YES YOUNG ECONOMISTS SEMINAR

HRVATSKA NARODNA BANKA
EUROSUSTAV

19th YOUNG ECONOMISTS' SEMINAR
TO THE 32nd DUBROVNIK ECONOMIC CONFERENCE

May 28 – 29, 2026, Dubrovnik, Croatia

Tanja Linta

Overreaction in Expectations with Endogenous Feedback

Draft version
Please do not quote

Overreaction in Expectations with Endogenous Feedback*

Tanja Linta†

Toulouse School of Economics

March 30, 2026

Abstract

This paper measures biases in expectations within environments characterized by feedback loops between expectations and outcomes. Through a forecasting experiment, I provide evidence that although individuals systematically overreact to recent information, this overreaction is mitigated by stabilizing general equilibrium feedback. A simple theoretical model incorporating costly information processing shows that such a mitigation is feasible only if agents recognize the existence of feedback and adjust their behavior accordingly, thereby amplifying its stabilizing effects. Within the New Keynesian framework, stronger stabilizing feedback that attenuates the forecasting bias accelerates the convergence of endogenous variables to the rational expectations equilibrium. However, it does not eliminate overreaction, resulting in excess volatility in inflation responses to exogenous shocks. Consequently, monetary policy needs to respond more aggressively to an inflationary shock to achieve the same stabilizing effects as under rational expectations.

*I am extremely grateful to Christian Hellwig, Nicolas Werquin, and Eugenia Gonzalez-Aguado for their invaluable advice, guidance, and support. I would also like to thank Ingela Alger, Fernando Alvarez, George-Marios Angeletos, Charles Brendon, Alexander Guembel, Nour Meddahi, Maximilian Müller, Anna Sanktjohanser, and seminar/workshop participants at the Toulouse School of Economics, the Federal Reserve Bank of Chicago, and the 2025 Experimental Finance Conference for their valuable comments, suggestions, and discussions. I gratefully acknowledge funding from Banque de France. The experiment was preregistered under ID AEARCTR-0014403.

†tanja.linta@tse-fr.eu

1 Introduction

What economic agents expect about the future shapes their behavior, which ultimately determines the overall state of the economy. Conversely, the current state of the economy influences agents' beliefs about the future, creating a feedback loop between expectations and actual outcomes that often propagates and amplifies the effects of aggregate shocks. Therefore, how agents form expectations shapes equilibrium dynamics. This mechanism is at the core of many models in macroeconomics and finance. For instance, in the New Keynesian framework, aggregate expectations drive the dynamics of inflation and the output gap, while monetary policy aims to contain the self-reinforcing dynamics and stabilize the economy.

While feedback loops are commonly addressed by assuming that agents form expectations rationally—they have access to all available information, can process it effectively, and hold beliefs that are, on average, correct—an extensive empirical literature has documented a range of deviations from the rational benchmark (D'Acunto et al., 2023; Dräger and Lamla, 2024). However, most of this evidence abstracts from the effects of endogenous feedback loops. In survey-measured expectations, the environment is exogenously given, and in experiments, biases are frequently assessed in single-agent settings.

This paper examines the interaction between individual biases in expectations and general equilibrium feedback effects. Specifically, do feedback effects amplify or mitigate these biases? Conversely, how do biases influence general equilibrium dynamics? To answer these questions, I elicit individual expectations in a laboratory experiment designed to isolate feedback loops between expectations and outcomes. Experimental data show that stabilizing general equilibrium feedback—such as monetary policy effects which counteract positive feedback loops—reduces individual biases in expectations, thereby reinforcing its stabilizing impact on equilibrium dynamics. While individuals tend to systematically overreact to the most recent information relative to rational expectations, this overreaction is significantly diminished with stabilizing general equilibrium feedback.

To rationalize the empirical findings, I employ a simple theoretical model incorporating costly information processing and showing that such attenuation in overreaction is feasible only if individuals adjust their forecasting behavior in response to the feedback. Further-

more, by applying insights from the experiment and theoretical model into a canonical New Keynesian framework, I show how the measured interaction between expectations biases and general equilibrium feedback influences monetary policy. Specifically, overreaction in expectations generates excess volatility in inflation responses to exogenous shocks, prompting monetary policy to react more aggressively to an inflationary shock, relative to the rational expectations benchmark. However, if agents are aware of the feedback, their behavioral adjustment reduces volatility and accelerates convergence.

To cleanly measure and compare biases in expectations, I first conduct a laboratory experiment that generates forecasting data in environments with and without feedback loops, and while varying the persistence of the underlying process. In the experiment, participants are randomly assigned to separate treatment groups and asked to forecast future values of a process over 45 consecutive periods. Treatments vary along two dimensions: feedback and persistence. In the *Baseline* treatment, there is no feedback; the process is a simple and stable AR(1). In the *Feedback* groups, the aggregate forecast of the experimental economy enters the law of motion and affects realized outcomes. Both the *Baseline* and *Feedback* groups are further divided into treatments with zero or high underlying persistence. Variation in persistence is included because existing empirical evidence shows that expectation biases vary with persistence in the absence of feedback (Bordalo et al., 2018; Afrouzi et al., 2023). Participants are incentivized to maximize the accuracy of their forecasts and are made aware of the presence and the sign of feedback.

The experimental design focuses on negative feedback due to its relevance in the context of monetary policy. For instance, if a central bank raises the policy rate in response to high inflation, it creates incentives that push the economy in the opposite direction, turning the mapping from expectations to outcomes negative, rather than reinforcing the positive link between inflation expectations and inflation, or output expectations and output. Even if agents face information frictions or are boundedly rational, as suggested by overwhelming empirical evidence (see e.g., Dräger and Lamla (2024) for a survey), they are typically aware of the existence of a central bank and its role in responding to inflation. The key question then is whether and how expectations adjust to the known presence and sign of policy feedback. If expectations adjust, responses to shocks and inflation may moderate; if not, behavior can work against policy and amplify fluctuations.

Data generated in the experiment yield three empirical facts. First, they strongly reject the rational expectations hypothesis; participants systematically overreact to the most recent observations. Second, the overreaction is stronger for transitory processes, which is consistent with earlier evidence from survey data (Bordalo et al., 2018) and experimental environments without feedback loops (Afrouzi et al., 2023). Third, negative feedback mitigates the degree of overreaction, with the reduction being more pronounced for transitory processes.

The second part of the paper introduces a forecasting model where agents form biased beliefs about a long-run feature of the underlying process, providing a mechanism that matches all three empirical facts. Building on the framework in Afrouzi et al. (2023), and extending the environment to allow feedback, the model rests on two assumptions: (i) agents cannot process all available information, and (ii) recent information is more salient and easier to use, whereas older signals require cognitive effort. Agents, therefore, extrapolate from recent data when forming a prior belief about the long-run mean, and face costly processing of older information when updating the prior to improve their forecasting accuracy. This mechanism then generates a systematic overreaction that changes with persistence and feedback in a way that mirrors the experimental evidence. The most important implication of the model is that the documented attenuation in overreaction emerges only if agents condition their forecasts on the presence and sign of feedback, and is not just a consequence of the mechanical change in the process.

Competing expectations models cannot account for the three empirical facts. If forecasters ignore feedback and use a naive forecasting rule as if feedback were absent, the model predicts *larger* overreaction under negative feedback. Without adjusting forecasts for the presence of negative feedback, forecasters overshoot more, not less, which is reinforcing the interpretation that agents revise their rules in response to feedback. Furthermore, standard adaptive and extrapolative rules, which typically predict an overreaction, suggest patterns similar to the naive forecasting rule and consistently overshoot the observed degree of overreaction. Lastly, diagnostic expectations, as standardly defined in Bordalo et al. (2018), do not predict an overreaction for fully transitory processes unless biases affect more permanent features of the process. Taken together, these alternatives cannot reproduce the joint dependence of overreaction on both feedback and persistence.

Lastly, the paper embeds the forecasting bias documented in the experiment in a version of the New Keynesian model without rational expectations (Branch and McGough, 2009) to examine policy implications. First, overreaction to recent information generates inertia in expectations, causing excess persistence in the response of inflation to an exogenous shock. A central bank targeting inflation then needs to respond more strongly to achieve the same convergence path as implied by rational expectations. However, if agents internalize negative feedback from policy and, in response, reduce the degree of overreaction, as suggested by the forecasting model, the required reaction of monetary policy to exogenous shocks is relatively weaker. Less overreaction mutes down inertia in expectations and accelerates convergence, amplifying the stabilizing effects of monetary policy.

Taken together, these findings argue for models that permit deviations from rationality but recognize context-dependent forecasting rules. The degree of overreaction in expectations declines with persistence and negative feedback, which is a consequence of a behavioral adjustment in expectations, as implied by the forecasting model. The New Keynesian framework then shows that this behavioral adjustment reduces the policy intensity required for stabilization in reaction to exogenous shocks. However, because overreaction persists, achieving the same convergence path still requires a policy coefficient on inflation above the rational-expectations benchmark, but below what fully backward-looking models would prescribe. Moreover, agents' adjustment in beliefs to the presence and sign of feedback that reduces overreaction to recent news underscores the importance of credibility and effective communication in central banking.

1.1 Literature review

A large body of literature in macroeconomics and finance utilizes survey-measured expectations to reject the rational expectations hypothesis, yet there is little consensus on the direction of deviation (e.g., Coibion and Gorodnichenko (2015); Bouchaud et al. (2019); Ma et al. (2024); Bordalo et al. (2019, 2020); Kohlhas and Walther (2021)). Competing models emphasize underreaction (Mankiw and Reis, 2002; Sims, 2003; Maćkowiak and Wiederholt, 2009; Gabaix, 2014) versus overreaction (Barberis et al., 2015; Bordalo

et al., 2018; Beutel and Weber, 2025) to recent information, and recent work investigates reasons for the lack of consensus (Angeletos et al., 2021; Kučinskas and Peters, 2022; Afrouzi et al., 2023; Broer and Kohlhas, 2024).

Stepping away from survey data towards controlled experiments, and stripping down the forecasting environment to a simple and stable AR(1), Afrouzi et al. (2023) show that participants overreact to recent observations and that overreaction varies with the persistence of the process and the horizon of the forecast.[1] They argue that the experimental approach allows for control over confounding factors in a way that is not feasible with survey data, which helps reconcile the conflicting empirical evidence by showing that biases vary with the characteristics of the environment. This paper builds directly on Afrouzi et al. (2023) by introducing expectations feedback into the experimental forecasting setting to measure whether and how overreaction patterns change in response. He and Kučinskas (2024) address a related question by adding a correlated variable to an AR(1) experimental setup but abstracting from expectations feedback.

To incorporate the expectations feedback, the experimental framework adopts the learning-to-forecast approach initially introduced by Marimon and Sunder (1993) and extended into macroeconomics by Adam (2007) (see, e.g., Bao et al. (2021) for a survey). My experimental design closely relates to univariate models as in Bao and Duffy (2016) and Evans et al. (2025), which falls between Afrouzi et al. (2023) with an AR(1) setting and macro-laboratory frameworks with multiple endogenous variables and several exogenous shocks. Bao and Duffy (2016) examine the speed of convergence to rational expectations within a Cobweb model (Evans and Honkapohja, 2001) under adaptive versus eductive learning, while implementing negative feedback and varying its strength. The learning-to-forecast literature generally finds that agents can achieve rational expectations equilibrium, even with biased expectations, given sufficient negative feedback in the system. Within the New Keynesian framework, literature examining the role of expectations in stabilization via monetary policy links negative feedback to the central bank's response strength and the Taylor principle (Pfajfar and Žakelj, 2018; Kryvtsov and Petersen, 2019; Assenza et al., 2021). In a similar setup, Evans et al. (2025) use exper-

[1] Other research involving forecasting stochastic processes in experimental frameworks includes, among others, Hey (1994), Asparouhova et al. (2009), Reimers and Harvey (2011), Beshears et al. (2013) and Frydman and Nave (2017).

imental evidence with feedback ranging from positive to negative to validate a learning model unifying several existing theories. The present paper instead measures and compares biases in expectations within environments with and without feedback loops. The objective is to isolate feedback and its effect on biases in expectations that arise when feedback is present relative to those in a simple AR(1) setting. This paper is the first to apply the learning-to-forecast methodology to measure expectation biases in the style of the survey data literature.

Documenting how individuals form expectations is particularly valuable for environments where these expectations influence the system and shape the impact of shocks. However, individual biases interact with the feedback in the environment, and it is not immediately obvious what the consequences of such interaction are. L'Huillier et al. (2024) and Bianchi et al. (2024) implement diagnostic expectations, a type of model generating overreaction, in a standard New Keynesian model and show that endogenous extrapolation can generate excess persistence and volatility in endogenous variables or lead to repeated boom-bust cycles in response to an exogenous shock. This paper contributes by directly measuring the effects of the interaction between bias and feedback. It provides evidence that the extrapolation parameter is responsive to policy changes, suggesting that the formation of expectations is dependent on policy, which in turn affects the model's dynamics in a way that depends on behavioral adjustments.

The rest of the paper is structured as follows. Section 2 describes the experimental design. Section 3 discusses the empirical results. Section 4 shows the forecasting model explaining the mechanism behind the empirical findings. Section 5 applies the findings to the environment of the New Keynesian model. Section 6 concludes.

2 Experiment

To cleanly measure differences in expectations biases between environments with and without feedback loops, we can generate forecasting data within an experimental framework that allows for control of the data-generating process and the information available to participants. This section describes the experimental environment, treatments, incentives created for participants, and the logistics of the experiment.

2.1 Experimental environment

The experiment consists of a forecasting task in which participants predict the future values of a process over 45 periods, indexed by t. The process is defined as:

$$y_t = (1-\rho)\mu + \rho y_{t-1} + \delta F_t y_{t+1} + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2), \tag{1}$$

where $F_t y_{t+1} \equiv \frac{1}{N}\sum_{i=1}^{N} f_t^i y_{t+1}$ represents the aggregate forecast at time t for the value of the process in the next period, y_{t+1}, formed while observing the history of realizations of the process up to y_{t-1}. The aggregate forecast is computed as the average of all individual forecasts, $f_t^i y_{t+1}$, submitted in period t within a group of N participants.

The parameter δ determines the strength of the aggregate forecast's influence on the realizations of the process. If $\delta = 0$, the process collapses to a simple AR(1), with the persistence parameter $\rho \in [0, 1)$. If $\delta > 0$, there is positive feedback between aggregate expectations and outcomes, meaning that a higher aggregate forecast results in a higher actual realization of the process. Consequently, if the aggregate forecast deviates significantly from past realizations or the long-run mean, positive feedback can further destabilize the process. Suppose the forecast relies on the past values of the process in any way, which is the case even under rational expectations for $\rho > 0$. In that case, the impact of exogenous shocks will be amplified by the positive feedback.

In contrast, if $\delta < 0$, there is negative feedback between expectations and outcomes, which creates a mean-reverting force. As the aggregate forecast deviates in one direction, negative feedback pushes the process in the opposite direction. A higher aggregate forecast in one period tends to result in a lower realization of the process in the following period. In such cases, negative feedback mitigates the effects of exogenous shocks.

Before the experiment starts, participants receive a qualitative description of the forecasting environment and the process, including whether feedback is present, its sign, and what it signifies regarding the relationship between the variables (i.e., whether a higher aggregate forecast corresponds to a higher or lower actual realization). When $\delta = 0$, participants only observe and guess, and their forecasts do not influence the process. However, when $\delta \neq 0$, they have to consider (i) that their own forecast, along with forecasts of all other participants in their group, will affect future realizations of the process, and (ii)

that the aggregate forecast will provide either positive or negative feedback, making it more likely for the following period's realization to be higher or lower as a result.

2.2 Treatments

Main treatments vary along two dimensions, and are implemented through variations in two parameters in Equation (1): δ and ρ. The first parameter, δ, governs the presence of the expectations feedback, indicating whether and how strongly the process is influenced by the aggregate forecast of its future values. The second dimension represents the underlying persistence of the process (ρ).

2.2.1 Feedback

The first dimension of treatments varies the presence of feedback between expectations and outcomes, where in one scenario the feedback is a part of the process (*Feedback*) and in the other it is not (*Baseline*). The objective is to compare how forecasting behavior changes in environments with feedback, relative to predicting a simpler, exogenous process, to determine whether individuals employ the same forecasting rules in both situations or if they adjust their strategies based on their awareness of feedback and its sign. More specifically, the key question is whether individuals base their expectations solely on past values of the process they are predicting, regardless of whether expectations influence that process, or if they can incorporate potential effects of feedback into their predictions. In the *Baseline* scenario, $\delta = 0$, so the process simplifies to an AR(1) with $\rho \in [0, 1)$. This case corresponds to the experiment conducted by Afrouzi et al. (2023), and here serves as a control group. In contrast, the *Feedback* condition imposes $\delta = -0.5$, narrowing down the focus to negative feedback because of its relevance in the context of monetary policy.

In the monetary policy context, when a central bank increases the policy rate in response to high inflation, it creates incentives that push the economy in the opposite direction. Through the lens of a standard New Keynesian model (Galí, 2015), monetary policy that satisfies the Taylor principle generates negative feedback between output expectations and inflation, rather than reinforcing the positive relationship between inflation expectations and inflation, or output expectations and output. Instead, higher

future output expectations lead to lower inflation in equilibrium. A rational agent, anticipating high inflation, adjusts their expectations by incorporating the general equilibrium effects of the central bank's reaction. The question then becomes whether individuals with biased expectations can incorporate feedback from policy measures counteracting inflationary pressures into their predictions if they recognize its existence.

