
$RL_{ANN,\,no\,lag}$	1.38	1.54	0.68	1.20	1.11	1.32	1.19	1.16	1.08
$RL_{ANN,\,one\,lag}$	0.94	0.86	0.76	**0.85**	0.81	1.10	1.00	**0.76**	**0.74**

Note: We calculate the unconditional variances for the nominal interest rate, inflation and 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.

3.5 Discussion

Taking one step back, we would like 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. It allows to capture nonlinearities while being agnostic about the specific functional form, bringing the model closer to the data compared to a standard SVAR, that 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 of course 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 want to criticize the Fed's monetary policy. Rather, our aim is to provide a proof of work of the RL algorithm and to contrast the policy rules in varying settings.

Relying on our estimated linear SVAR economy, reinforcement learning finds policies performing well in such a world. They outperform common rules and the actual Fed behavior, which also

produce solid results (Table 5). Applying the common rules in the ANN economy, however, produces inferior loss values. Only the BA seems to be able to stabilize both economies quite well (Table 6). Nevertheless, BA performs worse compared to both of our optimized rules $RL_{ANN,\,no\,lag}$ and $RL_{ANN,\,one\,lag}$.

In our DSGE model comparison exercise, which is conducted in linear (or linearized) economic models, most rules perform quite well. The balanced approach (BA), which seems to be robust over our linear and ANN economy does (on average) not transfer this stability to the DSGE model comparison. It performs only slightly better than TR93 but is inferior to most other rules. While in the historical counterfactual generally *no lag* versions achieve the better results, including lagged values of inflation and output gap seems to help reducing undesired variability in the DSGE model comparison. Given the assumed loss function with equal weights on inflation and output gap stabilization, our optimized policy $RL_{ANN,\,one\,lag}$ withstands best the model variations. Concerning the worse behavior of TR93 and NPP in our ANN economy, it seems as if these common policies perform fine in linear economy structures, while they rather fail when nonlinearities are present. Having the remarkable data fit of our ANN economy in mind, we would suggest also considering rules like the $RL_{ANN,\,onelag}$ that perform well in such a nonlinear environment.

Concerning linear versus nonlinear policy rules, the loss resulting from the historical counterfactual suggests that nonlinear optimized policies are even better economic stabilizers (Table 6). We do not challenge this result within a DSGE model comparison exercise due to the increased complexity. The robustness analysis of nonlinear policies' in the form of neural networks with respect to model uncertainty is left for future research. Still, the results from the dynamic counterfactual are promising and central banks could consider nonlinear policies in the form of ANNs, as well. We have to admit, however, that these policies are more difficult to communicate due to their more complex structure. Since credibility is essential for a central bank, certainly a trade-off between policy improvement and transparency exists with these kind of rules. We show how partial dependence plots can address this black-box critique.

4 Conclusion

This paper provides a new machine learning based approach to finding optimal monetary policy reaction functions given preferences of a central bank. In the first step, we show how ANNs can

be used as a modeling device for transition equations, capturing nonlinear interdependencies and thereby bringing the setup closer to the data compared to a standard SVAR. Then, we apply reinforcement learning, which is a computational approach of goal-directed learning from interaction and optimal control.

Reinforcement learning is flexible enough to account for different kinds of nonlinear constraints like the ZLB, convex Phillips/ IS curves and asymmetric preferences. Further, it does not require perfect knowledge of the environment and can therefore mitigate the problem of model-uncertainty. However, learning requires experimenting with the economy's reaction to policy actions.

Relying on a nonlinear ANN representation of the economy, our optimized linear rules are shown to reduce the central bank's loss by 35 % and the nonlinear ones by over 43 % compared to values implied by actual data. The results of our DSGE model comparison, where we evaluate the linear optimized rules across several models, further indicate the approach's robustness with respect to model uncertainty. Robustness checks concerning model uncertainty of the optimized nonlinear rules are left for future research.

Revising a central bank's monetary policy strategy certainly involves many more aspects than we touch upon in this paper. The Fed recently announced that it will switch to average inflation targeting and review its strategy roughly every 5 years. Future research could therefore consider loss functions that incorporate an average inflation target. One could further think of combining the advantages of economic modeling and RL by deviating from rational expectations and combining a learning central bank with learning private agents within a DSGE model. While we focus on a reaction function for the nominal interest rate in this paper, one might also consider reaction functions for unconventional monetary policy measures like asset purchases. The most obvious step for future research is to add more variables to broaden the representation of the economy as well as the controlled variables of the central bank. Compared to standard DP 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 suggest adding reinforcement learning to the central bankers' toolkit for determining optimal monetary policy reaction functions. When implementing our method, one should update the algorithm regularly. This means, that the neural network representing the economy is estimated with the latest data, which incorporates possible modified social behaviour or structural changes.

References

Adam, K. and Billi, R. M. (2006). Optimal Monetary Policy under Commitment with a Zero Bound on Nominal Interest Rates. *Journal of Money, Credit and Banking*, 83(7):1877–1905.

Aras, S. and Kocakoç, İ. D. (2016). A New Model Selection Strategy in Time Series Forecasting with Artificial Neural Networks: IHTS. *Neurocomputing*, 174:974–987.