This question can be explored through an experiment based on a comprehensive macroeconomic model developed within the New Keynesian framework. In the learning-to-forecast literature (Bao et al., 2021), many experiments investigate the speed of convergence to equilibria and the effectiveness of monetary policy, considering varying signs and strengths of feedback in a complete macroeconomic, often New Keynesian, setup. However, the primary focus of this paper is to specifically isolate the marginal effect of feedback on the forecasting rules individuals adopt, as well as the biases in expectations that arise. Utilizing a large macroeconomic model would considerably complicate the measurement of how biases change with and without feedback, as one would need to control for, e.g., the effects of multiple exogenous shocks, correlated variables, and complex dynamics interacting with feedback effects.

2.2.2 Persistence

The second dimension of treatments involves variation in the underlying persistence of the process, ranging from fully transitory to highly persistent. Persistence is indicated by the parameter ρ, which takes the value $\rho = 0$ in the *Transitory* treatment group, and $\rho = 0.9$ in the *Persistent*. Previous research (Afrouzi et al., 2023; He and Kučinskas, 2024) has documented differences in expectation formation patterns across environments with varying levels of persistence; however, these studies did not incorporate expectations feedback. The purpose of the second dimension of treatments is to determine whether the variations in biases with persistence previously recorded in the literature, in similar but entirely exogenous environments, also apply to settings that include feedback. Afrouzi et al. (2023) also show that these differences in biases with respect to persistence can help rule out potential explanations for the observed patterns in the data, i.e., certain models of expectation formation deviating from rational expectations that cannot simultaneously

account for how biases vary with persistence.

The variation in feedback and persistence results in four treatment groups, which encompass all combinations of *Baseline*, *Feedback*, *Transitory*, and *Persistent*. For simplicity, the parameter μ is set to zero in all treatments, and $\sigma_\epsilon^2 = 2$. There is only one sequence of shocks that remains constant across all treatments. Figure E.2 illustrates the specific processes that participants face in the experiment, based on the given sequence of shocks. In the *Feedback* condition, the values of the process in the Figure are simulated under the assumption that expectations are formed rationally. However, participants in the experiment observe the realized values that are influenced by the aggregate forecast of their group. Under rational expectations in the transitory case, feedback has no impact, resulting in the two processes overlapping perfectly. In both scenarios, the best guess remains zero in every period, regardless of feedback. Conversely, in the persistent case, negative feedback reduces the amplitude of the variation in the process.

2.3 Logistics

The experimental design integrates two approaches that have been used separately in the existing literature. The first one is applied to the *Baseline* condition and resembles the setup in, among others, Afrouzi et al. (2023), where participants predict an AR(1) process. Since the process is exogenous, participants' forecasts do not influence the outcomes. In each period t, participants individually observe the history of realizations of the process up to period $t-1$ and provide incentivized predictions for the process's realization in the next period, $t+1$. Within the *Baseline* condition, participants are randomly assigned to either *Transitory* or *Persistent* group.

In the *Feedback* condition, the design follows the learning-to-forecast approach (Bao et al., 2021). Similar to the *Baseline*, participants are split into two separate persistence groups. However, within the two persistence groups, participants are further divided and randomly assigned to groups of six, where each group represents one experimental economy that remains consistent throughout the experiment. They similarly observe all realizations of the process up to period $t-1$ and individually provide predictions for the realization of the process in the following period. The key difference in the *Feedback*

condition is that each period, after participants individually submit their predictions, they are aggregated within groups such that the mean defines $F_t y_{t+1}$ in Equation (1). Consequently, the group's aggregate forecast affects the actual realizations of the process. This feedback from the aggregate forecast is why the process can vary between different groups of six, even if they face precisely the same sequence of shocks.

Because the aggregate forecast formed in period t defines the actual realization of the process in that same period, participants in *Feedback* treatments cannot observe the time t realizations of the process. To maintain consistency, the same restriction is also imposed in the *Baseline* condition. Participants can only observe the previous realizations of the process up to $t-1$ and their own previously submitted forecasts; they do not have access to the predictions made by others or to the realizations of shocks.

Before starting the experiment, participants read a qualitative description of the underlying environment. In the *Baseline* condition, they are asked to "predict the future values of a random process over 45 periods". In the *Feedback* condition, participants receive the same basic instructions as in the *Baseline*, but with added information: (i) that they are predicting in a group, (ii) that each period the computer will calculate the average of their individual predictions within the group, and (iii) that this average prediction will *negatively* feed back into the process they are predicting. The primary objective is to ensure that forecasters understand the process is not entirely exogenous; rather, their forecasts now influence the outcomes they aim to predict. Participants do not observe the equations or the values of the underlying parameters, to avoid explicitly providing them with forecasting rules that the experiment seeks to uncover. In both conditions, participants are informed about the information available on their screen and how their payoffs will be calculated. Full experimental instructions can be found in Appendix C. The interface was programmed using oTree software (Chen et al., 2016), and a screenshot is provided in Figure E.1.

2.4 Procedures

Participants are compensated based on the accuracy of their forecasts. In each period t, their forecast score is computed as $100/(1+|f^i_{t-1}y_t - y_t|)$, where $f^i_{t-1}y_t$ is the individual

forecast at time $t-1$ for the value of the process in period t, y_t. To incentivize accuracy, the forecast score decreases with the distance from the realized value. The final score for each participant is the sum of their forecast scores across all periods of the experiment, which is then converted to euros at the rate of 75 cents for every 100 points. Additionally, each participant receives a participation fee of 3 EUR. The average payout per participant is approximately 15 EUR.

The experiment uses a between-subjects design, meaning that each participant is assigned to only one treatment group. In the *Feedback* condition, as is standard in the learning-to-forecast literature, there are six groups of six participants in each of the two persistence groups, with one group of six representing a single experimental economy. In the *Baseline*, to match the participant numbers to those in the *Feedback* condition, there are 36 participants within each persistence treatment. Each session lasted on average 45 minutes. The participant pool consisted of 143 predominantly undergraduate students at the Toulouse School of Economics, where the experiment took place.[2] Furthermore, the experiment was conducted in a laboratory setting. Due to the nature of the game, particularly in the *Feedback* condition where individual forecasts are aggregated within groups each period, participants need to play at the same time, which is impossible to achieve in online experiments, even if there is an added benefit of a more diverse subject pool.

3 Empirical findings

The empirical analysis consists of two main components: (i) measuring the degree of overreaction in forecasting behavior across treatment groups, and (ii) investigating how the observed overreaction interacts with negative feedback in the propagation of exogenous shocks and exploring the implications of the forecasting bias for the dynamics of the process. The final sample includes a total of 5,567 observations, after excluding the first five periods and a few extreme outliers, as explained in Appendix D.

[2]Data for one participant in the *Baseline* treatment was excluded due to technical difficulties during the experiment.

3.1 Overreaction in individual forecasts

3.1.1 Forecast errors

Table 1 reports summary statistics for individual forecast errors across treatment groups, which, as an indicator of forecasting accuracy, will be a key element in measuring overreaction. They are defined as $y_{t+1} - f_t^i y_{t+1}$, where y_{t+1} is the actual realization of the process in period $t+1$ and $f_t^i y_{t+1}$ is the individual forecast for the same period of a participant i submitted at time t.

Table 1: **Forecast Errors Summary Statistics**

Treatment		Mean	Median	SD	Min	Max	N
Baseline	Transitory	-0.8	-1.0	2.8	-7.9	9.3	1365
	Persistent	-1.0	-1.3	3.5	-12.6	9.5	1404
Feedback	Transitory	-0.7	-0.8	3.1	-16.3	20.5	1401
	Persistent	-0.6	-0.7	3.2	-15.6	15.0	1397

Note: Forecast errors are defined as $y_{t+1} - f_t^i y_{t+1}$, where y_{t+1} is the actual realization of the process in period $t+1$ and $f_t^i y_{t+1}$ is the individual forecast for the same period of a participant i submitted at time t. Treatment groups vary in the presence of feedback (Baseline versus Feedback) and in persistence (Transitory versus Persistent).

Under rational expectations, forecast errors average out around zero, irrespective of feedback or persistence. However, in the data, the errors are on average negative across all treatments, indicating a systematic tendency to overpredict, where individual forecasts overshoot the actual outcomes. Negative feedback helps reduce this bias in both persistence groups, resulting in less negative forecast errors, with a stronger attenuation for higher persistence. Interestingly, the role of high persistence varies across feedback groups. In the absence of feedback, high persistence increases both the bias and the variability of forecast errors, relative to the transitory case. Conversely, when negative feedback is present, higher persistence instead reduces the bias, suggesting a stabilizing interaction between the two features. Figures E.3 and E.4 illustrate the histograms of individual forecasts and forecast errors, respectively, across all treatment groups.

To quantify the differences in forecast error means observed between treatment groups and assess their significance, we can estimate the treatment effects of high persistence and

feedback on forecast errors, relative to the transitory *Baseline* as:

$$y_{t+1} - f_{it}^{i} y_{t+1} = \alpha + \beta_1 \text{Persistent} + \beta_2 \text{Feedback} + \beta_3 (\text{Persistent} \times \text{Feedback}) + \epsilon_t, \quad (2)$$

where *Persistent* and *Feedback* are indicator variables for treatments with high persistence and negative feedback, respectively. Table 2 shows the estimates.

Table 2: **Treatment effects on Forecast Errors**

	Intercept	Persistent	Feedback	Persistent × Feedback
Coefficient	−0.82	−0.19	0.10	0.29
	(0.05)	(0.09)	(0.08)	(0.15)

Note: The table reports the estimates from a pooled regression of forecast errors, $y_{t+1} - f_{it} y_{t+1}$, on indicator variables where $Persistent = 1$ indicates groups with high persistence and $Feedback = 1$ treatments with negative feedback. The intercept reports the mean forecast error in the Baseline-Transitory group. Standard errors, in parentheses, are clustered at the individual level.

In the transitory *Baseline*, the mean forecast error is around –0.8. While negative feedback alone helps to reduce the errors towards zero, the effect is not statistically significant. In contrast, high persistence makes them significantly more negative. The interaction between persistence and feedback is positive and has the most significant impact among all categories. When both persistence and feedback are considered, the combined effect compared to the transitory *Baseline* leads to a notable decrease in average forecast errors. Overall, feedback helps reduce bias in forecast errors, with the most substantial reduction occurring when the process is highly persistent.

3.1.2 Measuring overreaction

To measure overreaction in forecasting behavior relative to the rational expectations benchmark, we can estimate the following equation:

$$y_{t+1} - f_t^i y_{t+1} = \alpha_i + \beta y_{t-1} + \nu_{it}, \tag{3}$$

which is regressing the individual forecast error $(y_{t+1} - f_t^i y_{t+1})$ on the most recent observation available to participants at the time of making the forecast, y_{t-1}.

This approach to measuring underreaction or overreaction allows us to compare forecasting behavior to the rational expectations benchmark. According to the rational expectations hypothesis, the coefficient β should be zero, indicating that forecast errors are not systematically predictable by the information available to forecasters at the time of providing the forecast. A non-zero coefficient signifies a deviation from the rational benchmark, and the sign reveals its nature. $\beta < 0$ occurs if and only if individual forecasts react more strongly to recent information, relative to the rational expectations forecast, suggesting an overreaction. Conversely, $\beta > 0$ signals a relative underreaction.[3] Additionally, the magnitude of the deviation provides insight into how strongly forecasters rely on the last observation when making their forecasts across different treatment groups.

There are several methods to measure overreaction in this context that have been previously used in the literature. Initially, Coibion and Gorodnichenko (2015) measured the correlation between forecast errors and forecast revisions, arguing that revisions reflect how forecasts react to new information. Revisions are often defined as the difference between forecasts for the same future value of the process made in two consecutive periods $(f_t^i y_{t+1} - f_{t-1}^i y_{t+1})$. However, Afrouzi et al. (2023) critique this approach and suggest instead measuring overreaction through "forecast-implied persistence", which refers to the sensitivity of individual forecasts to the most recently observed information and is straightforward to estimate within their AR(1) framework. The main criticism they raise is two-fold. First, estimation can be challenging, especially for transitory processes when expectations are close to rational. In such cases, revisions tend to be near zero, and the regression coefficient is not well estimated. Second, if forecasts are measured with noise,

[3] See e.g., Broer and Kohlhas (2024, p. 1338–1339) for the formal proof.

because they appear on both sides of the regression equation, they can make the estimated coefficient mechanically negative.

Even though forecasts in this paper are also measured in a controlled experimental environment, adding expectations feedback to the data-generating process makes it significantly more complicated, so the estimation of forecast-implied persistence is not as straightforward or clean. The approach adopted here, as defined in equation (3), is similar to those in Kohlhas and Walther (2021) or Broer and Kohlhas (2024), and is not subject to estimation difficulties mentioned above.

Table 3: **Overreaction estimates by treatment**

	Transitory		**Persistent**	
	Baseline	Feedback	Baseline	Feedback
y_{t-1}	-0.65	-0.49	-0.24	-0.20
	(0.05)	(0.05)	(0.03)	(0.05)
Intercept	-0.72	-0.73	-0.84	-0.54
	(0.05)	(0.07)	(0.09)	(0.10)
Differences in y_{t-1} coefficients:				
B - F	-0.15		-0.04	
	(0.07)		(0.06)	
N	1330	1365	1368	1361

Note: The table shows estimates of the correlation between forecast errors $(y_{t+1} - f_t^i y_{t+1})$ and the last observation available to the participants, y_{t-1}, as defined in equation (3), separately for each of the four treatment groups. The bottom panel displays the differences in estimated coefficients between the Baseline and Feedback conditions in both persistence groups. Standard errors, in parentheses, are clustered at the individual level.

Table 3 reports the estimates of β from equation (3) for individual forecasts, categorized by treatment group. First, the coefficients are negative and highly significant, which is inconsistent with rational expectations and indicates an overreaction in forecasting behavior.

Second, the overreaction in the *Baseline* group is more pronounced for transitory processes, as evidenced by more negative coefficients, which is a pattern documented previously in research using both experimental and survey data (Bordalo et al., 2020; Afrouzi et al., 2023). Table F.1 compares the estimates from the *Baseline* case in Table

3 with the corresponding estimates using data from Afrouzi et al. (2023) for the closest available persistence levels. Their experiment was conducted online with the general population, whereas the data here was collected in a laboratory setting. The results in Table F.1 indicate that the estimates are quite similar. While some discrepancies are expected due to differences in shock sequences between the two settings and because the closest persistence level corresponding to $\rho = 0.9$ in this paper is $\rho = 0.8$ in their setting, the signs of the coefficients remain the same, and the magnitudes are comparable.

A novel finding is the evidence that overreaction and its variation with persistence extend to environments with negative feedback, where overreaction is still significantly more pronounced when the process is transitory. Most importantly, relative to the *Baseline*, the degree of overreaction is weaker with negative feedback, regardless of the persistence of the process. Notably, the decrease is both stronger and statistically significant for transitory process. In summary, four empirical facts emerge: relative to the rational benchmark, (i) forecasters tend to overreact, (ii) the overreaction is stronger for transitory processes, (iii) negative feedback mitigates the degree of overreaction, and (iv) the reduction in overreaction with negative feedback is more pronounced for transitory processes.

3.1.3 Average forecasts

The literature often shows that estimating forecast error regressions using individual forecasts, instead of average or consensus data, yields different results, particularly in the signs of the coefficients, which are difficult to reconcile. For instance, using data from the Survey of Professional Forecasters, Coibion and Gorodnichenko (2015) find a positive correlation between forecast errors and forecast revisions, indicating that average forecasts tend to underreact to the most recent information. In contrast, Bordalo et al. (2020), among others, using the same data but focusing instead on individual forecasts, find a negative correlation that suggests an overreaction in forecasting behavior. Furthermore, Kohlhas and Walther (2021) show a negative correlation between forecast errors and the last realization of the process even in the average forecasts.

Angeletos et al. (2021) show that these empirical patterns, when examined dynamically, can be explained by a combination of an initial underreaction and a delayed overre-

action to the same information. Table F.2 presents the same estimates as in Table 3 but using aggregate forecasts instead of individual ones, showing that the results remain essentially unchanged. In aggregate form, the estimates remain negative and have a similar magnitude, suggesting an overreaction also in the average forecasts. In the experimental framework of this paper, all information is public, and the realizations of the process are observed without noise. If noise were present, it might generate an underreaction in either or both individual and average forecasts.

3.1.4 Overreaction over time

The objective of this section is to evaluate whether the overreaction in forecasts to one piece of news continues across subsequent periods. To understand the potential consequences of this bias, it is helpful to determine how persistent the bias is over time, and specifically whether it accumulates over multiple periods or decays quickly after an exogenous shock.

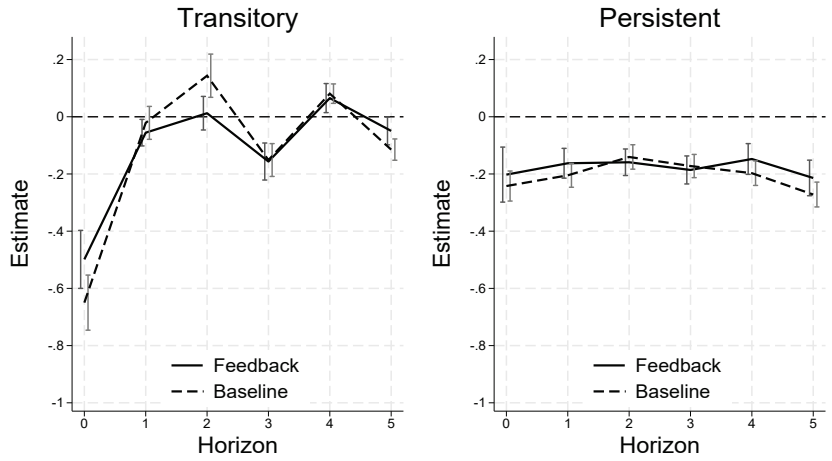
For horizon $h = 0, 1, ..., H$, we can estimate the following:

$$y_{t+h} - f_t^i y_{t+h} = \alpha_i + \beta^h y_{t-1} + \nu_{it+h} \tag{4}$$

where β^h represents co-movement between forecast errors h periods ahead to the same past realization y_{t-1}. Estimates at $h = 0$ correspond to the estimates in Table 3, while estimates for $h > 0$ trace how the influence of the past observation persists over time. Figure 1 displays the estimated β^h for each of the four treatment groups.

As discussed in the previous section, the overreaction at $h = 0$ is stronger in transitory cases, both with and without feedback, compared to persistent. However, the overreaction is smaller when feedback is present relative to the *Baseline* scenario, while there is no significant difference in the persistent case regardless of feedback. Over time, although the overreaction is initially more substantial in the transitory environment, it quickly decays toward zero after $h = 0$ as forecast errors cease to be systematically linked to y_{t-1} after the first period. Additionally, the subsequent variation in the transitory scenario is more or less the same in both *Feedback* and *Baseline* conditions. In contrast, the coefficients for both *Baseline* and *Feedback* in the persistent case show little decay across

Figure 1: **Overreaction over time**



Note: The figure plots estimates of β^h from horizon-h regressions of forecast errors, $y_{t+h} - f_t^i y_{t+h}$, on the last observed value y_{t-1}, separately for each treatment. The left (right) panel shows the Transitory (Persistent) treatment group; solid lines denote responses in the Feedback groups and dashed lines responses in the Baseline. Points are coefficient estimates with 95% confidence intervals. Standard errors are clustered at the individual level.

horizons, so the impact of the same past observation remains present for several periods ahead, however with only modest differences throughout.

The observed changes of overreaction over time suggest that the same mechanism operates differently depending on persistence. When shocks are short-lived, the key concern is volatility; forecasters initially react strongly to the latest observation but then adjust away from it. In contrast, in persistent environments, the main issue is inertia; while the overreaction is milder, it persists over time. Negative feedback reduces the initial overshooting, but this effect is only evident in the transitory environment.

3.2 Dynamic responses to shocks

The second part of the empirical analysis examines how the observed overreaction in forecasting behavior interacts with negative feedback and its consequences for the dynamics of the process across different treatment groups. Negative feedback influences both forecasting behavior, including biases, and the process itself. This section compares dynamic responses of the process and the forecasts to the same exogenous shocks across treatment groups, to understand whether and how shocks propagate differently based on

variations in feedback and persistence.

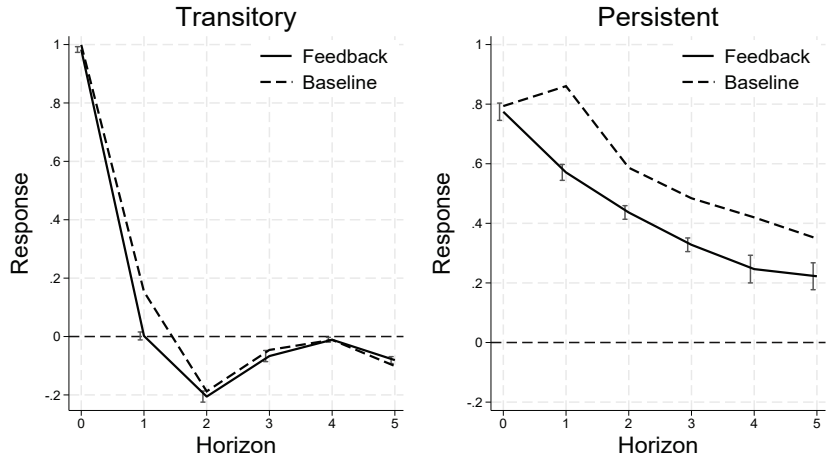
3.2.1 Response of outcomes to shocks

To quantify how biased forecasts and negative feedback shape the propagation of exogenous shocks to outcomes, we can estimate the following impulse response functions:

$$y_{t+h} = \alpha + \gamma_{1h}\epsilon_t + u_{t+h} \tag{5}$$

which for $h = 0, 1, ..., H$ measures the reaction of y_{t+h} to the exogenous ϵ_t from the experimental process across the four treatment groups. The coefficients γ_{1h} are interpreted as the response of y_{t+h} to a one-unit shock in ϵ_t. Figure 2 shows the estimates.

Figure 2: **Response of y_{t+h} to ϵ_t**



Note: The figure plots estimates of γ_{1h} from horizon-h local-projection impulse response regressions, as specified in equation (5), run separately by treatment group. The left (right) panel shows the Transitory (Persistent) group; solid lines denote responses in the Feedback groups and dashed lines responses in the Baseline. Points show the estimated response of y_{t+h} to a one-unit innovation in ϵ_t. Vertical bars are 95% confidence intervals. Standard errors are clustered at the individual level.

In the transitory case, both with and without feedback, the initial impact of shocks is larger compared to the persistent scenarios, but it decays rapidly, essentially vanishing after the first period. The *Feedback* curve lies below the *Baseline*, indicating that negative feedback reduces the pass-through effect of shocks. In the persistent environment, the response is relatively smaller on impact than in the transitory case, and it declines only gradually over time. However, negative feedback shifts the entire path down significantly more than in the transitory case, suggesting a more negligible and faster-decaying effect

of shocks on the process when forecasts feed back negatively into outcomes, as opposed to when there is no feedback.

The overall implication is that negative feedback dampens the effect of shocks on y_t, with a proportionally stronger attenuation in the persistent setting. At least in the transitory context, it is straightforward to see that this pattern is consistent with forecasting behavior amplifying the stabilizing effect of negative feedback. Under rational expectations, negative feedback in theory should not have an impact on the outcomes of the process, and the two lines on the left-hand side panel of Figure 2 should overlap.

3.2.2 Response of forecasts to shocks

Similarly to the analysis of outcomes, to quantify the response of individual forecasts to shocks, we can estimate the following:

$$f^i_{t+h} y_{t+h+1} = \alpha + \gamma_{2h} \epsilon_t + u_{t+h} \tag{6}$$

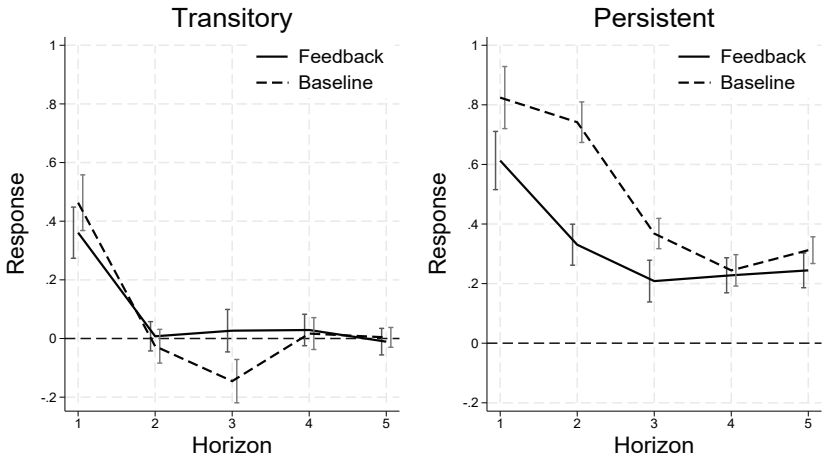
which for horizon $h = 1, 2, ..., H$ measures the reaction of $f^i_{t+h} y_{t+h+1}$ to the exogenous ϵ_t from the experiment across the treatment groups. The coefficients γ_{2h} are interpreted as the response of individual forecasts $f^i_{t+h} y_{t+h+1}$ to a one-unit shock at time t.[4] Figure 3 shows the estimates.

In contrast to the response of outcomes to exogenous shocks, the response of forecasts in the transitory case is smaller than in the persistent case, both with and without feedback, but it also decays rapidly and vanishes after the first period. There are almost no differences between *Feedback* and *Baseline* in the transitory environment. In the persistent case, the initial impact is relatively larger than in the transitory case, but it decays more gradually. Similarly to the responses in outcomes, negative feedback in the persistent case shifts the entire path down, implying a smaller effect of shocks on the forecasts with negative feedback, relative to the *Baseline*.

Figure E.5 compares the dynamic responses of both outcomes and forecasts, as shown separately in Figures 2 and 3, across the treatment groups. An alternative definition of

[4]Horizon $h = 0$ response is omitted because participants in the experiment do not observe the current realization of the process (y_t).

Figure 3: **Response of f^i_{t+h} to ϵ_t**



Note: The figure plots estimates of γ_{2h} from horizon-h local-projection impulse response regressions, as specified in equation (6), run separately by treatment group. The left (right) panel shows the Transitory (Persistent) group; solid lines denote responses in the Feedback groups and dashed lines responses in the Baseline. Points show the estimated response of $f^i_{t+h} y_{t+h+1}$ to a one-unit innovation in ϵ. Responses start at horizon $h = 1$ because participants in the experiment observe realizations of the process only up to period $t - 1$. Vertical bars are 95% confidence intervals. Standard errors are clustered at the individual level.

under- and overreaction comes from comparing whether forecasts react to shocks more or less than the process itself reacts to the same shock, as in, for example, Angeletos et al. (2021). If forecasts overshoot outcomes, they respond more strongly to shocks than the process itself, defining an overreaction. Overreaction, as described in section 3.1, compares reactions of individual forecasts to the rational expectations benchmark. Here, the definition is slightly different, but the figure shows that the conclusion is the same. In the transitory case, individual forecasts react to shocks more strongly than the process itself, while both the response of the process and the forecasts are smaller in *Feedback* relative to *Baseline*. In the persistent case, the reaction of forecasts and outcomes is generally at the same level, indicating that overreaction, measured in this way, does not change significantly between environments with and without feedback when the process is persistent.

3.3 Response times

To further examine whether forecasting behavior differs between groups with and without feedback, and whether participants forecast differently under these conditions, we can

measure the amount of time they spend in each period or on each forecasting page during the experiment. In the *Feedback* condition, participants are aware of the presence of feedback and its impact on the process they are forecasting. However, nothing prevents them from disregarding the extra information and basing their forecasts on previous observations in the same way as participants in the *Baseline* group. If they do take feedback into account, they face a slightly more complex task that requires somewhat more effort.

Table 4 compares the average time in seconds participants spent on each forecasting page between *Baseline* and *Feedback* conditions, and shows that participants facing a process with feedback spent about 7 seconds more per page than those in the *Baseline*, indicating a 37% increase in thinking time. The averages exclude instruction screens and waiting times within groups.

Table 4: **Average Response Times**

	Baseline	**Feedback**	**F–B**
Seconds per round	19.12	26.12	7.01
	(0.95)	(0.98)	(1.36)
MW p-value			< 0.001

Note: Means are computed at the participant level over all rounds and then averaged within treatment. "F–B" reports the difference in means (Feedback minus Baseline). MW p-values is from a two-sample Mann–Whitney test comparing the distributions of participant-level average times across treatments and p<0.001 rejects equality of distributions. Waiting times within groups and instruction screens are excluded. Standard errors, in parentheses, are clustered at the individual level.

Because one page equals one period, this difference amounts to roughly 4.7 additional minutes over 40 forecasting periods. This gap indicates that the *Feedback* condition required more deliberation when forming forecast and that participants appear to have adjusted their behavior in response, instead of ignoring the presence of feedback.

4 Forecasting model

To put structure on the documented patterns in the data, this section presents a forecasting model that provides a mechanism that matches all empirical facts. The empirical findings described in section 3 show a systematic deviation from rational expectations: similar to previous research, forecasters tend to overreact to recent information when

predicting an AR(1) process. Additionally, the overreaction also occurs in the context of expectations feedback, but is notably weaker when the feedback is negative, and more so when there is less persistence.

Afrouzi et al. (2023) develop a model in which forecasters face costly processing of past information when forming beliefs about the long-run mean of a stable AR(1) process they are predicting. They show that this friction accounts for the variation in overreaction across different persistence levels and forecast horizons. Moreover, they show that alternative models cannot simultaneously explain all recorded facts.

The extension introduced here incorporates the same friction in an environment with expectations feedback, showing that it similarly accounts for variations in overreaction across different feedback regimes. The description of the environment follows the steps outlined in their paper, with the core elements summarized before introducing the feedback extension.

4.1 Environment

The process that the agent is forecasting is given by:

$$y_t = (1-\rho)\mu + \rho y_{t-1} + \delta F_t y_{t+1} + \nu_t, \tag{7}$$

where $F_t y_{t+1}$ is the aggregate forecast at time t of y_{t+1}, δ captures the strength of expectations feedback, and $\nu_t \sim i.i.d.N(0, \sigma_\nu^2)$ is an exogenous shock. Agents are rewarded for forecast accuracy, with payoff $-(f_t^i y_{t+1} - y_{t+1})^2$, where $f_t^i y_{t+1}$ denotes agent i's time t forecast of y_{t+1}.

Agents lack complete information about μ but can acquire it, at a cost, to improve their forecast accuracy. In every period t, the agent observes the most recent realization y_{t-1}, forms a prior $\mu \sim N(y_{t-1}, \tau^{-1})$, where τ denotes its precision, and then decides whether or not to process more information to update the prior. However, processing information is subject to a cost that increases with the amount of information that is processed, denoted S_t, which includes y_{t-1}. Information processing entails a convex cost, specified as

$$C_t(S_t) = \omega \frac{\exp\left(\gamma \cdot \mathbb{I}(S_t; \mu \mid y_{t-1})\right) - 1}{\gamma}, \tag{8}$$

where $\omega \geq 0$ scales the cost, $\gamma \geq 0$ determines convexity, and $\mathbb{I}(S_t; \mu \mid y_{t-1})$ is Shannon's mutual information, measuring the expected reduction in uncertainty about μ, given S_t.[5] Rational expectations are nested as a special case when information is either costless or fully processed. Having the prior centred around the last observation generates the overreaction to y_{t-1}, which decreases as the agent processes more information and gets closer to the true μ.

The model is based on two key assumptions: (i) agents can only process a subset of available information, and (ii) recent information is easier to use, while any additional processing incurs a cognitive cost. The first assumption is widely discussed in the literature on working memory (see, e.g., Cowan (2017) for a survey), suggesting that individuals are often unable to effectively process all signals in their environment and thereby focus on only one part of it. The second assumption draws from the psychology literature that explores how information enters working memory (Evans, 2008; Hitch et al., 2018). The first method is centered around recency, highlighting the importance of the most recent information, which quickly enters working memory. The second method requires conscious deliberation and cognitive effort to select and process additional information.[6]

4.1.1 Agent's problem

The agent solves the problem of choosing an information set S_t that balances forecast accuracy against the cost of information acquisition:

$$\min_{S_t} \mathbb{E}\left[\min_{f_t^i} \mathbb{E}\left[(f_t^i y_{t+1} - y_{t+1})^2 | S_t\right] + C_t(S_t)\right], \tag{9}$$

subject to $y_{t-1} \subseteq S_t \subseteq \mathcal{A}_t$, where $\mathcal{A}_t$ is the set of all available signals at time t.

[5] The cost function is defined exactly as in Afrouzi et al. (2023), who build on literature in rational inattention (e.g., Woodford (2014) or Kőszegi and Matějka (2020)), where processing costs traditionally represent cognitive constraints. Their setting diverges in that the processing costs are related to the processing of past observations, and it is the closest to da Silveira et al. (2024), who model the optimal choice of the structure of memory. The cost specification is equivalent to the one in Sims (2003) when $\gamma \to 0$ and the cost is linear in Shannon's mutual information function.

[6] See Afrouzi et al. (2023) for a detailed overview of the literature.

The aggregate forecast is conjectured to take the form

$$F_t y_{t+1} = \psi y_{t-1}, \tag{10}$$

which simplifies the process to

$$y_t = (1-\rho)\mu + (\rho + \delta\psi) y_{t-1} + \nu_t, \tag{11}$$

with the stability condition $|\rho+\delta\psi|< 1$. Afrouzi et al. (2023) show that the agent's problem in (9) can be reduced to choosing the precision of the posterior belief about μ, denoted τ, which minimizes expected forecast errors and the information costs. Incorporating feedback, this problem becomes

$$\min_{\underline{\tau}\leq\tau\leq\bar{\tau}} \left[\frac{(1+\rho+\delta\psi)^2(1-\rho)^2}{\tau} + \omega \frac{\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma} \right] \tag{12}$$

where $\tau \equiv \text{Var}(\mu \mid S_t)^{-1}$ is the precision of the posterior, $\underline{\tau}$ is the prior precision, and $\bar{\tau}$ is the maximum precision attainable under the full information set. The upper bound is assumed not to bind when $\mathcal{A}_t$ is sufficiently large.

Fixing $\xi \equiv (\mu, \rho, \delta, \omega, \gamma, \underline{\tau})$, the resulting optimal posterior precision is

$$\tau^*(\psi;\xi) = \underline{\tau} \max\left\{1, \left(\frac{(1+\rho+\delta\psi)^2(1-\rho)^2}{\omega\underline{\tau}}\right)^{\frac{1}{1+\gamma}}\right\}. \tag{13}$$

The solution shows that posterior precision depends jointly on the process characteristics, persistence (ρ) and feedback (δ), as well as on the aggregate forecast parameter ψ. While in Afrouzi et al. (2023) the solution yields a fixed τ^*, here τ is a function of ψ, and ψ itself is determined in equilibrium. In equilibrium, the optimal posterior precision τ^* and the form of individual forecasts jointly pin down ψ. See Appendix A.1 for a formal proof.