Ball, L. and Mazumder, S. (2019). A Phillips Curve with Anchored Expectations and Short-Term Unemployment. *Journal of Money, Credit and Banking*, 51(1):111–137.

Bellman, R. (1957a). A Markovian Decision Process. *Journal of Mathematics and Mechanics*, 6(5):679–684.

Bellman, R. (1957b). *Dynamic Programming*. Princeton University Press.

Bernanke, B. S., Gertler, M., and Gilchrist, S. (1999). The Financial Accelerator in a Quantitative Business Cycle Framework. *Handbook of Macroeconomics*, 1:1341–1393.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-dynamic Programming*. Athena Scientific.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, 23(5):408–422.

Carabenciov, I., Freedman, M. C., Garcia-Saltos, M. R., Laxton, M. D., Kamenik, M. O., and Manchev, M. P. (2013). GPM6:The Global Projection Model with 6 Regions. IMF Working Paper 13-87, International Monetary Fund.

Castro, P. S., Desai, A., Du, H., Garratt, R., and Rivadeneyra, F. (2021). Estimating policy functions in payment systems using reinforcement learning. Staff Working Paper 2021-7, Bank of Canada.

Charpentier, A., Elie, R., and Remlinger, C. (2021). Reinforcement learning in economics and finance. *Computational Economics*, pages 1–38.

Chen, M., Joseph, A., Kumhof, M., Pan, X., Shi, R., and Zhou, X. (2021). Deep Reinforcement Learning in a Monetary Model. *arXiv preprint arXiv:2104.09368*.

Christiano, L. J., Motto, R., and Rostagno, M. (2014). Risk Shocks. *American Economic Review*, 104(1):27–65.

Cochrane, J. H., Taylor, J. B., and Wieland, V. (2019). Evaluating Rules in the Feds Report and Measuring Discretion. In *Hoover Institution, Strategies for Monetary Policy: A Policy Conference*.

Cogan, J. F., Cwik, T., Taylor, J. B., and Wieland, V. (2010). New Keynesian versus Old Keynesian Government Spending Multipliers. *Journal of Economic Dynamics and Control*, 34(3):281–295.

Coibion, O. and Gorodnichenko, Y. (2015). Is the Phillips Curve Alive and Well After All? Inflation Expectations and the Missing Disinflation. *American Economic Journal: Macroeconomics*, 7(1):197–232.

Cúrdia, V. and Woodford, M. (2009). Credit Frictions and Optimal Monetary Policy. BIS Working Paper 278, JBIS.

Debortoli, D., Kim, J., Lindé, J., and Nunes, R. (2019). Designing a Simple Loss Function for Central Banks: Does a Dual Mandate Make Sense? *The Economic Journal*, 129(621):2010–2038.

Del Negro, M., Giannoni, M. P., and Schorfheide, F. (2015). Inflation in the Great Recession and New Keynesian Models. *American Economic Journal: Macroeconomics*, 7(1):168–96.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 74(366a):427–431.

Dolado, J. J., María-Dolores, R., and Naveira, M. (2005). Are Monetary-Policy Reaction Functions Asymmetric?: The Role of Nonlinearity in the Phillips Curve. *European Economic Review*, 49(2):485–503.

Dolado, J. J., Pedrero, R. M.-D., and Ruge-Murcia, F. J. (2004). Nonlinear Monetary Policy Rules: Some New Evidence for the US. *Studies in Nonlinear Dynamics & Econometrics*, 8(3).

Fernández-Villaverde, J., Guerrón-Quintana, P., Kuester, K., and Rubio-Ramírez, J. (2015). Fiscal Volatility Shocks and Economic Activity. *American Economic Review*, 105(11):3352–84.

Gertler, M. and Karadi, P. (2011). A model of Unconventional Monetary Policy. *Journal of Monetary Economics*, 58(1):17–34.

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

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

Hansen, L. and Sargent, T. J. (2001). Robust Control and Model Uncertainty. *American Economic Review*, 91(2):60–66.

Hawkins, R. J., Speakes, J. K., and Hamilton, D. E. (2015). Monetary Policy and PID Control. *Journal of Economic Interaction and Coordination*, 10(1):183–197.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes.* MIT Press.

Iacoviello, M. and Neri, S. (2010). Housing Market Spillovers: Evidence from an Estimated DSGE Model. *American Economic Journal: Macroeconomics*, 2(2):125–64.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root. *Journal of Econometrics*, 54:159–178.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer.

Levenberg, K. (1944). A Method for the Solution of Certain Non-linear Problems in Least Squares. *Quarterly of Applied Mathematics*, 2(2):164–168.

Levin, A., Wieland, V., and Williams, J. C. (2003). The Performance of Forecast-based Monetary Policy Rules under Model Uncertainty. *American Economic Review*, 93(3):622–645.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N. M. O., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. (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*, pages 19–46.

Marquardt, D. W. (1963). An Algorithm for Least-squares Estimation of Nonlinear Parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with Deep Reinforcement Learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533.

Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 807–814.

Nikolsko-Rzhevskyy, A., Prodan, R., and Papell, D. H. (2014). Deviations from Rules-Based Policy and Their Effects. *Journal of Economic Dynamics and Control*, 49:4–17.