The optimal posterior precision τ^*, resulting from the agent's information choice, defines the weight that agent assigns to the most recent observation y_{t-1} when forming their beliefs about μ, which characterizes the optimal individual forecast and the resulting

degree of overreaction measured by the correlation between forecast errors and y_{t-1}.

Given the information choice, the conditional mean of the posterior $\mathbb{E}[\mu|S_t]$ is a convex combination of the true μ and the last observation y_{t-1},

$$\mathbb{E}[\mathbb{E}[\mu|S_t]|\mu, y_{t-1}] = (1-\alpha)\mu + \alpha y_{t-1}, \quad \alpha(\psi;\xi) \equiv \frac{\tau}{\tau^*(\psi;\xi)} \in (0,1] \tag{14}$$

where α denotes the ratio of the precisions of the prior and the posterior. The frictionless benchmark corresponds to the limit $\alpha \to 0$, i.e., $\tau^*(\psi;\xi) \to \infty$, while $\alpha = 1$ corresponds to no information processing. The optimal individual forecast then is

$$f_t^{i*} y_{t+1} = (1+\rho+\delta\psi)(1-\rho)\left[(1-\alpha)\mu + \alpha y_{t-1}\right] + (\rho+\delta\psi)^2 y_{t-1} + u_t, \tag{15}$$

where $\mathbb{E}[u_t|y_{t-1}] = 0$. See Appendix A.2 for a formal proof.

4.1.2 Equilibrium

Fixing primitives $\xi \equiv (\mu, \rho, \delta, \omega, \gamma, \underline{\tau})$ and given a conjectured aggregate forecasting rule $F_t y_{t+1} = \psi y_{t-1}$, the agent's information processing choice delivers posterior precision $\tau^*(\psi;\xi)$ as defined in (13), which defines the learning parameter $\alpha(\psi;\xi)$ as in (14) and the individual optimal forecast as in (15), with the weight $\lambda(\rho, \delta, \psi)$ as in (18). A symmetric equilibrium is a pair (ψ^*, α^*) such that:

$$\text{(E1) } \alpha^* = \alpha(\psi^*;\xi) \quad \text{(Information optimality)} \tag{16}$$

$$\text{(E2) } \psi^* = \lambda(\rho, \delta, \psi^*) \quad \text{(Aggregation consistency)} \tag{17}$$

The conditions (16) and (17) are the only equilibrium requirements: (i) information processing is optimal given ψ^*, (ii) the individual's best response weight equals the aggregate coefficient. The equilibrium is locally stable if $|\lambda'(\rho, \delta, \psi^*)| < 1$. Equilibrium ψ^* as function of τ^* is derived in Appendix A.3.

4.2 Theoretical predictions

This section provides detailed explanations for what the model predicts for the behavior and dynamics of (i) individual forecasts and (ii) overreaction and how it relates to the data. Overreaction is defined based on the relationship between forecast errors and the last observation, as in the empirical section. The analysis considers variations in persistence and feedback for the values set in the experiment, while also exploring the model's predictions across a broader range of these parameters.

4.2.1 Extrapolation weights

Extrapolation weights in this section are defined as the weight that the optimal individual forecast assigns to the most recent available observation, y_{t-1}, as the best response when predicting the value of y_{t+1}. For $\mu = 0$, as in the experiment, the weight, evaluated holding ψ, and thus $\alpha(\psi;\xi)$, fixed, is

$$\lambda(\rho,\delta;\psi) = \alpha(1-\rho)(1+\rho+\delta\psi) + (\rho+\delta\psi)^2 \tag{18}$$

and describes how the agent's forecast adjusts to changes in the primitives before ψ adjusts through the aggregate feedback loop.

Proposition 1. For $\delta \in \mathbb{R}$ and $\rho \in [0,1)$, fix $\psi > 0$ and let $\alpha = \alpha(\psi;\xi) \in (0,1]$. The weight $\lambda(\rho,\delta;\psi)$ defined in (18) satisfies:

i. $\lambda > 0$ for all δ.

ii. λ is strictly convex in δ, with a unique minimum at $\delta^{\min} = -\frac{2\rho+\alpha(1-\rho)}{2\psi} < 0$.

iii. With positive feedback ($\delta > 0$), λ is strictly increasing in δ, and $\lambda(\rho,\delta;\psi) > \lambda(\rho,0;\psi)$.

iv. With negative feedback ($\delta < 0$), λ is strictly increasing in δ above the threshold $\delta > \delta^{\min} = -\frac{2\rho+\alpha(1-\rho)}{2\psi}$. For $\delta < \delta^{\min}$, λ is decreasing in δ. There is a threshold $\delta_0 = -\frac{2\rho+\alpha(1-\rho)}{\psi} < \delta^{\min}$ such that $\lambda(\rho,0;\psi) > \lambda(\rho,\delta;\psi)$ for all $\delta \in (\delta_0, 0)$.

v. Both thresholds $\delta^{\min}$ and δ_0 are strictly decreasing in ρ.

vi. The marginal effect of feedback on λ is strictly increasing in ρ and α.

Proof. See Appendix A.4.

The key takeaway from the model's theoretical predictions is that the optimal individual forecast adjusts both to the presence of feedback and to the underlying persistence. Proposition 1 outlines the model's predictions regarding these variations.

Directly in relation to the experiment, two important model predictions can be validated by the data. First, $\lambda > 0$ indicates that the optimal individual forecast assigns a strictly positive weight to the last observation across all levels of persistence and feedback. Under rational expectations, if the process is fully transitory, the optimal weight is zero. Second, the weight decreases in response to negative feedback, provided that the feedback is not too strong, and this decrease is more pronounced with higher persistence.

Table F.3 shows estimates of the extrapolation weight derived from the experimental data for the four treatment groups, from

$$F_t y_{t+1} = \lambda y_{t-1} + \epsilon_t, \tag{19}$$

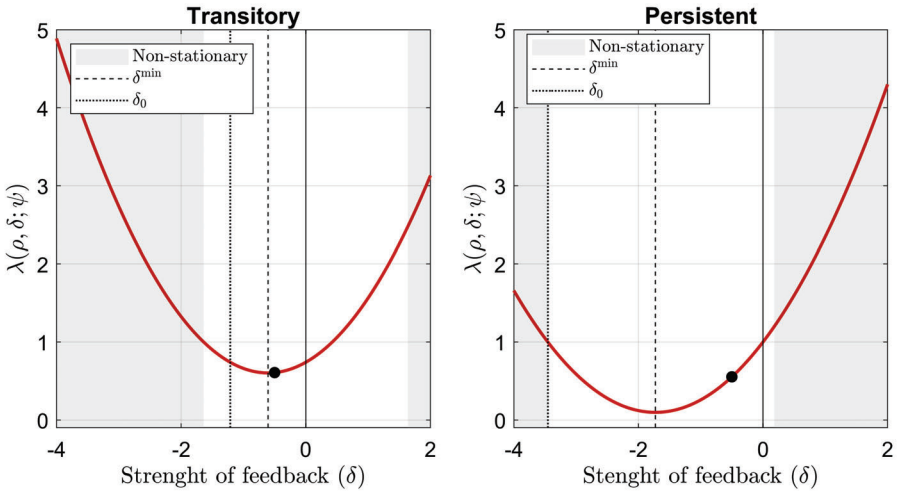
where $F_t y_{t+1} \equiv \frac{1}{N}\sum_{i=1}^{N} f_t^i y_{t+1}$ is the aggregate forecast, defined as the average of the individual forecasts within a group. The estimates demonstrate that the weight is strictly positive in all scenarios, it decreases with negative feedback, and more so when the process is persistent. These estimates are consistent with the model predictions in Proposition 1.

Beyond the data, the model provides additional predictions about how the weight λ in the optimal individual forecast changes in response to both negative and positive feedback, as well as for varying degrees of each. Generally, with respect to feedback δ, the weight is U-shaped with a unique minimum at a strictly negative $\delta^{\min}$. That means that increasing feedback, i.e., making it more positive or less negative, results in a higher weight in the optimal forecast, indicating that forecasters extrapolate more. However, making the feedback more negative reduces the extrapolation up to a certain point ($\delta^{\min}$), but once negative feedback exceeds the threshold (i.e., becomes "too negative"), the slope reverses and more negative feedback leads to more extrapolation in individual forecasts.

Figure 4 illustrates these dynamics, i.e., shows how the weight λ, as defined in equation (18), changes with the feedback parameter δ, separately for the transitory ($\rho = 0$) and

persistent ($\rho = 0.9$) environments, as in the experiment. In each plot, the black dot indicates the equilibrium $\psi^* = \lambda(\rho, \delta; \psi^*)$ for $\delta = -0.5$, while the red line depicts how the weight changes, evaluated holding ψ^* constant, with the strength and sign of the feedback. For reference, the grey shaded areas indicate values of δ, given ψ, that violate the stationarity condition $|\rho + \delta\psi| < 1$.

Figure 4: **Extrapolation weights**



Note: The figure shows the extrapolation weight $\lambda(\rho, \delta; \psi)$ that the optimal individual forecast places on the most recent observation as the strength of feedback δ varies. The left panel shows the transitory environment with $\rho = 0$, and the right panel the persistent one with $\rho = 0.9$. The black dot in each panel indicates the equilibrium $\psi^* = \lambda(\rho, \delta; \psi^*)$, and the red line shows λ changes with feedback, holding ψ^* fixed. Grey shading indicates values of δ that violate the stationarity condition $|\rho + \delta\psi| > 1$.

The dashed vertical line in the figure represents the $\delta^{\min}$ threshold in both cases, indicating the point up to which the extrapolation weight decreases with more negative feedback. Beyond this threshold, stronger negative feedback leads to sharper oscillatory dynamics between consecutive periods. Since agents forecast y_{t+1} based on y_{t-1}, these alternating dynamics occurring with strong negative feedback make the two-period-lagged observation relatively more informative for predicting y_{t+1}, which justifies a higher optimal weight on the last observed value. In an environment with high persistence, internal momentum counteracts these oscillations, meaning that a stronger negative feedback is necessary before oscillatory forces dominate, thereby shifting $\delta^{\min}$ further to the left. Consequently, in a more persistent environment, there is a greater scope for increasing negative feedback before it turns counterproductive.

However, even though strong negative feedback beyond the $\delta^{\min}$ threshold raises the extrapolation weight, the weight remains lower than it would be without any feedback, up to the second threshold, δ_0, which is marked by the dotted vertical line in Figure 4. This indicates that, up to the δ_0 point, increasing negative feedback keeps the extrapolation weight below what it would be without feedback.

Moreover, higher persistence shifts both thresholds $\delta^{\min}$ and δ_0 to the left, thereby enlarging the range of negative feedback over which the weight decreases or remains below its no-feedback level. Additionally, feedback bites more strongly when the process is persistent and when agents rely more on their prior, meaning that positive feedback increases the weight more and negative feedback decreases it more.

4.2.2 Overreaction

Overreaction is defined through the relationship between forecast errors and the last available observation of the process, y_{t-1}, as in the empirical section. Using the optimal individual forecast defined in equation (15), the corresponding forecast error is defined as:

$$y_{t+1} - f_t^{i*} y_{t+1} = -\left[\alpha(1-\rho)(1+\rho+\delta\psi)\right] y_{t-1} + (\rho+\delta\psi)\nu_t + \nu_{t+1} \tag{20}$$

where

$$\beta(\rho, \delta; \psi) \equiv \alpha(1-\rho)(1+\rho+\delta\psi) \tag{21}$$

corresponds to the overreaction coefficient defined in Section 3.1.2.

Proposition 2. For $\rho \in [0,1)$ and $\delta > -\frac{1+\rho}{\psi}$, fix $\psi > 0$ and let $\alpha = \alpha(\psi;\xi) \in (0,1]$. Then $\beta(\rho, \delta; \psi)$ as defined by (21) satisfies:

i. $\beta > 0$ and the slope $-\beta < 0$.

ii. β is strictly increasing in α and δ.

iii. β is strictly concave in ρ with a unique maximum at $\rho^{\max} = -\frac{\delta\psi}{2}$.

iv. For $\delta \geq 0$, β is strictly decreasing in ρ. For $\delta < 0$, β is strictly increasing in ρ on $[0, \rho^{\max})$ and strictly decreasing in ρ on $(\rho^{\max}, 1)$. For any $0 \leq \rho_L < \rho_H < 1$, $\beta(\rho_H) < \beta(\rho_L)$ iff $\delta > -\frac{\rho_L + \rho_H}{\psi}$.

v. The marginal effect of feedback on β is strictly increasing in α, and strictly decreasing in ρ.

Proof. See Appendix A.5.

Similar to the dynamics in the optimal individual forecast, the overreaction coefficient also changes with persistence and feedback, and Proposition 2 describes the comparative statics. The negative slope confirms that there is indeed an overreaction, which is larger at lower persistence unless the negative feedback is sufficiently strong. With respect to feedback, when negative, it attenuates overreaction, and the attenuation is larger with low persistence. These implications map directly into the three empirical facts documented in the experimental data.

Beyond the data, the model predicts that positive feedback amplifies overreaction and that the amplification is stronger when persistence is low. Moreover, both the amplification with positive and the attenuation with negative feedback are stronger with less learning (higher α). In general, when agents process less information (higher α), the slope is more negative, i.e., overreaction is stronger.

The overreaction term is strictly concave in ρ, and with positive or zero feedback, falls monotonically as persistence increases. With negative feedback, there is a single hump with a maximum $\rho^{\max}$ at low values of ρ, where overreaction initially rises slightly with persistence and declines thereafter. Stronger negative feedback pushes $\rho^{\max}$ to the higher end of the persistence range, and if sufficiently strong, it shifts such that the overreaction term monotonically *increases* with persistence for all values of ρ. However, a reversal of that nature would require feedback of large magnitude ($-\delta\psi \approx 2$), which is outside the $|\rho + \delta\psi| < 1$ stability region. If comparing two persistence levels, the overreaction at the lower one will be relatively higher for weak enough negative feedback.

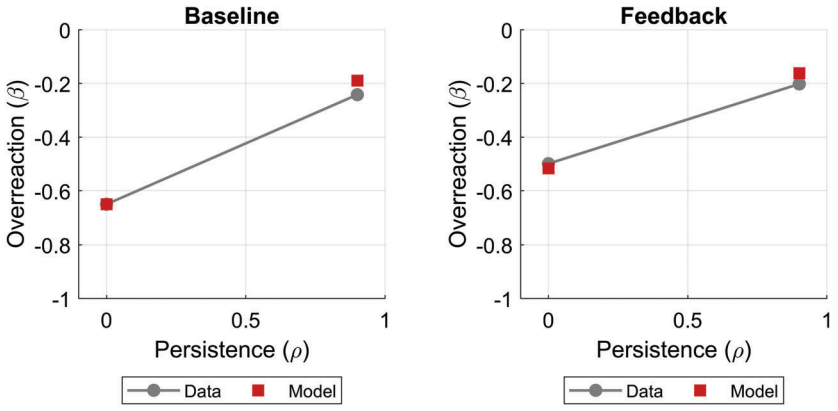
4.3 Quantitative results

The model admits a closed form solution for the corner case $\alpha = 1$ (any δ), and for $\delta = 0$ (any α). Outside the corner, the equilibrium is computed numerically. The experimental design pins down persistence, $\rho \in \{0, 0.9\}$, and feedback, $\delta \in \{-0.5, 0\}$, with $\mu = 0$. The remaining parameters are the information-cost scale (ω) and curvature (γ), and the prior

precision $\underline{\tau}$. The parameters are estimated to minimize the mean-squared error between the equilibrium overreaction in (22) and the corresponding empirical estimates in the *Baseline* condition from Table 3, and are then cross-validated in the *Feedback* case. Only the product $\theta \equiv \omega\underline{\tau}$ is identified from the targeted moments, which, with γ, means that there are two targets, and with two persistence cases in the *Baseline* condition, there are two moments. Note the implicit assumption that the cost function parameters and the prior precision do not vary with persistence or feedback, only the benefit of processing information.

The objective is flat in a small neighbourhood. The set achieving near-minimum MSE is $\theta \in \{0.03, 0.4\}$ and $\gamma \in \{0, 6.7\}$ with MSE of 0.0006. Using a tolerance of 10^{-4}, the numerical minimum is at $\gamma \simeq 4.9$ and $\theta \simeq 0.1$. Model fit is stable within the reported neighbourhood; raising γ above 2 improves the *Baseline* fit marginally. The process yields two equilibria for each (ρ, δ) pair when $\delta \neq 0$, where only one is fixed-point stable. The alternative, unstable solution is documented for completeness but not used in the quantitative fit because its implications are not validated by the experimental data.

Figure 5: **Model fit**



Note: This figure shows the model fit. Each panel reports the overreaction coefficient β plotted against persistence $\rho \in \{0, , 0.9\}$ for two environments: Baseline ($\delta = 0$) and Feedback ($\delta < 0$). Grey circles are the empirical estimates; red squares are the model-implied values using parameters estimated in the *Baseline* condition, and held fixed in Feedback for cross validation.

Figure 5 illustrates the fit of the model by plotting the overreaction coefficient β against persistence ρ for both the *Baseline* and *Feedback* conditions. The grey circles represent the empirical estimates, while the red squares indicate the model-implied values based on

the estimated parameters. The model accounts well for the cross-section of overreaction across (ρ, δ), capturing the systematic increase in β with higher persistence and its shift towards zero under negative feedback.