Nikolsko-Rzhevskyy, A., Prodan, R., and Papell, D. H. (2018). Policy Rules and Economic Performance. Working paper.

Orphanides, A. (2001). Monetary Policy Rules Based on Real-Time Data. *American Economic Review*, 91(4):964–985.

Orphanides, A. (2003). Monetary Policy Evaluation with Noisy Information. *Journal of Monetary economics*, 50(3):605–631.

Orphanides, A. and Wieland, V. (2000). Inflation Zone Targeting. *European Economic Review*, 44(7):1351–1387.

Primiceri, G. E. (2005). Time Varying Structural Vector Autoregressions and Monetary Policy. *The Review of Economic Studies*, 72(3):821–852.

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

Rudebusch, G. and Svensson, L. E. (1999). Policy Rules for Inflation Targeting. In *Monetary Policy Rules*, pages 203–262. University of Chicago Press.

Schaling, E. (2004). The Nonlinear Phillips Curve and Inflation Forecast Targeting: Symmetric versus Asymmetric Monetary Policy Rules. *Journal of Money, Credit and Banking*, 36(3):361–386.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. (2014). Deterministic Policy Gradient Algorithms. In *International Conference on Machine Learning*.

Sims, C. A. and Zha, T. (2006). Were there Regime Switches in US Monetary Policy? *American Economic Review*, 96(1):54–81.

Smets, F. and Wouters, R. (2007). Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach. *American Economic Review*, 97(3):586–606.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

Svensson, L. (1997). Inflation Forecast Targeting: Implementing and Monitoring Inflation Targets. *European Economic Review*, 41(6):1111–1146.

Svensson, L. E. (2020). Monetary Policy Strategies for the Federal Reserve. *International Journal of Central Banking*, 16(1):133–193.

Tambakis, D. N. (2009). Optimal Monetary Policy with a Convex Phillips Curve. *The BE Journal of Macroeconomics*, 9(1).

Taylor, J. B. (1993). Discretion versus Policy Rules in Practice. In *Carnegie-Rochester Conference Series on Public Policy*, volume 39, pages 195–214.

Taylor, J. B. (2007). The Explanatory Power of Monetary Policy Rules. *Business Economics*, 42(4):8–15.

Taylor, J. B. and Wieland, V. (2012). Surprising Comparative Properties of Monetary Models: Results from a New Model Database. *Review of Economics and Statistics*, 94(3):800–816.

Taylor, J. B. and Williams, J. C. (2010). Simple and robust rules for monetary policy. In Friedman, B. M. and Woodford, M., editors, *Handbook of Monetary Economics*, volume 3, pages 829–859. Elsevier.

Tetlow, R. J. and Von zur Muehlen, P. (2001). Robust Monetary Policy with Misspecified Models: Does Model Uncertainty Always Call for Attenuated Policy? *Journal of Economic Dynamics and Control*, 25(6-7):911–949.

Watson, M. W. (2014). Inflation Persistence, the NAIRU, and the Great Recession. *American Economic Review*, 104(5):31–36.

Wieland, V. (2000). Monetary Policy, Parameter Uncertainty and Optimal Learning. *Journal of Monetary Economics*, 46(1):199–228.

Wieland, V., Afanasyeva, E., Kuete, M., and Yoo, J. (2016). New Methods for Macro-financial Model Comparison and Policy Analysis. In *Handbook of Macroeconomics*, volume 2, pages 1241–1319. Elsevier.

Wieland, V., Cwik, T., Müller, G. J., Schmidt, S., and Wolters, M. (2012). A New Comparative Approach to Macroeconomic Modeling and Policy Analysis. *Journal of Economic Behavior & Organization*, 83(3):523–541.

Woodford, M. (2001). The Taylor Rule and Optimal Monetary Policy. *American Economic Review*, 91(2):232–237.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D. C., and Socher, R. (2020). The AI Economist: Improving Equality and Productivity with AI-Driven Tax Policies.

C Appendix C

C.1 DDPG Algorithm

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 $j = 1, \ldots, q$ and $i = 0, \ldots, q$ are initialized. Among them, the parameters $[\beta_j, \delta_j]$ are randomly initialized using the *He*-approach, which is the recommended when working with ReLu layers as shown in He et al. (2015).[40] The biases α_i are initialized to zero (default). The critic parameters θ^Q are also randomly initialized using the *glorot* function of Glorot and Bengio (2010).[41] 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 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 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 20 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 by our economy representations, we also initialize these from the data.

[40] The *He*-initializer samples from a normal distribution with zero mean and variance $2/InputSize$, where $InputSize$ corresponds to the number of variables entering the respective layer of the neural network. Depending on our policy function this equals either 2 or 4.

[41] 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.

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

$$i_t = P(x_t|\theta_t^P) + \mathcal{N}_t. \tag{21}$$

g) The previously chosen action and the state enters the environmental equations, i.e. our linear ((6) and (7)) or ANN economy ((4) and (5)). 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 the one of the estimated shocks.

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 being 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

$$h_j = r(x_j, i_j) + \gamma Q'(x_{j+1}, P'(x_{j+1}|\theta^{P'})|\theta^{Q'}). \tag{22}$$

It is composed of the reward in j plus the discounted future reward, presuming the 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 $\theta^{Q'}$.

$$L = \frac{1}{N}\sum_j (h_j - Q(x_j, i_j|\theta^Q))^2 \tag{23}$$

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 0.0001 (slower than the default value of 0.01).

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:

$$\nabla_{\theta^P} J \approx \frac{1}{N} \sum_j [\nabla_i Q(x, i|\theta^Q)|_{x=x_j, i=P(x_j)} \nabla_{\theta^P} P(i|\theta^P)|_{x=x_j}], \tag{24}$$

where J denotes the expected 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 with 0.0001.

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.

Table 8: 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: 0.0001
			Gradient Threshold: 1
			Optimizer: adam
Actor	Linear Version	imageInputLayer	
		fullyConnectedLayer	Nodes: 1
		reluLayer	
	Non-lin. Version	imageInputLayer	
		fullyConnectedLayer	Nodes: n
		tanhLayer	
		fullyConnectedLayer	Nodes: 1
		reluLayer	
	General		Initializer: He
			Learn Rate: 0.0001
			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 2. 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.

C.2 Further Results

C.2.1 Estimation Results

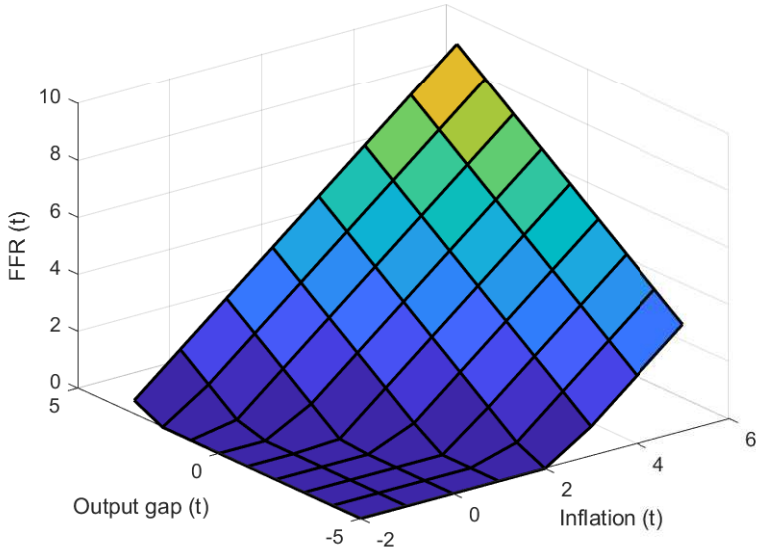
Table 9: Estimation Results of Restricted SVAR

Parameters	Estimates	p-Values
Output gap		
C^y	0.3834	0.0351
$a^y_{y,1}$	0.9084	0.0000
$a^y_{\pi,1}$	-0.1437	0.1409
$a^y_{i,1}$	0.2726	0.0661
$a^y_{i,2}$	-0.2896	0.0313
$\bar{R}^2$	0.9100	
MSE	0.2108	
$\hat{\sigma}^2_{\varepsilon_1}$	0.2136	
DW	1.8206	
$LM(1)$	3.1037	0.1644
Inflation		
C^π	0.1035	0.1659
$a^\pi_{y,0}$	-0.0655	0.1578
$a^\pi_{y,1}$	0.1970	0.0048
$a^\pi_{y,2}$	-0.1121	0.0163
$a^\pi_{\pi,1}$	1.2970	0.0000
$a^\pi_{\pi,2}$	-0.3116	0.0076
$a^\pi_{i,1}$	-0.0122	0.4174
$\bar{R}^2$	0.9450	
MSE	0.0326	
$\hat{\sigma}^2_{\varepsilon_2}$	0.0330	
DW	2.1095	
$LM(1)$	2.5542	0.0696