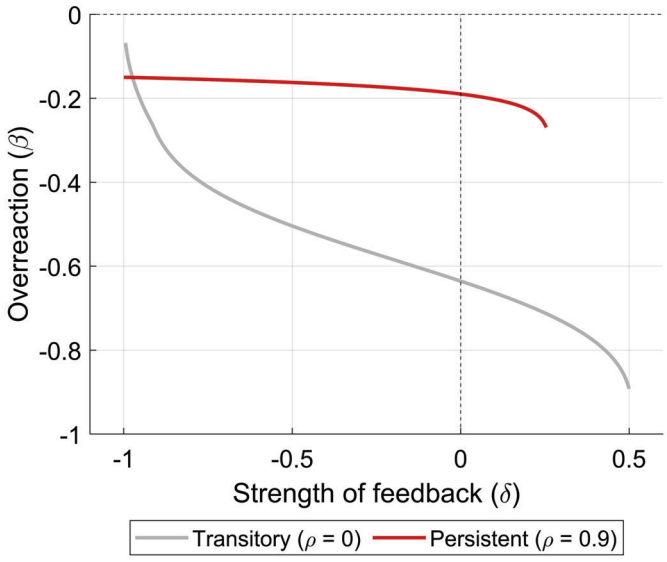
4.3.1 Equilibrium overreaction

For any (ρ, δ), the model then implies the equilibrium overreaction coefficient

$$\beta^*(\rho, \delta, \psi^*) = \alpha(\psi^*; \xi)(1 - \rho)(1 + \rho + \delta\psi^*). \tag{22}$$

where ψ^* is the solution to the fixed-point problem and $\alpha(\cdot)$ is the induced information choice. The empirical counterpart is the slope in the regression of the forecast errors on y_{t-1}, which corresponds to $-\beta^*(\rho, \delta, \psi^*)$. Figure 6 plots equilibrium overreaction against the strength of feedback δ, holding ρ fixed at the two experimental values, $\rho = 0$ and $\rho = 0.9$. The equilibrium is recomputed for each δ on a fine grid.

Figure 6: **Overreaction curves**



Note: This figure shows equilibrium overreaction β as a function of feedback strength δ. The grey curve shows the transitory case with $\rho = 0$, and the red curve the persistent case with $\rho = 0.9$. For each δ in the displayed range, the equilibrium is recomputed on a fine grid with model parameters fixed.

A few features emerge from the model's equilibrium. First, the degree of overreaction is higher for transitory processes for most values of feedback. Second, the slope is

decreasing in δ. More positive feedback makes the slope more negative, amplifying the degree of overreaction, while more negative feedback makes it less negative, attenuating the overreaction. Third, the curve is much steeper in the transitory case than in the persistent. The absolute sensitivity of overreaction to feedback is larger when persistence is lower. In the figure, the grey line of the transitory case falls sharply with δ, while in the persistent case it declines modestly. These patters mirror the facts documented in the data and also shown in Figure 5.

As $\alpha(\psi)$ decreases with $(1 + \rho + \delta\psi)$, positive feedback induces more information processing (lower α), which partially offsets amplification. Conversely, negative feedback raises α and partly offsets attenuation. This endogenous adjustment generates the mild curvature visible in both lines; the transitory curve is markedly nonlinear, while the persistent curve is comparatively flat.

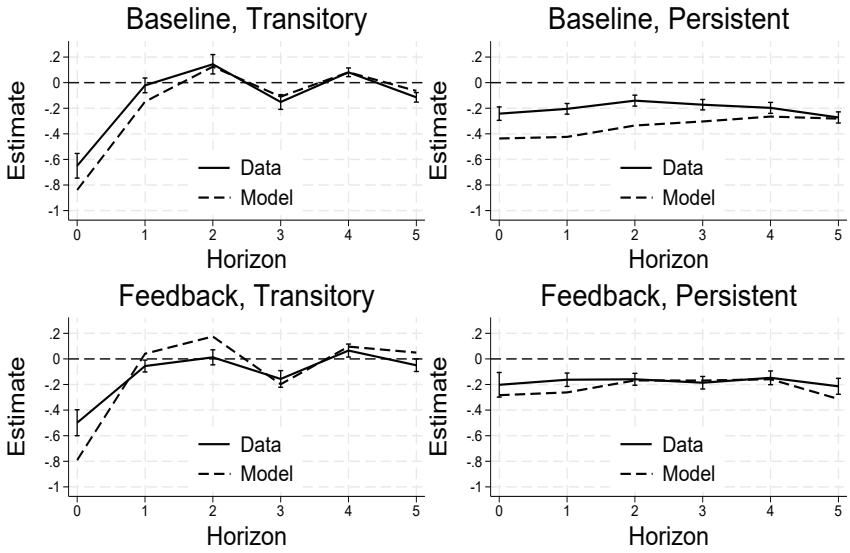
4.4 Additional model validation

To further validate the model, this section compares the empirical estimates presented in sections 3.1.4 and 3.2.1 with the corresponding estimates derived from the simulated data of the model. In the appendix, Figure E.6 illustrates the average forecasts across all treatment groups in the experiment alongside the average forecasts from the model-simulated data, and Figure E.7 shows the comparisons of the impulse response functions of forecasts to the exogenous shocks between the experimental and model-simulated data.

4.4.1 Overreaction over time

Section 3.1.4 of the empirical part examines whether the documented overreaction in forecasting behavior persists over subsequent periods. Specifically, it explores how long the overreaction to y_{t-1} is present after period t and whether the difference in duration relates to feedback and persistence. The overreaction over time is estimated as in equation (4), where β^h represents co-movement between forecast errors, $(y_{t+h} - f_t^i y_{t+h})$, h periods ahead to the same past realization y_{t-1}. Figure 7 displays the same estimates as in Figure 1, matched with the estimates of the same coefficients, but on model-simulated data. The model's data is simulated using the same sequence of shocks as in the experiment.

Figure 7: **Overreaction over time: Data and Model**



Note: The figure plots estimates of β^h from horizon-h regressions of forecast errors, $y_{t+h} - f_t^i y_{t+h}$, on the last observed value y_{t-1}, separately for each treatment, and compares them to similar estimates using model-simulated data. In Baseline, only $f_t^i y_{t+h}$ are simulated, while y_{t+h} remains unchanged. In Feedback, both $f_t^i y_{t+h}$ and y_{t+h} as simulated. All series from the model are simulated using the same sequence of shocks that participants faced in the lab. The left (right) panels show the Transitory (Persistent) treatment group; the top (bottom) panels show the Baseline (Feedback) groups. Solid lines represent estimates in the experimental data, and dashed lines represent responses in the model-simulated data. Points are coefficient estimates; vertical bars in the experimental data estimates are 95% confidence intervals. Standard errors are clustered at the individual level.

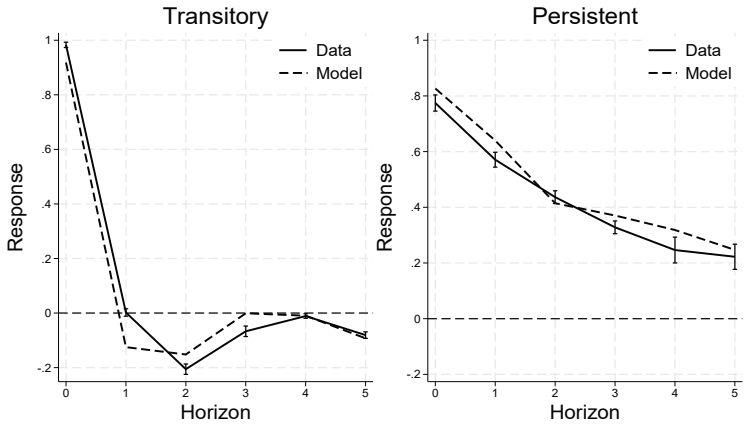
In general, the model's predictions follow the original empirical estimates fairly well. The initial overreaction is more substantial in the transitory case, but it decays quickly. In the persistent case, it is initially significantly smaller but remains at the same level throughout the observed horizon. The overreaction measured through the model-simulated data slightly overestimates it on impact in the transitory case, both in the *Baseline* and *Feedback*, and for most periods in the persistent *Baseline*. At the same time, it matches the empirical estimates in the persistent *Feedback* case almost perfectly.

4.4.2 Dynamic response to shocks

Section 3.2.1 measures the dynamic responses of outcomes to exogenous shocks across all treatment groups, to understand how shocks propagate through the process and how the propagation varies with feedback and persistence. It is estimated as in equation (5), where the estimated coefficient γ_{1h} measures the response of y_{t+h} to a one-unit shock in

ϵ_t. Figure 8 compares the estimates of γ_{1h} shown in Figure 2 to the same coefficients, but estimated on model-simulated data. As in the previous section, the model series are estimated using the same sequence of shocks as in the experiment.

Figure 8: **Response of y_{t+h} to ϵ_t: Data and model**



Note: The figure plots estimates of γ_h from horizon-h local-projection impulse response regressions, as specified in (5), run separately by treatment group, and compares them to similar estimates using model-simulated data. It shows estimates only for the Feedback condition; there are no differences in Baseline by definition. The left (right) panel shows the Transitory (Persistent) group; solid lines represent responses in the experimental data and dashed lines responses in the model simulated data. All series from the model are simulated using the same sequence of shocks that participants faced in the lab. Points show the estimated response of y_{t+h} to a one-unit innovation in ϵ. Vertical bars in the experimental data estimates are 95% confidence intervals.

The experimental data show that in the transitory case, both with and without feedback, the initial impact is larger when persistent, but it decays quickly and vanishes after the first period. The response in *Feedback* is below the response in the *Baseline*, showing that negative feedback reduces the pass-through of shocks. In the persistent case, the response is smaller on impact, declines only gradually, and negative feedback shifts the entire path down significantly more than in the transitory case, implying a smaller effect of shocks relative to the environment without feedback. Figure 2 shows that the same response of outcomes estimated on model-simulated data matches the empirical estimates very well, and follows the whole path closely over the entire horizon. Overall, all four validation exercises strongly support the model by showing that it matches fairly well both the point estimates and general dynamics documented in the empirical results.

4.5 Comparison with alternative forecasting models

To further support the forecasting model discussed so far, this section compares its predictions and how well they match the empirical facts to those of alternative expectation formation models. First, it examines a counterfactual scenario where agents forecast the future as if there were no feedback in the environment. In this case, the feedback present in the environment would push the bias in the opposite direction from what we observe in the data. This pattern suggests that the attenuation in overreaction can only occur if agents modify their forecasting behavior in response to negative feedback.

Second, this section compares the forecasting model to three common alternatives that typically predict an overreaction in expectations: adaptive, extrapolative, and diagnostic expectations. Similar to the first case, these alternative models cannot match all empirical facts simultaneously.

4.5.1 Overreaction with a naive forecasting rule

Endogenous feedback affects outcomes through two channels: (i) it alters the law of motion via the effective AR(1) coefficient $\rho + \delta\psi$, and (ii) it changes optimal forecast through the information choice $\alpha(\cdot)$ and the induced weight on y_{t-1}. To separate these forces, we can consider a counterfactual in which forecasters ignore feedback and apply the same forecasting rule as in the Baseline ($\delta = 0$) while the process itself is subject to feedback ($\delta \neq 0$). This isolates the mechanical effect of feedback from the behavioral adjustment in the rule.

Optimal individual forecast in the *Baseline* condition is

$$f_t^{NR} y_{t+1} = (1-\rho^2)[(1-\alpha)\mu + \alpha y_{t-1}] + \rho^2 y_{t-1} \tag{23}$$

which in the feedback environment defines the forecast error as

$$y_{t+1} - f_t^{NR} y_{t+1} = -\left[(\rho^2 + (1-\rho^2)\alpha) - (\rho+\delta\psi)^2\right] y_{t-1} + (\rho+\delta\psi)\nu_t + \nu_{t+1} \tag{24}$$

with the corresponding overreaction coefficient defined by

$$\beta^{NR}(\rho, \delta; \psi) \equiv (\rho^2 + (1-\rho^2)\alpha) - (\rho+\delta\psi)^2. \tag{25}$$

Proposition 3. For $\rho \in [0,1)$ and $\delta \in \mathbb{R}$, fix $\psi > 0$ and let $\alpha = \alpha(\psi;\xi) \in (0,1]$. Then $\beta^{NR}(\rho,\delta;\psi)$ as defined by (25) satisfies:

i. At $\delta = 0$, $\beta^{NR} = \alpha(1-\rho^2)$ which is consistent with *Baseline* overreaction.

ii. β^{NR} is concave in δ with a unique maximum at $\delta^{\max} = -\frac{\rho}{\psi} < 0$.

iii. With positive feedback ($\delta > 0$), β^{NR} is strictly decreasing in δ, and $\beta^{NR}(\rho,\delta;\psi) < \beta^{NR}(\rho,0;\psi)$

iv. With negative feedback ($\delta < 0$), for $\delta > \delta^{\max}$, β^{NR} is strictly decreasing in δ, and $\beta^{NR}(\rho,\delta;\psi) > \beta^{NR}(\rho,0;\psi)$. For $\delta < \delta^{max}$, β^{NR} is increasing in δ. For $\rho > 0$, there is a threshold $\delta_0 = -\frac{2\rho}{\psi} < \delta^{\max} < 0$ such that $\beta^{NR}(\rho,\delta;\psi) > \beta^{NR}(\rho,0;\psi)$ holds for $\delta \in (\delta_0, 0)$. For $\rho = 0$, $\delta^{max} = \delta_0 = 0$.

v. Both $\delta^{\max}$ and δ_0 are decreasing in ρ.

Proof. See Appendix A.6.

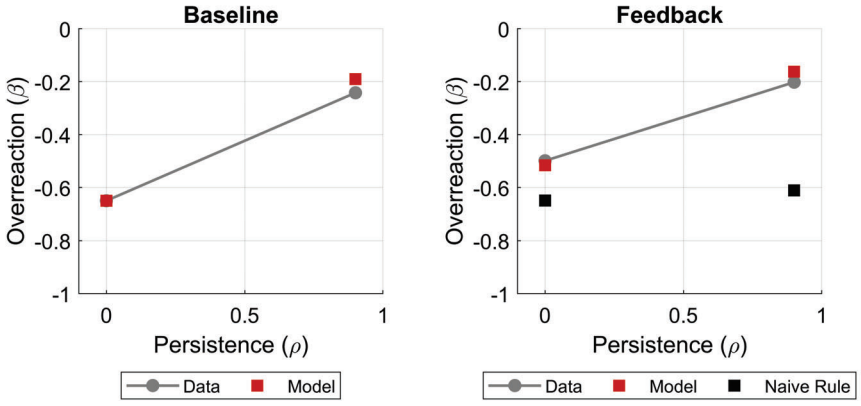
The key takeaway from the comparative statics of the overreaction coefficient generated with the naive forecasting rule, as described by Proposition 3, is that changes in overreaction with negative feedback move in the exactly the opposite direction from what is observed in the data and predicted by the original forecasting model. If agents do not take the effects of feedback into account, their forecasts will overshoot relatively more than in the Baseline, instead of less, as observed in the data. Moreover, there is little no variation in the overreaction with persistence, which is also not in line with the data.

With the naive rule, there is no behavioral adjustment to the feedback, the information weight α and the forecast's dependence on y_{t-1} are the same as in the *Baseline*, while the data-generating process changes with feedback. When feedback is mildly negative, $\delta \in (\delta^{\max}, 0)$, making feedback more negative increases overreaction. Only after feedback crosses the turning point, does further negativity reduce β^{NR}. Relative to the no-feedback benchmark, the naive rule predicts the overreaction above the Baseline on $\delta \in (\delta_0, 0)$, and below the Baseline once $\delta < \delta_0$. In the transitory case, both thresholds collapse to zero, so any $\delta < 0$ lies to the left of both. Overreaction is then strictly below its Baseline level and declines monotonically with more negative feedback. In this case, the naive rule explains the decrease in overreaction with negative feedback that we observe in the data, but only for the fully transitory case. Both thresholds get more negative as persistence increases,

so the decrease in overreaction that we observe in the data for the highly persistent case cannot be explained by the model with the naive forecasting rule. Positive feedback has the opposite implication under the naive rule. As feedback becomes more positive, β^{NR} falls and the overreaction is everywhere below the Baseline value. The attenuation under positive and amplification under negative feedback are the consequences from the feedback altering the law of motion.

Negative feedback dampens the dynamics; for fixed weights in the forecast, realized outcomes are mechanically pulled toward zero, and more so when persistence is high, while forecasts do not adjust. When feedback is not too negative, the mismatch makes the forecast error conditionally larger and makes the slope more negative, implying a stronger overreaction relative to Baseline. Under positive feedback, the same mechanics work in reverse as outcomes are pushed away from zero while forecasts stay anchored at the Baseline. These predictions contrast with the patterns in the data, where negative feedback attenuates the overreaction, suggesting that forecasters do adjust their rules when feedback is present.

Figure 9: **Model fit with the naive rule**



Note: This figure shows the model fit for the naive-rule counterfactual. Each panel reports the overreaction coefficient β plotted against persistence $\rho \in \{0, , 0.9\}$ for two environments: *Baseline* ($\delta = 0$) and *Feedback* ($\delta < 0$). Grey circles are the empirical estimates; red squares are the model-implied values using parameters estimated in the *Baseline* condition, and held fixed in *Feedback* for cross validation. Black squares report the naive-rule counterfactual obtained by applying the *Baseline* forecasting rule while the process is subject to feedback; in *Baseline*, it coincides with the model and overlays the red marker.

Figure 9 shows the model fit after incorporating the naive rule counterfactual in the *Feedback* condition, using the same parameters. In the *Baseline* panel, the naive rule aligns

with the model by construction. According to the model's predictions, the naive rule with negative feedback and high persistence yields a coefficient that is more negative than both the base model and the observed data, as well as compared to the *Baseline* condition. In the persistent case, negative feedback interacts with high ρ to pull realizations further below the *Baseline* path, while forecasts remain unchanged. In the transitory case, there is a barely noticeable reduction in overreaction with negative feedback, relative to the *Baseline*. This reduction is not substantial enough to account for the observed empirical variation.