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

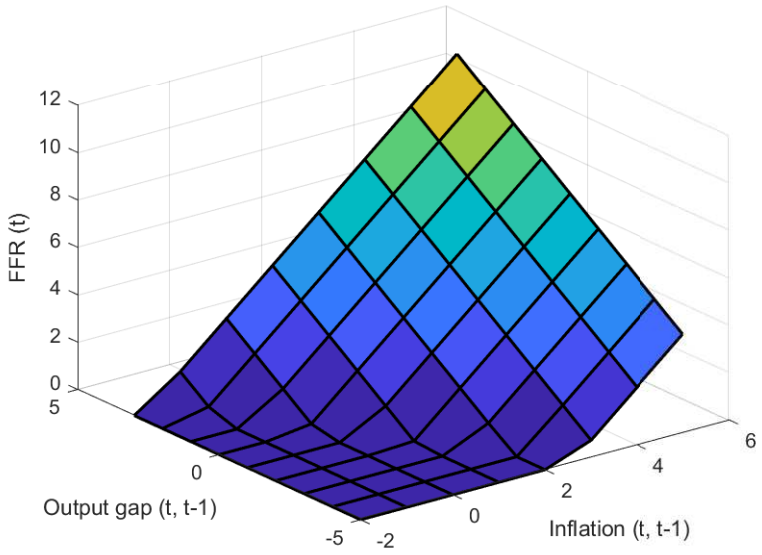
C.2.2 Partial Dependence Plots

Figure 13: Partial Dependence Surface Plot - $RL_{ANN,\,no\,lag}$



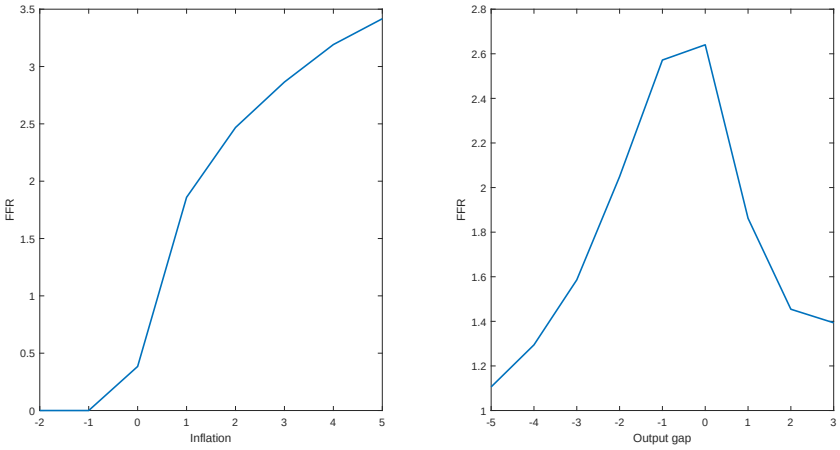
Note: This figure shows the partial dependence of the nominal interest rate, i_t, (FFR) on inflation, π_t, and on the output gap, y_t, under $RL_{ANN,\,no\,lag}$.

Figure 14: Partial Dependence Surface Plot - $RL_{ANN,\,no\,lag}$



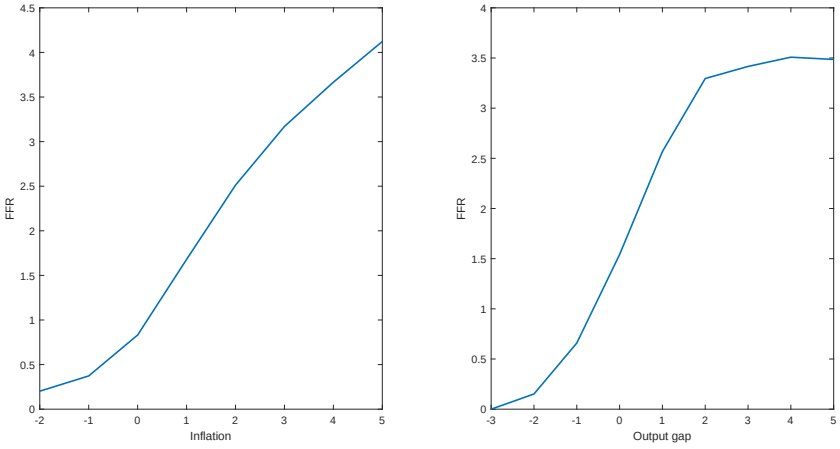
Note: This figure shows the partial dependence of the nominal interest rate, i_t, (FFR) on inflation, π_t, and on the output gap, y_t, assuming $\pi_t = \pi_{t-1}$ and $y_t = y_{t-1}$ under $RL_{ANN,\,one\,lag}$.

Figure 15: Partial Dependence Line Plots - $RL_{ANN,\,no\,lag,\,nonlin}$



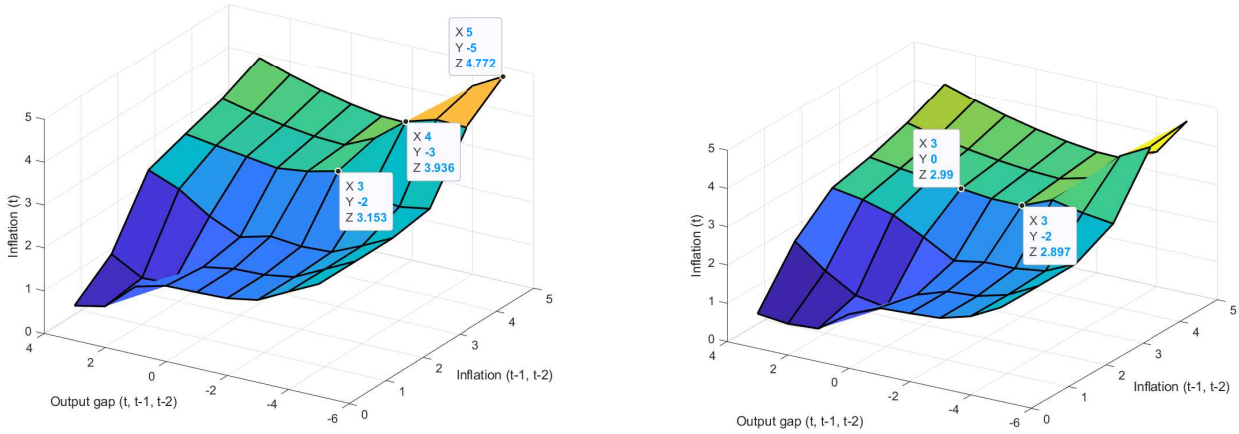
Note: This figure shows the partial dependence of the nominal interest rate, i_t, (FFR) on inflation, π_t, (left) and on the output gap, y_t, (right) under $RL_{ANN,\,no\,lag,\,nonlin}$, marginalizing over the respective remaining variable.