In the experiment, participants are aware of the presence and sign of feedback. The fact that overreaction in the data is smaller with negative feedback indicates that forecasters incorporate it into their forecasts. Absent such internalization, the overreaction would be equal or larger, especially with high persistence. The evidence suggests that participants recognize the sign of the negative feedback even if precise parameter values are uncertain, and they appear to infer a persistence above the true value, consistent with the overreaction bias.

4.5.2 Overreaction with other expectations models

Among the alternatives to rational expectations, the three most common ones predict an overreaction to recent information: adaptive, extrapolative, and diagnostic expectations. Adaptive and extrapolative expectations refer to forecasting rules that rely solely on past data, without considering the underlying characteristics of the environment, such as persistence or feedback, and are, in nature, similar to the naive rule discussed above. Under diagnostic expectations, in contrast, forecasts adapt to the environment and are, by nature, relatively close to rational but assume that agents overreact to information that is already informative about the future. This section shows that neither of the options can simultaneously account for all four empirical facts observed in the experiment. Models in the family of sticky and noisy expectations (e.g., Mankiw and Reis (2002) or Woodford (2003)) or cognitive discounting (Gabaix, 2014) have been shown to predict an underreaction to recent information, and as such will not be considered here.

Adaptive expectations. As the historically most used alternative to rational ex-

pectations, they have been introduced as early as the 1950s (Cagan, 1956). They are standardly defined as

$$f_t^{i,A} y_{t+1} = \lambda y_{t-1} + (1-\lambda) f_{t-1}^i y_t \tag{26}$$

where the forecast for the following period is the weighted average of the last observation and the last period's forecast.

Extrapolative expectations. Often used in finance to explain persistent overreaction to recent shocks (Hong and Stein, 1999; Barberis and Shleifer, 2003), they can be defined as:

$$f_t^{i,E} y_{t+1} = y_{t-1} + \phi(y_{t-1} - y_{t-2}) \tag{27}$$

where the forecast depends on the last observation and extrapolates from the recent trend at the degree of ϕ.

Diagnostic expectations. Introduced by Bordalo et al. (2018) and based on the representativeness heuristic of Kahneman and Tversky (1972), it contains rational expectations, implying that forecasters are forward-looking and incorporate the characteristics of the environment into their forecasts. Standardly, it is defined as:

$$f_t^{i,D} y_{t+1} = \mathbb{E}_t\, y_{t+1} + \theta(\mathbb{E}_t\, y_{t+1} - \mathbb{E}_{t-1}\, y_{t+1}) \tag{28}$$

where θ measures the degree of overreaction to the most recent surprise, i.e., the difference in rational expectations between the last two periods.

Table 5: **Alternative models parameter estimation**

Model	**Estimate**
Adaptive (λ)	0.69
	(0.01)
Extrapolative (ϕ)	-0.23
	(0.01)
Diagnostic (θ)	0.20
	(0.03)

Note: The table shows parameter estimates for the three alternative expectation models. Parameters are estimated by constrained least squares on pooled forecasting data from all four treatments. Standard errors, in parentheses, are clustered at the individual level.

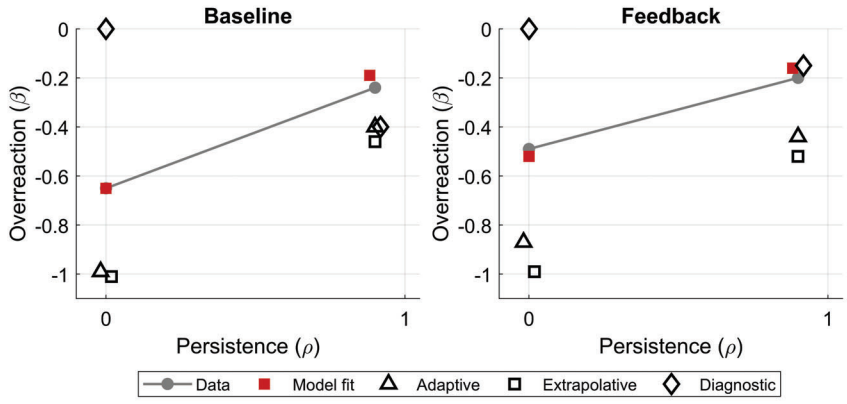
To test whether alternative models can match the empirical patterns observed in the experiment, we can use the experimental forecasting data to estimate the parameters of the three models as defined by equations (26) to (28). Then, using the fitted values from these estimations, we can recompute overreaction as the correlation between forecast errors and the last observation, y_{t-1}, in all three cases, and for the four processes varying in persistence and feedback.

As the first step, the three models are estimated on pooled data from all four treatments, i.e., for both values of persistence and feedback, using constrained least squares. The estimates are shown in Table 5. All three models confirm an overreaction. Both adaptive and extrapolative expectations estimates show a high weight on the last observation; 0.69 and 0.77 respectively. The estimate of $\theta = 0.2$ in the diagnostic model means that forecasts react 20% too much to the latest news. Then, using the fitted values from these estimations, we can reestimate equation (3) as:

$$y_{t+1} - f_t^{i,j} y_{t+1} = \alpha_i + \beta^j y_{t-1} + \nu_{it} \tag{29}$$

for $j = \{A, E, D\}$ corresponding to the three expectations models. Figure 10 compares the estimates with the original ones and the model predictions, for all four treatment groups.

Figure 10: **Model fit with other expectations models**



Note: This figure shows the model fit alongside alternative expectations models. Each panel reports the overreaction coefficient β plotted against persistence $\rho \in \{0, , 0.9\}$ for two environments: *Baseline* ($\delta = 0$) and *Feedback* ($\delta < 0$). Grey circles are the empirical estimates; red squares are the model-implied values using parameters estimated in the *Baseline* condition, and held fixed in *Feedback* for cross validation. Open markers plot the alternative models using parameters fitted on the pooled data: triangles for Adaptive (λ), squares for Extrapolative (ϕ), and diamonds for Diagnostic (θ), with β re-estimated from each rule's fitted forecasts.

All three expectation models predict more or less the same level of overreaction in the persistent *Baseline* case, which is in line with previous findings. Diagnostic expectations, both in nature and in the estimates, are closest to the model here, which is evident in the persistent case. However, they fail in the fully transitory case. By definition, diagnostic expectations predict zero overreaction to news that is truly uninformative, as is the case in the fully transitory environment, which does not match the overreaction documented in the data.

Adaptive and extrapolative models consistently predict overreaction that is stronger than what is observed in the data. What they get right is the decrease with persistence, but they do not match the changes with feedback. Similarly to the predictions of the model with the naive rule, they predict a slight decrease in overreaction with negative feedback for the transitory process, but a significant increase in the persistent case.

5 Application: New Keynesian model

In the standard New Keynesian (NK) environment (e.g., Galí (2015)), a monetary policy that satisfies the Taylor principle generates negative general-equilibrium feedback from inflation expectations to the output gap. The intuition is as follows: an increase in inflation expectations, holding the nominal rate fixed, results in a lower ex-ante real interest rate, which brings spending forward and increases aggregate demand, marginal costs, and thus inflation. A central bank that targets inflation will increase nominal interest rates more than one-to-one with current inflation, which increases the real interest rate, relieves pressure on demand, and reduces the output gap, thereby stabilizing inflation as well. When the Taylor principle holds, an increase in inflation expectations will lead to a decrease in the output gap, and the strength of this negative feedback increases with the aggressiveness of the policy response.

That same negative feedback also affects firms' strategic behavior. Cornand and Heinemann (2022) show that monetary policy that responds to inflation and satisfies the Taylor principle turns prices into strategic substitutes. Absent general-equilibrium reasoning, a firm that expects others to raise prices has an incentive to do the same. However, with model-consistent expectations, firms internalize general equilibrium effects, anticipate the

induced demand contraction, and have an incentive to reduce prices instead.

Under rational expectations, there is no strategic uncertainty; firms and households internalize the policy effects and immediately coordinate on the unique equilibrium following a shock. Convergence is then not a learning process but an immediate alignment of beliefs with the unique equilibrium. Conversely, if agents are boundedly rational or information is noisy, they may over-extrapolate from recent inflation or fail to internalize policy feedback adequately. Then there is a transition path where best responses iterate towards an equilibrium rather than achieving it instantaneously. Cornand and Heinemann (2022) argue that by influencing the degree of strategic complementarity, the central bank can shape the adjustment path towards equilibrium.

Experimental literature (Bao et al., 2021) shows substantial evidence that agents can reach the rational expectations equilibrium even with biased expectations, provided there is adequate negative feedback in the system. Even if agents know very little about their environment and form entirely backward-looking expectations, negative feedback generates a self-correcting mechanism that facilitates convergence along the learning path by correcting deviations of expectations from actual outcomes. With negative feedback, higher expectations lead to lower outcomes. This oscillatory behavior has been shown to enable quicker convergence towards equilibrium compared to scenarios where feedback is positive.

However, the way agents form expectations and how much their beliefs extrapolate from the past also matters for how forcefully the general-equilibrium stabilizer operates. The experiment isolates precisely this mechanism. It removes the complexity of the New Keynesian model by presenting participants with a reduced-form process that includes a feedback term, capturing the general equilibrium effects between expectations and the model's outcomes. The objective is not to determine whether forecasters can fully track the interactions between inflation, output, and the interest rate, but rather to assess whether they modify their forecasts in response to feedback, if they are aware of its presence and sign.

Data from the experiment and the forecasting model indicate that individuals adjust their forecasting behavior to match the characteristics of their environment. When faced with negative feedback, they tend to extrapolate less from past values compared to situa-

tions with no feedback, which amplifies its stabilizing effects. Although negative feedback stabilizes even without behavioral adjustments, such adjustments significantly influence the speed of convergence toward the rational expectations equilibrium. The following sections demonstrate the logic of this mechanism through a simplified version of the standard New Keynesian model.

5.1 Model

The standard New Keynesian model (e.g., Galí (2015), Walsh (2010) or Woodford (2003)) can be extended away from rational expectations to allow for incorporating other expectation formation models (see e.g., Branch and McGough (2009)). Consider the standard three equations:

$$x_t = F_t x_{t+1} - \sigma(i_t - F_t \pi_{t+1}) + g_t \tag{30}$$

$$\pi_t = \kappa x_t + \beta F_t \pi_{t+1} \tag{31}$$

$$i_t = \phi_\pi \pi_t \tag{32}$$

where (30) is the dynamic IS curve, (31) is the New Keynesian Phillips curve and (32) is the Taylor-type policy rule targeting inflation. x_t denotes the output gap, π_t inflation, both in deviations from the steady state, i_t is the nominal interest rate and g_t is an i.i.d. exogenous shock.

Away from the rational expectations benchmark, we can assume expectations to be backward looking, defined as

$$F_t x_{t+1} = \psi_x x_{t-1} \tag{33}$$

$$F_t \pi_{t+1} = \psi_\pi \pi_{t-1} \tag{34}$$

and substitute (33), (32) and (31) into (30) to obtain:

$$x_t = \frac{\psi_x}{1 + \sigma\phi_\pi\kappa} x_{t-1} + \frac{\sigma(1 - \phi_\pi\beta)}{1 + \sigma\phi_\pi\kappa} F_t \pi_{t+1} + \frac{1}{1 + \sigma\phi_\pi\kappa} g_t \tag{35}$$

which makes (35) resemble the process in the experiment. Under standard calibration

(Clarida et al., 2000), $\sigma = 1$ is the intertemporal elasticity of substitution, $\kappa = 0.3$ is the slope of the Phillips curve, $\beta = 0.99$ is the discount factor and $g_t \sim N(0, \sigma_g)$ with $\sigma_g = 2$ as in the experiment. ϕ_π measures the response of nominal interest rate to deviations in inflation from the steady state.

Parameters ψ_x and ψ_π in (33-34) define the weight agents put on the last observation when forming output and inflation expectations, respectively. Defining $\rho \equiv \frac{\psi_x}{1+\sigma\phi_\pi\kappa}$ and $\delta \equiv \frac{\sigma(1-\phi_\pi\beta)}{1+\sigma\phi_\pi\kappa}$, we can rewrite (35) as

$$x_t = \rho x_{t-1} + \delta F_t \pi_{t+1} + \rho g_t \tag{36}$$

and calibrate $\psi \equiv \psi_x = \psi_\pi = f(\rho(\psi), \delta)$ externally from the forecasting model in the previous section. The forecasting model delivers a prediction for the weight the aggregate forecast puts on the last observation for a given degree of persistence (ρ) and feedback (δ). For the given base parameters, $\rho > 0$ and $\delta < 0$ if the Taylor principle holds ($\phi_\pi > 1$).

The key difference with respect to the standard backward-looking models, such as adaptive or extrapolative expectations, is the endogeneity of the extrapolation parameter (ψ) to the persistence and feedback in reduced-form version of the model. If the primitives of the model change, e.g., the slope of the Phillips curve (κ) or the Taylor rule coefficient (ϕ_π), the persistence and feedback change as well, and imply a different degree of extrapolation from the past. For example, holding all else constant, more aggressive monetary policy (higher ϕ_π) makes the feedback parameter more negative, and persistence lower, which implies relatively less extrapolation. In contrast, a flatter Phillips curve (lower κ), makes the feedback more negative as well but steeply increases the persistence, which makes a decrease in extrapolation relatively smaller.

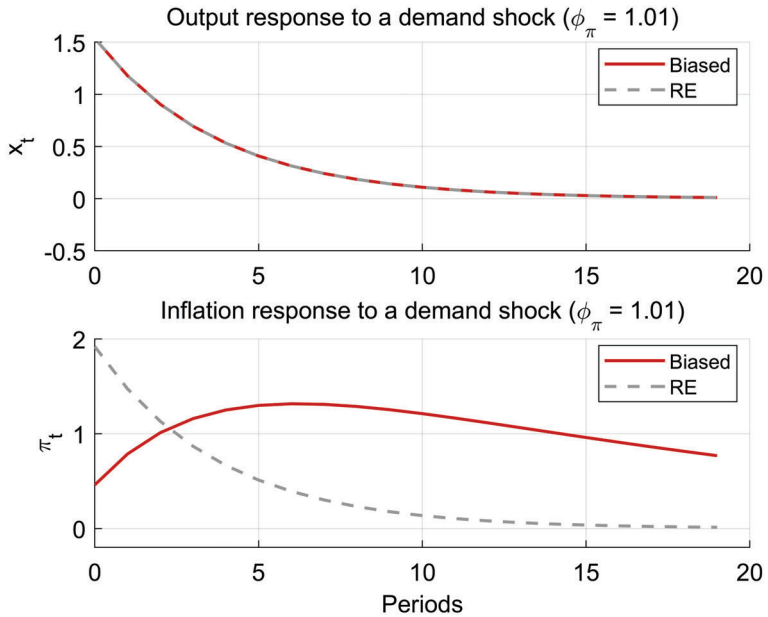
To isolate the most relevant mechanism in the context of this paper, explained through three different cases varying the strength of monetary policy responses to inflation, we can set $\psi_x = 1$ and focus on inflation expectations. The rational expectations benchmark used for comparisons retains the same assumption, with only inflation expectations formed rationally, to ensure symmetry in the general degree of persistence between the models. Appendix B shows a similar analysis where this assumption is relaxed, imposing $\psi_x = \psi_\pi$ instead, and allowing both inflation and output gap expectations to be formed

rationally in the benchmark. The alternative approach confirms that the main conclusions remain valid. The following sections demonstrate how the model dynamics change with varying levels of policy feedback, to highlight the role of adjustments in extrapolation in determining the speed of convergence to the rational expectations equilibrium.

5.1.1 Case 1: $\phi_\pi = 1.01$

Taylor rule coefficient that satisfies the Taylor principle but is only slightly above 1 ($\phi_\pi = 1.01$) translates to zero feedback ($\delta = 0$) and high persistence ($\rho = 0.76$), implying the extrapolation at the level of $\psi_\pi = 0.96$. Figure 11 shows the impulse response functions of output gap and inflation to a demand shock in that scenario, and for comparison the response in the rational expectation benchmark case.

Figure 11: **IRFs with zero feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.01$, over 20 periods. Solid red lines depict the biased-expectations model with $\psi_x = 1$ and $\psi_\pi = 0.96$ extrapolation weights, while the grey dashed line represents the response in the rational benchmark with $F_t x_{t+1} = 1$, and only inflation expectations formed rationally.

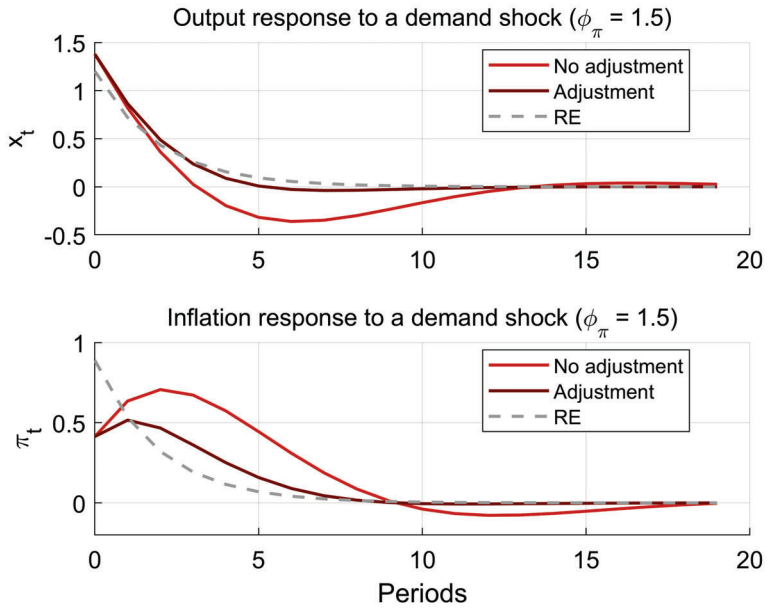
With $\delta = 0$ and effectively no feedback from inflation expectations, the output gap response is identical in both cases. In contrast, the response of inflation with biased expectations is smaller initially, but peaks with a lag and is significantly more persistent,

with no convergence to the rational expectations equilibrium even after 20 periods. Under rational expectations, a non-negative δ is sufficient for convergence, even if delayed, which is not the case in the alternative scenario.

5.1.2 Case 2: $\phi_\pi = 1.5$

A more aggressive monetary policy characterized by $\phi_\pi = 1.5$ leads to negative feedback, with a value of $\delta = -0.33$, and slightly lower persistence at $\rho = 0.69$. Both the decrease in persistence and the more negative feedback reduce the extent of extrapolation in expectations, resulting in $\psi_\pi = 0.63$.

Figure 12: **IRFs with negative feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.5$, over 20 periods. The grey dashed line represents the response in the rational benchmark with $F_t x_{t+1} = 1$, and only inflation expectations formed rationally. The solid red line ('No adjustment' depicts the biased-expectations model with $\psi_x = 1$ and $\psi_\pi = 0.96$ held fixed; the dark red line ('Adjustment') depicts the biased-expectations model with $\psi_x = 1$ and $\psi_\pi = 0.63$.

Figure 12 illustrates the same impulse response functions of the output gap and inflation to a demand shock as in Case 1. The grey dashed line represents the responses in the rational benchmark, while the other two lines depict the responses in the model with biased expectations. The 'Adjustment' line reflects the response with $\psi_\pi = 0.63$ in the

inflation expectations forecasting rule, assuming that the agents' behavior changes with the primitives of the model. In contrast, the 'No adjustment' line shows the responses of an agent if they do not alter their behavior and use the same weight in their forecasts as if δ remains close to zero.

A stronger policy response that generates more negative feedback reduces inflation inertia, regardless of the adjustment in expectations. Relative to the weak-policy case, a more aggressive rule itself reduces persistence and accelerates convergence. However, if the tendency to extrapolate remains high, negative feedback can lead to oscillations around the rational expectations path. When agents overweight recent observations, their forecasts can push the system too far in one direction, prompting a corrective move in the opposite direction next period. When forecasting behavior adjusts to the strength of policy, agents extrapolate less, which dampens these oscillations. As a result, both output and inflation responses align more closely with the rational expectations paths. However, even with the adjustment, some overreaction to the past observation remains, causing inflation to stay above the rational convergence path for a while.

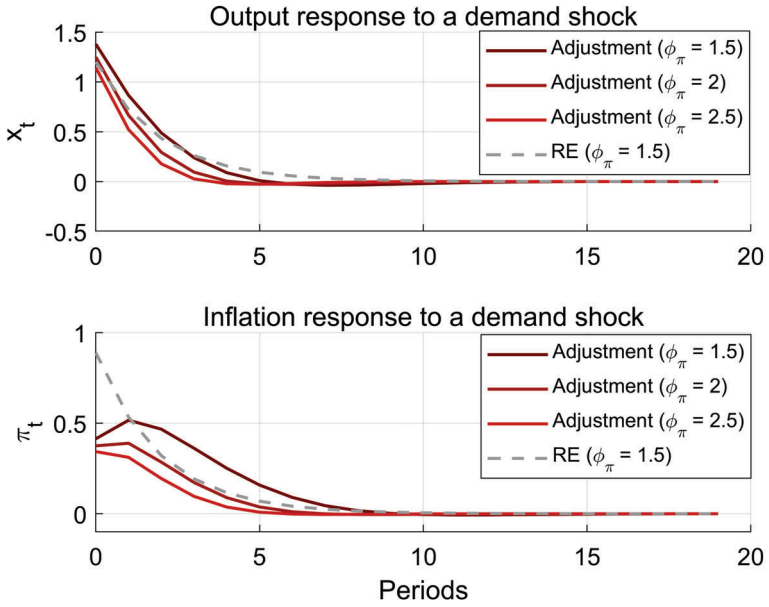
On impact, the output gap reacts more strongly with biased expectations compared to the rational benchmark, because backward-looking beliefs do not account for the future effects of shocks as effectively as forward-looking expectations do, resulting in a stronger initial effect. Over time, it falls slightly below the rational expectations path. With a strong enough monetary policy, an increase in the nominal rate will lead to a higher real rate, thereby lowering the aggregate demand. With relatively more persistence in inflation expectations, the rise in the nominal interest rate translates to a relatively larger increase in the real rate, which reduces demand more. The stronger the tendency to extrapolate from the past data, the more pronounced this effect becomes, which is why the the output response in the 'No adjustment' case dips further below the rational path.

5.1.3 Case 3: $\phi_\pi^{RE} = 1.5$ and $\phi_\pi^B = \{1.5, 2, 2.5\}$

Even with an adjustment in expectation, some inertia in inflation remains, suggesting that a stronger monetary policy reaction is necessary to achieve the same convergence path as in the rational expectations benchmark. Figure 13 compares impulse response

functions of output and inflation in the rational benchmark under the standard Taylor rule coefficient ($\phi_\pi^{RE} = 1.5$) with responses with biased expectations under $\phi_\pi^B = \{1.5, 2, 2.5\}$. Increasing ϕ_π^B implies a stronger negative feedback (higher δ), reduced persistence ρ and thus less extrapolation at $\psi_\pi = 0.51$ and $\psi_\pi = 0.47$ with $\phi_\pi = 2$ and $\phi_\pi = 2.5$, respectively.

Figure 13: **IRFs with negative feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.5$, over 20 periods. The grey dashed line represents the response in the rational benchmark with $F_t x_{t+1} = 1$, and only inflation expectations formed rationally. The solid red lines ("Adjustment") depict the biased-expectations model with $\psi_x = 1$ for $\phi_\pi \in 1.5, 2, 2.5$, using the corresponding inflation-expectations weights $\psi_\pi \in 0.63, 0.51, 0.47$, respectively.

The figure shows that with the Taylor coefficient of $\phi_\pi = 2$ the response of inflation is weaker on impact and approximately follows the path of the rational benchmark, with a faster convergence in the output gap. Increasing the strength of monetary policy reaction to inflation any further speeds up convergence of both variables even more, but the additional effect is relatively smaller than in the initial increase.

In summary, greater inertia in expectations slows down the adjustment of inflation and intensifies the effect of the real interest rate, resulting in a faster decrease in output as monetary policy becomes more aggressive. If individuals extrapolate less when aware of the policy feedback, they dampen these effects, enabling inflation to adjust more quickly. These patters suggest that a stronger monetary policy response is required than implied

by rational expectations, although not as much as suggested by fixed rules, such as adaptive or extrapolative expectations. Awareness of the feedback mechanism can therefore stabilize the system even if agents remain backward-looking and do not process the effects of shocks in the same way a rational agent would. This highlights the role and importance of credibility and effective communication in central banking.

6 Conclusion

This paper investigates the biases in expectations that emerge within environments characterized by feedback loops between agents' expectations and economic outcomes. Through a forecasting experiment, the empirical evidence highlights a systematic overreaction to recent information among participants, which is significantly mitigated by the presence of negative feedback.

A forecasting model extends the framework of Afrouzi et al. (2023) to allow for feedback. Agents form beliefs about a long-run feature of the process they are predicting that extrapolate from the most recent observation, and face costly information processing in updating their beliefs to improve the forecast accuracy. The model suggests that the decrease in overreaction is contingent upon agents incorporating feedback into their forecasting behavior. Embedding the documented bias and its variation with feedback into a New Keynesian framework shows that overreaction generates inertia in expectations, causing excess persistence in the response of inflation to an exogenous shock, necessitating a stronger monetary policy reaction compared to rational expectations. However, if agents recognize the presence of feedback and reduce the degree of overreaction, as implied by the forecasting model, the required reaction of monetary policy is relatively weaker.

Overall, the findings suggest the need for models that allow for deviations from rationality while acknowledging that forecasting rules are context-dependent. In practical terms, achieving effective stabilization requires a policy coefficient that is higher than what would be suggested by rational expectations, yet not as strong as implied by fully backward-looking models. Agents' adjustment of forecasting behavior highlights the importance of credibility and effective communication in central banking, ensuring individuals are aware of feedback mechanisms.

References

Adam, K. (2007). Experimental Evidence on the Persistence of Output and Inflation. *The Economic Journal*, 117(520):603–636.

Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., and Thesmar, D. (2023). Overreaction in Expectations: Evidence and Theory. *Quarterly Journal of Economics*, 138:1713–1764.

Angeletos, G.-M., Huo, Z., and Sastry, K. A. (2021). Imperfect Macroeconomic Expectations: Evidence and Theory. *NBER Macroeconomics Annual*, 35:1–86.

Asparouhova, E., Hertzel, M., and Lemmon, M. (2009). Inference from Streaks in Random Outcomes: Experimental Evidence on Beliefs in Regime Shifting and the Law of Small Numbers. *Management Science*, 55(11):1766–1782.

Assenza, T., Heemeijer, P., Hommes, C. H., and Massaro, D. (2021). Managing Self-Organization of Expectations through Monetary Policy: A Macro Experiment. *Journal of Monetary Economics*, 117:170–186.

Bao, T. and Duffy, J. (2016). Adaptive versus Eductive Learning: Theory and Evidence. *European Economic Review*, 83:64–89.

Bao, T., Hommes, C., and Pei, J. (2021). Expectation Formation in Finance and Macroeconomics: A Review of New Experimental Evidence. *Journal of Behavioral and Experimental Finance*, 32:100591.

Barberis, N., Greenwood, R., Jin, L., and Shleifer, A. (2015). XCAPM: An Extrapolative Capital Asset Pricing Model. *Journal of Financial Economics*, 115:1–24.

Barberis, N. and Shleifer, A. (2003). Style Investing. *Journal of Financial Economics*, 68(2):161–199.

Beshears, J., Choi, J. J., Fuster, A., Laibson, D., and Madrian, B. C. (2013). What Goes Up Must Come Down? Experimental Evidence on Intuitive Forecasting. *American Economic Review*, 103(3):570–574.

Beutel, J. and Weber, M. (2025). Beliefs and Portfolios: Causal Evidence. Chicago Booth Research Paper 22-08.

Bianchi, F., Ilut, C., and Saijo, H. (2024). Diagnostic Business Cycles. *The Review of Economic Studies*, 91(1):129–162.

Bordalo, P., Gennaioli, N., La Porta, R., and Shleifer, A. (2018). Diagnostic Expectations and Credit Cycles. *Journal of Finance*, 73:199–227.

Bordalo, P., Gennaioli, N., La Porta, R., and Shleifer, A. (2019). Diagnostic Expectations and Stock Returns. *The Journal of Finance*, 74(6):2839–2874.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, Attention, and Choice. *Quarterly Journal of Economics*, 135:1399–1442.

Bouchaud, J.-P., Krüger, P., Landier, A., and Thesmar, D. (2019). Sticky Expectations and the Profitability Anomaly. *The Journal of Finance*, 74:639–674.

Branch, W. A. and McGough, B. (2009). A New Keynesian Model with Heterogeneous Expectations. *Journal of Economic Dynamics and Control*, 33(5):1036–1051.

Broer, T. and Kohlhas, A. N. (2024). Forecaster (Mis-)Behavior. *The Review of Economics and Statistics*, 106(5):1334–1351.

Cagan, P. (1956). The Monetary Dynamics of Hyperinflation. In *Studies in the Quantity Theory of Money*, pages 25–117. University of Chicago Press.

Chen, D. L., Schonger, M., and Wickens, C. (2016). oTree—An Open-source Platform for Laboratory, Online, and Field Experiments. *Journal of Behavioral and Experimental Finance*, 9(C):88–97.

Clarida, R., Galí, J., and Gertler, M. (2000). Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory. *The Quarterly Journal of Economics*, 115(1):147–180.

Coibion, O. and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, 105:2644–2678.

Cornand, C. and Heinemann, F. (2022). Monetary Policy Obeying the Taylor Principle Turns Prices Into Strategic Substitutes. *Journal of Economic Behavior and Organization*, 200:1357–1371.

Cowan, N. (2017). The Many Faces of Working Memory and Short-Term Storage. *Psychonomic Bulletin & Review*, 24(4):1158–1170.

da Silveira, R. A., Sung, Y., and Woodford, M. (2024). Optimally Imprecise Memory and Biased Forecasts. *American Economic Review*, 114(10):3075–3118.

D'Acunto, F., Malmendier, U., and Weber, M. (2023). What do the data tell us about inflation expectations? In *Handbook of Economic Expectations*, chapter 5, pages 133–161. Academic Press.

Dräger, L. and Lamla, M. J. (2024). Consumers' Macroeconomic Expectations. *Journal of Economic Surveys*, 38(2):427–451.

Evans, G. W., Gibbs, C. G., and McGough, B. (2025). A Unified Model of Learning to Forecast. *American Economic Journal: Macroeconomics*, 17(2):101–33.

Evans, G. W. and Honkapohja, S. (2001). *Learning and Expectations in Macroeconomics*. Princeton University Press.

Evans, J. S. B. T. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, 59:255–278.

Frydman, C. and Nave, G. (2017). Extrapolative Beliefs in Perceptual and Economic Decisions: Evidence of a Common Mechanism. *Management Science*, 63:2340–2352.

Gabaix, X. (2014). A Sparsity-Based Model of Bounded Rationality. *Quarterly Journal of Economics*, 129:1661–1710.

Galí, J. (2015). *Monetary Policy, Inflation, and the Business Cycle: An Introduction to the New Keynesian Framework and Its Applications*. Princeton University Press, 2nd edition.

He, S. and Kučinskas, S. (2024). Expectation Formation with Correlated Variables. *The Economic Journal*, 134(660):1517–1544.

Hey, J. D. (1994). Expectations Formation: Rational or Adaptive or . . . ? *Journal of Economic Behavior & Organization*, 25:329–349.

Hitch, G. J., Hu, Y., Allen, R. J., and Baddeley, A. D. (2018). Competition for the Focus of Attention in Visual Working Memory: Perceptual Recency versus Executive Control. *Annals of the New York Academy of Sciences*, 1424:64–75.

Hong, H. and Stein, J. C. (1999). A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets. *Journal of Finance*, 54(6):2143–2184.

Kahneman, D. and Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*, 3(3):430–454.

Kohlhas, A. N. and Walther, A. (2021). Asymmetric Attention. *American Economic Review*, 111:2879–2925.

Kőszegi, B. and Matějka, F. (2020). Choice Simplification: A Theory of Mental Budgeting and Naive Diversification. *Quarterly Journal of Economics*, 135(2):1153–1207.

Kryvtsov, O. and Petersen, L. (2019). Expectations and Monetary Policy: Experimental Evidence.

Kučinskas, S. and Peters, F. S. (2022). Measuring Under- and Overreaction in Expectation Formation. *The Review of Economics and Statistics*, pages 1–45.

L'Huillier, J.-P., Singh, S. R., and Yoo, D. (2024). Incorporating Diagnostic Expectations into the New Keynesian Framework. *The Review of Economic Studies*, 91(5):3013–3046.

Ma, Y., Ropele, T., Sraer, D., and Thesmar, D. (2024). A Quantitative Analysis of Distortions in Managerial Forecasts.

Mankiw, N. G. and Reis, R. (2002). Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve. *Quarterly Journal of Economics*, 117:1295–1328.

Marimon, R. and Sunder, S. (1993). Indeterminacy of Equilibria in a Hyperinflationary World: Experimental Evidence. *Econometrica*, 61(5):1073–1107.

Maćkowiak, B. and Wiederholt, M. (2009). Optimal Sticky Prices under Rational Inattention. *The American Economic Review*, 99:769–803.

Pfajfar, D. and Žakelj, B. (2018). Inflation Expectations and Monetary Policy Design: Evidence from the Laboratory. *Macroeconomic Dynamics*, 22(4):1035–1075.

Reimers, S. and Harvey, N. (2011). Sensitivity to Autocorrelation in Judgmental Time Series Forecasting. *International Journal of Forecasting*, 27:1196–1214.

Sims, C. A. (2003). Implications of Rational Inattention. *Journal of Monetary Economics*, 50:665–690.

Walsh, C. E. (2010). *Monetary Theory and Policy*. MIT Press, 3rd edition.

Woodford, M. (2003). *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

Woodford, M. (2014). Stochastic Choice: An Optimizing Neuroeconomic Model. *American Economic Review*, 104(5):495–500.

A Proofs

Proofs in sections A.1 and A.2 below closely follow the logic of the proof in Afrouzi et al. (2023) (Online Appendix B.2-3) and are adjusted to incorporate the feedback extension. The main difference, aside from the aggregate forecast entering the law of motion, is that agents observe y_{t-1} realizations in every period, instead of y_t, which is a direct consequence of the feedback in the environment.

A.1 Optimal posterior precision

Proof. The agent's problem consists of two decisions: (i) selecting the information set $S_t \in \mathcal{A}_t$ that minimizes the processing costs and the ex-ante forecast error, and (ii) choosing the optimal forecast $f_t^i y_{t+1}$ given the chosen S_t. The problem is solved backwards by defining the optimal individual forecast for any S_t, and then solving for the optimal S_t, given the optimal forecast.