Figure 16: Partial Dependence Line Plots - $RL_{ANN,\,one\,lag,\,nonlin}$



Note: This figure shows the partial dependence of the nominal interest rate, i_t, (FFR) on inflation, $\pi_,t$ (left) and on the output gap, y_t, (right) under $RL_{ANN,\,one\,lag,\,nonlin}$, assuming $\pi_t = \pi_{t-1}$ and $y_t = y_{t-1}$ and marginalizing over the respective remaining variable.

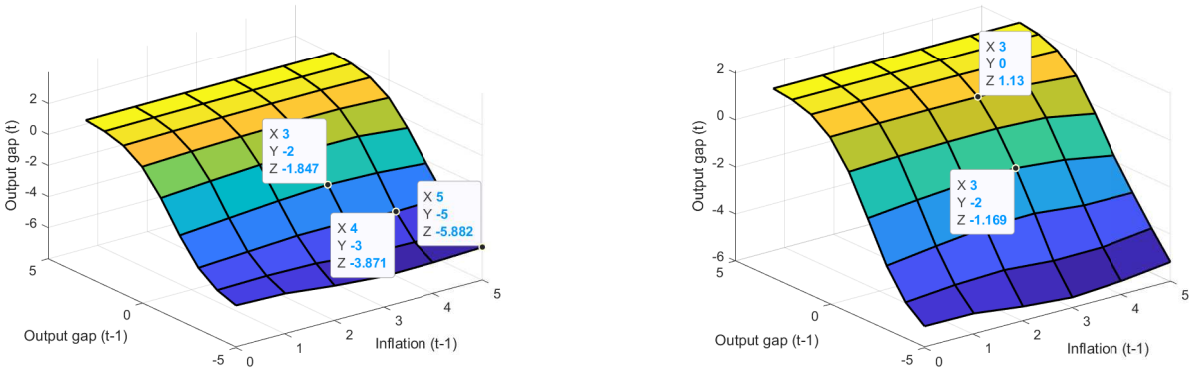
Figure 17: Partial Dependence Plots - Inflation Dynamics under TR93 and BA



(a) at $i_{t-1} = 5$ (TR93) **(b) at $i_{t-1} = 3$ (BA)**

Note: This figure illustrates the inflation dynamics of the historical counterfactual in the ANN economy under TR93 (a) and BA (b). It plots the partial dependence of inflation π_t, on last period's inflation $\pi_{t-1} = \pi_{t-2}$ and the output gap $y_t = y_{t-1} = y_{t-2}$, by assuming constant values across lags and fixing the nominal interest rate, i_{t-1}, at its respective value. The highlighted points represent the dynamics over time.

Figure 18: Partial Dependence Plots - Ouput Gap Dynamics under TR93 and BA



(a) at $i_{t-1} = 5$ (TR93) **(b) at $i_{t-1} = 3$ (BA)**

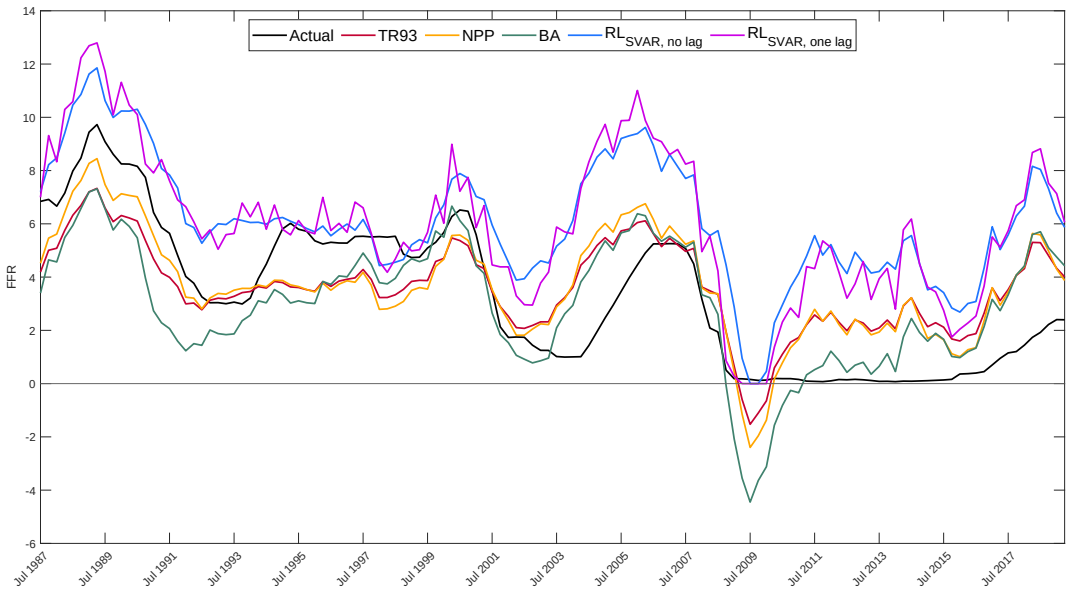
Note: This figure illustrates the output gap dynamics of the historical counterfactual in the ANN economy under TR93 (a) and BA (b). It plots the partial dependence of output gap y_t, on last period's output gap y_{t-1} and inflation π_{t-1}, by fixing the nominal interest rate, i_{t-1}, at its respective value. The highlighted points represent the dynamics over time.

C.2.3 Static Counterfactual

By plugging in the quarterly data on inflation and the output gap into each rule without considering any feedback mechanisms, we can compare the different interest rate prescriptions to the actual behaviour of the Fed.[42] This approach is taken very often to compute measures of discretion (see e.g. Nikolsko-Rzhevskyy et al. (2014, 2018) and Cochrane et al. (2019)). However, it does not provide information on which policy is preferable.

Although only data from 1987 until 2007 was used for the reinforcement learning part, we plot the static counterfactuals until 2019, in order to compare policy prescriptions in crisis periods where the actual nominal interest rate was stuck at the effective lower bound. It can be seen that the optimized rules account for the lower bound by never reaching negative territory. Moreover, all rules prescribe larger interest rates between 2003-2005 compared to the actual path, supporting the too low for too long argument. All rules also show an earlier lift-off from the ZLB after the great financial crisis, where under $RL_{ANN,\,one\,lag,nonlin}$, the ZLB is binding longest until 2014.

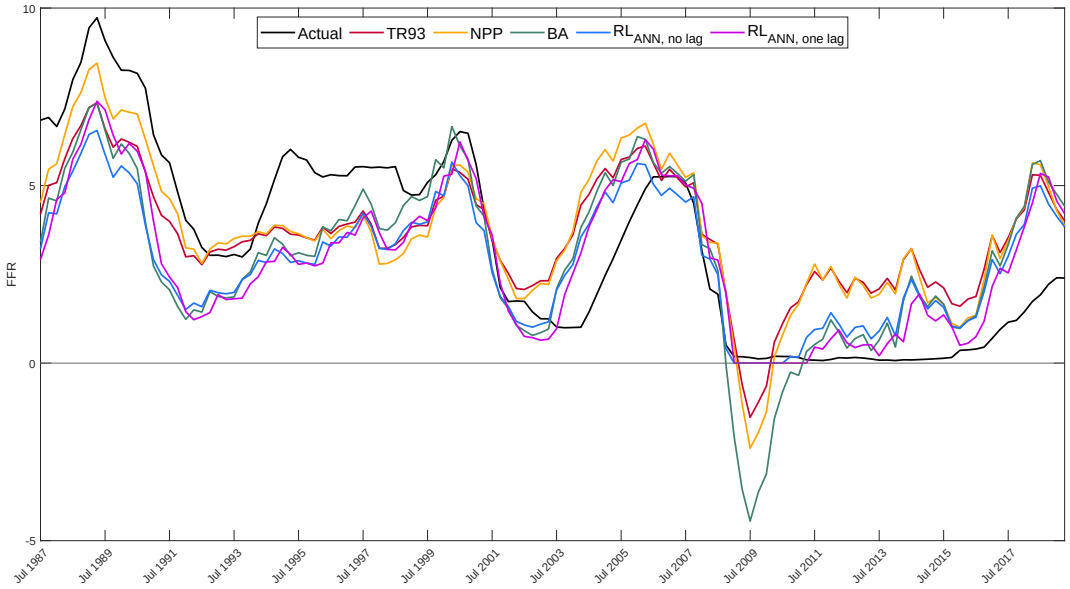
Figure 19: FFR and Prescriptions from Common and RL Rules based on SVAR Economy



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 results of 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 ($RL_{SVAR,\,one\,lag}$).

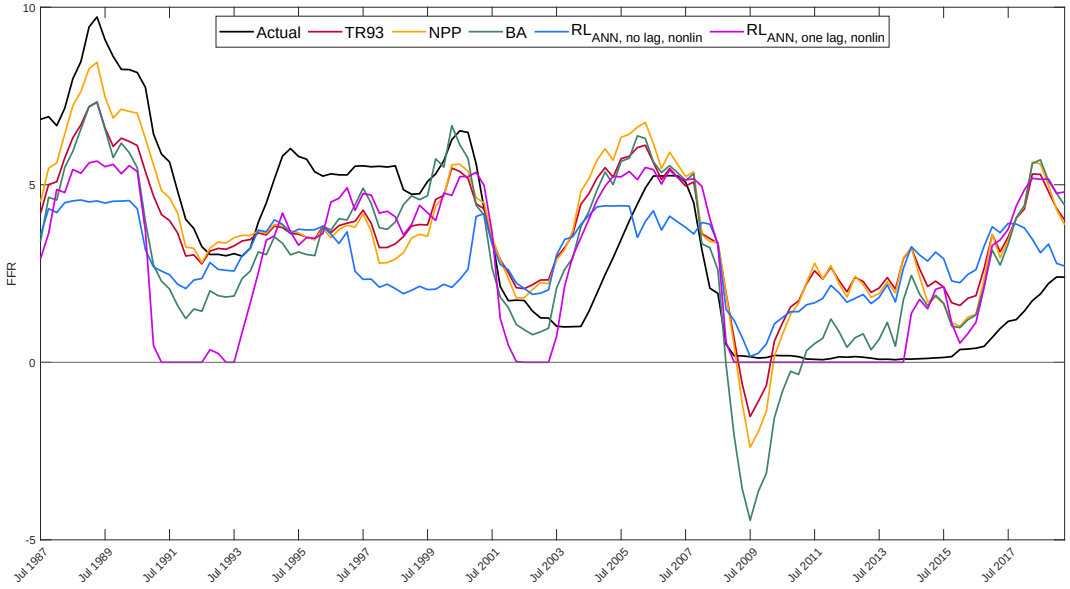
[42] For the counterfactual analyses, the natural rate of interest r^* and the inflation target π^* are set equal to 2 in TR93, BA and NPP.

Figure 20: FFR and Prescriptions from Common and Linear RL Rules based on ANN Economy



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

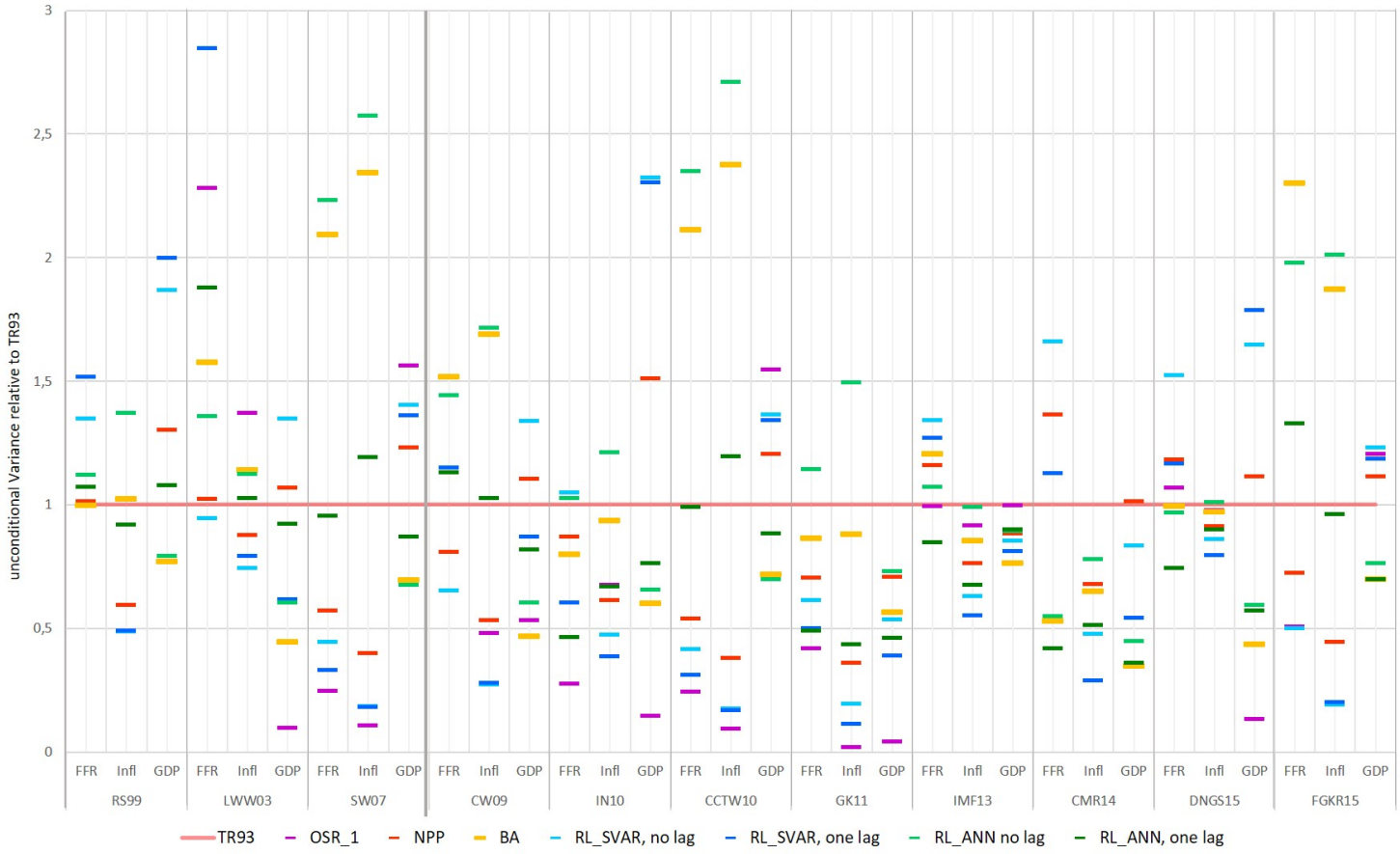
Figure 21: FFR and Prescriptions from Common and Nonlinear RL Rules based on ANN Economy



Note: This figure shows the static counterfactual of several common and the optimized policy rules within the ANN economy. *Actual*, *TR93*, *NPP* and *BA* are the same as before. Our ANN-based optimized nonlinear policy rules are depicted in blue ($RL_{ANN,\,no\,lag,\,nonlin}$) and purple ($RL_{ANN,\,one\,lag,\,nonlin}$).

C.2.4 Model Comparison

Figure 22: 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 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 10: 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 11: 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.