With a quadratic loss function, the optimal individual forecast for a given S_t is the unbiased expectation of y_{t+1} conditional on S_t. Formally,

$$f_t^{i*} y_{t+1}(S_t) \equiv \arg\min_{f_t^i y_{t+1}} \mathbb{E}[(f_t^i y_{t+1} - y_{t+1})^2 | S_t], \tag{37}$$

where $f_t^{i*} y_{t+1}(S_t)$ denotes the optimal individual forecast for a given S_t. (37) then implies $f_t^{i*} y_{t+1}(S_t) = \mathbb{E}[y_{t+1}|S_t]$, and it follows that the variance of y_{t+1} conditional on S_t is the loss from an inaccurate forecast:

$$\mathbb{E}[f_t^{i*} y_{t+1} - y_{t_1})^2 | S_t] = \text{var}(y_{t+1}|S_t). \tag{38}$$

Decomposing the variance into uncertainty about μ and the short-run fluctuations yields

$$\text{var}(y_{t+1}|S_t) = \text{var}((1-\rho)(1+\rho+\delta\psi)\mu + (\rho+\delta\psi)^2 y_{t-1} + (\rho+\delta\psi)\nu_t + \nu_{t+1}|S_t) \tag{39}$$

$$= (1-\rho)^2(1+\rho+\delta\psi)^2 \text{var}(\mu|S_t) + (1+\rho+\delta\psi)\sigma_\nu^2, \tag{40}$$

where $\nu_{t+j} \perp \mathcal{A}_t, \forall j \geq 0$, and $\text{var}(y_{t-1}|S_t) = 0$ since $y_{t-1} \in S_t$ by assumption. The second

term in (40) is independent of S_t. We can then rewrite the agent's problem as

$$\min_{S_t} \mathbb{E}[(1-\rho)^2(1+\rho+\delta\psi)^2 \text{var}(\mu|S_t) + C(S_t)|y_{t-1}] \tag{41}$$

$$\text{s.t. } \{y_{t-1}\} \subseteq S_t \subseteq \mathcal{A}_t, \tag{42}$$

where the expectation is taken conditional on y_{t-1} because the choice of S_t happens after the agent observes the last realization but before information is processed.

The second part of the proof is to show that the distribution of $\mu|S_t$ will be Gaussian under the optimal use of information. Afrouzi et al. (2023) prove this is the case by showing that for any arbitrary $S_t \in \mathcal{A}_t$, there exists $\hat{S}_t \in \mathcal{A}_t$ that yields a Gaussian posterior and a weakly lower value of the objective function, relative to S_t. The proof carries over verbatim when conditioning on y_{t-1} because (i) by assumption the prior $\mu|y_{t-1}$ is Gaussian, and (ii) by construction the processing cost is increasing in the conditional mutual information $\mathbb{I}(S_t; \mu \mid y_{t-1})$. This allows for redefining the cost function as

$$C(S_t) = \omega \frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma} \tag{43}$$

where $\tau(S_t) \equiv \text{var}(\mu|S_t)^{-1}$ is the posterior precision of belief about μ implied by S_t, and $\underline{\tau} \equiv \text{var}(\mu|y_{t-1})^{-1}$ the precision of the prior. The agent's problem can then be rewritten as

$$\min_{S_t} \mathbb{E}\left[\frac{(1-\rho)^2(1+\rho+\delta\psi)^2}{\tau(S_t)} + \omega \frac{\left(\frac{\tau(S_t)}{\underline{\tau}}\right)^{\gamma} - 1}{\gamma} \;\middle|\; y_{t-1} \right] \tag{44}$$

$$\text{s.t. } \{y_{t-1}\} \subseteq S_t \subseteq \mathcal{A}_t, \tag{45}$$

The problem can be reduced to the choice of the optimal posterior precision τ because the objective depends only on that precision, induced by S_t, as defined in (12), where the bounds are implied by the constraint on S_t. The optimal posterior precision is bounded below by the precision of the prior, $\underline{\tau}$, and above by $\bar{\tau}_t \equiv \text{var}(\mu|y^{t-1})^{-1}$, i.e., the precision achievable after processing all available information, which is arbitrarily small by assumption so we can assume it does not bind. Solving (12) for τ yields the following

K-T conditions:

$$-\frac{(1-\rho)^2(1+\rho+\delta\psi)^2}{\tau}+\frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma}\geq 0, \tag{46}$$

$$\tau\geq\underline{\tau}, \tag{47}$$

$$\left(-\frac{(1-\rho)^2(1+\rho+\delta\psi)^2}{\tau}+\frac{\omega}{\tau}\left(\frac{\tau}{\underline{\tau}}\right)^{\gamma}\right)(\tau-\underline{\tau})=0. \tag{48}$$

which yields the expression for the optimal posterior precision as defined in (13). □

A.2 Optimal individual forecast

Proof. To derive the optimal individual forecast, we need to find the optimal set $S_t\supseteq\{y_{t-1}\}$ that implies the optimal posterior precision as defined in (13). If we define $\mu_t\equiv\mathbb{E}[f_t^i y_{t+1}|y_{t-1},\mu]$ as the conditional mean of the individual forecast given the most recent observation y_{t-1} and the true μ, two cases arise. First, if

$$\left(\frac{\omega\underline{\tau}}{(1+\rho+\delta\psi)^2(1-\rho)^2}\right)\geq 1 \tag{49}$$

then $\text{var}(\mu|S_t)=\text{var}(\mu|y_{t-1})$ and the agent does not use more information than what is implied by their prior, which means that $\mathbb{E}[\mu|S_t]=\mathbb{E}[\mu|y_{t-1}]=y_{t-1}$, and

$$\begin{aligned}\mu_t &\equiv \mathbb{E}[\mathbb{E}[y_{t+1}|S_t]|y_{t-1},\mu] && (50)\\ &= (1-\rho)(1+\rho+\delta\psi)\,\mathbb{E}[\mathbb{E}[\mu|S_t]|y_{t-1},\mu]+(\rho+\delta\psi)^2\,\mathbb{E}[\mathbb{E}[y_{t-1}|S_t]|y_{t-1},\mu] && (51)\\ &= \left((1-\rho)(1+\rho+\delta\psi)+(\rho+\delta\psi)^2\right)y_{t-1} && (52)\end{aligned}$$

Second, if

$$\left(\frac{\omega\underline{\tau}}{(1+\rho+\delta\psi)^2(1-\rho)^2}\right)<1 \tag{53}$$

than the optimal set S_t contains more information than implied by the prior. Afrouzi et al. (2023) show that for any set $\tilde{S}_t$ that delivers the optimal posterior precision, a set $\hat{S}_t\equiv\{y_{t-1},\mathbb{E}[\mu|\tilde{S}_t]\}$ is a sufficient statistic, as both contain Gaussian variables and, by the law of total variance, both generate the same posterior variance.

Then, from Bayesian updating:

$$\mathbb{E}[\mu|S_t] = \mathbb{E}[\mu|\tilde{S}_t] = \mathbb{E}[\mu|y_{t-1}] + \frac{\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|y_{t-1})}{\text{var}(\mathbb{E}[\mu|\tilde{S}_t]|y_{t-1})}(\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|y_{t-1}]) \tag{54}$$

which implies that

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|y_{t-1}) = \text{var}(\mathbb{E}[\mu|\tilde{S}_t]|y_{t-1}) = \underline{\tau}^{-1} - \tau^{-1} \tag{55}$$

since $\mathbb{E}[\mu|\tilde{S}_t] - \mathbb{E}[\mu|y_{t-1}] \neq 0$ almost surely, with the last equality coming from the law of total variance. Then, we can decompose $\mathbb{E}[\mu|\tilde{S}_t]$ as

$$\mathbb{E}[\mu|\tilde{S}_t] = a\mu + by_{t-1} + \varepsilon_t, \tag{56}$$

where a and b are constants, and ε_t is orthogonal to y_{t-1} and μ conditional on $\tilde{S}_t$. From the agent's prior

$$y_{t-1} = \mathbb{E}[\mu|y_{t-1}] = \mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|y_{t-1}] = a\,\mathbb{E}[\mu|y_{t-1}] + by_{t-1} = (a+b)y_{t-1} \tag{57}$$

which makes $(a+b) = 1$. Furthermore,

$$\text{cov}(\mu, \mathbb{E}[\mu|\tilde{S}_t]|y_{t-1}) = a\text{var}(\mu|y_{t-1}), \tag{58}$$

which makes $a = 1 - \frac{\underline{\tau}}{\tau}$ and

$$\mathbb{E}[\mathbb{E}[\mu|\tilde{S}_t]|\mu, y_{t-1}] = \left(1 - \frac{\underline{\tau}}{\tau}\right)\mu + \frac{\underline{\tau}}{\tau}y_{t-1} \tag{59}$$

which characterizes μ_t as

$$\mu_t \equiv \mathbb{E}[\mathbb{E}[y_{t+1}|\tilde{S}_t]|\mu, y_{t-1}] \tag{60}$$

$$= (1-\rho)(1+\rho+\delta\psi)\left[\left(1 - \frac{\underline{\tau}}{\tau}\right)\mu + \frac{\underline{\tau}}{\tau}y_{t-1}\right] + (\rho+\delta\psi)^2 y_{t-1} \tag{61}$$

Defining $u_t \equiv f_t^i y_{t+1} - \mu_t$, yields the optimal individual forecast as defined in (15):

$$f_t^i y_{t+1} = (1-\rho)(1+\rho+\delta\psi)\left[\left(1-\frac{\tau}{\tau}\right)\mu + \frac{\tau}{\tau}y_{t-1}\right] + (\rho+\delta\psi)^2 y_{t-1} + u_t \tag{62}$$

with $\mathbb{E}[u_t|y_{t-1},\mu] = 0$ and $\alpha \equiv \frac{\tau}{\tau}$. □

A.3 Equilibrium ψ

The optimal posterior precision τ^*, aggregate forecast $F_t y_{t+1} = \psi y_{t-1}$ and the optimal individual forecast jointly pin down ψ in equilibrium. Defining $\alpha \equiv \frac{\tau}{\tau^*}$, and when $\mu = 0$ the optimal individual forecast is defined as

$$f_t^{i*} y_{t+1} = \left(\alpha(1-\rho)(1+\rho+\delta\psi) + (\rho+\delta\psi)^2\right) y_{t-1}. \tag{63}$$

Then, from the aggregate forecast it follows:

$$\psi = \alpha(1-\rho)(1+\rho+\delta\psi) + (\rho+\delta\psi)^2, \tag{64}$$

and

$$\delta^2\psi^2 + \left[\delta(2\rho+(1-\rho)\alpha) - 1\right]\psi + \left[\rho^2 + (1-\rho^2)\alpha\right] = 0 \tag{65}$$

which defines $\psi(\alpha)$ as

$$\psi(\alpha) = \frac{1-\delta(2\rho+(1-\rho)\alpha) \;\pm\; \sqrt{\left[1-\delta(2\rho+(1-\rho)\alpha)\right]^2 - 4\delta^2\left[\rho^2+(1-\rho^2)\alpha\right]}}{2\delta^2}. \tag{66}$$

A.4 Proposition 1

Proof. For $\delta \in \mathbb{R}$ and $\rho \in [0,1)$, fix $\psi > 0$ and let $\alpha = \alpha(\psi;\xi) \in (0,1]$.

i. Set $a \equiv (\rho+\delta\psi)$ and $b \equiv \alpha(1-\rho)$, where $b \in (0,1]$, since $\alpha \in (0,1]$ and $(1-\rho) \in (0,1]$. Then

$$\lambda = a^2 + b(1+a) = \left(a+\frac{b}{2}\right)^2 + b\left(1-\frac{b}{4}\right), \tag{67}$$

where $\left(a+\frac{b}{2}\right)^2 > 0$ and $\left(1-\frac{b}{4}\right) > 0$ for $b \in (0,1]$, so $\lambda > 0$. In the frictionless limit

$\alpha \to 0$, $\lambda = (\rho + \delta\psi)^2 \geq 0$ with $\lambda = 0$ only on the edge case of $\delta = -\frac{\rho}{\psi}$.

ii. Strict convexity follows from $\frac{\partial^2 \lambda}{\partial \delta^2} = 2\psi^2 > 0$ for $\psi > 0$. From the first order condition:

$$\frac{\partial \lambda}{\partial \delta} = 0 \quad \Rightarrow \quad \delta^{\min} = -\frac{2\rho + \alpha(1-\rho)}{2\psi} < 0. \tag{68}$$

iii. For $\delta > 0$, it is straightforward to see that

$$\frac{\partial \lambda}{\partial \delta} = \alpha\psi(1-\rho) + 2\psi(\rho + \delta\psi) > 0, \tag{69}$$

which implies $\lambda(\rho, \delta; \psi) - \lambda(\rho, 0; \psi) = \delta\psi(\delta\psi + \alpha(1-\rho) + 2\rho) > 0$.

iv. For $\delta < 0$, (ii.) implies

$$\frac{\partial \lambda}{\partial \delta} \begin{cases} > 0, & \text{if } \delta > \delta^{\min}, \\ < 0, & \text{if } \delta < \delta^{\min}. \end{cases} \tag{70}$$

Moreover,

$$\lambda(\rho, 0; \psi) = \lambda(\rho, \delta; \psi) \quad \Rightarrow \quad \delta_0 = -\frac{2\rho + \alpha(1-\rho)}{\psi} \tag{71}$$

so $\lambda(\rho, 0; \psi) > \lambda(\rho, \delta; \psi)$ for all $\delta \in (\delta_0, 0)$. In addition, $\delta_0 = 2\delta^{\min}$, so $\delta_0 < \delta^{\min} < 0$.

v. It follows from:

$$\frac{\partial \delta^{\min}}{\partial \rho} = -\frac{2-\alpha}{2\psi} < 0 \tag{72}$$

$$\frac{\partial \delta_0}{\partial \rho} = -\frac{2-\alpha}{\psi} < 0. \tag{73}$$

Both thresholds are strictly decreasing in ρ for the given parameters.

vi. Similarly, it follows from:

$$\frac{\partial^2 \lambda}{\partial \alpha \partial \delta} = \psi(1-\rho) > 0 \tag{74}$$

$$\frac{\partial^2 \lambda}{\partial \rho \partial \delta} = \psi(2-\alpha) > 0 \tag{75}$$

Therefore, the marginal effect of δ is strictly increasing both in α and ρ for the given parameters. □

A.5 Proposition 2

Proof. For $\rho \in [0,1)$ and $\delta > -\frac{1+\rho}{\psi}$, fix $\psi > 0$ and let $\alpha = \alpha(\psi;\xi) \in (0,1]$.

i. Since $\alpha > 0$ and $(1-\rho) > 0$,

$$\beta > 0 \quad \Leftrightarrow \quad (1+\rho+\delta\psi) > 0 \quad \Leftrightarrow \quad \delta > -\frac{1+\rho}{\psi}, \tag{76}$$

where the last inequality coincides with the left boundary of the stationarity condition:

$$|\rho+\delta\psi| < 1 \quad \Leftrightarrow \quad \delta \in \left(-\frac{1+\rho}{\psi}, \frac{1-\rho}{\psi}\right). \tag{77}$$

ii. Under the same restriction on δ as in (76), it follows that

$$\frac{\partial\beta}{\partial\alpha} = (1-\rho)(1+\rho+\delta\psi) > 0. \tag{78}$$

Moreover, for any $\delta \in \mathbb{R}$:

$$\frac{\partial\beta}{\partial\delta} = \alpha(1-\rho)\psi > 0. \tag{79}$$

iii. Strict concavity follows from $\frac{\partial^2\beta}{\partial\rho^2} = -2\alpha < 0$ for $\alpha > 0$. From the first order condition:

$$\frac{\partial\beta}{\partial\rho} = 0 \quad \Rightarrow \quad \rho^{\max} = -\frac{\delta\psi}{2}. \tag{80}$$

iv. For $\delta \geq 0$ and $\alpha > 0$, it is straightforward to see that

$$\frac{\partial\beta}{\partial\rho} = \alpha(-2\rho-\delta\psi) < 0, \tag{81}$$

and $\rho^{\max} < 0$. For $\delta < 0$, (iii.) implies

$$\frac{\partial\beta}{\partial\rho} \begin{cases} > 0, & \text{for } \rho \in [0, \rho^{\max}), \\ < 0, & \text{for } \rho \in (\rho^{\max}, 1). \end{cases} \tag{82}$$

Moreover, $\rho^{\max}$ is strictly decreasing in δ,

$$\frac{\partial \rho^{\max}}{\partial \delta} = -\frac{1}{2}\psi < 0, \tag{83}$$

so a stronger negative feedback pushes the maximum to a higher ρ. Furthermore, for any $0 \leq \rho_L < \rho_H < 1$,

$$\beta(\rho_H) - \beta(\rho_L) = \alpha\left[-(\rho_H - \rho_L)((\rho_H + \rho_L) + \delta\psi)\right]. \tag{84}$$

Since $\alpha > 0$ and $\rho_H - \rho_L > 0$,

$$\beta(\rho_H) < \beta(\rho_L) \quad \Leftrightarrow \quad (\rho_H + \rho_L) + \delta\psi \quad \Leftrightarrow \quad \delta > -\frac{\rho_L + \rho_H}{\psi}. \tag{85}$$

v. It follows from:

$$\frac{\partial^2 \beta}{\partial \alpha \partial \delta} = \psi(1 - \rho) > 0 \tag{86}$$

$$\frac{\partial^2 \beta}{\partial \rho \partial \delta} = -\alpha\psi < 0. \tag{87}$$

Therefore, the marginal effect of δ is strictly increasing in α, and strictly decreasing in ρ for the given parameters. □

A.6 Proposition 3

Proof. For $\rho \in [0, 1)$ and $\delta \in \mathbb{R}$, fix $\psi > 0$ and let $\alpha = \alpha(\psi; \xi) \in (0, 1]$.

i. Follows directly from (25) for $\delta = 0$.

ii. Strict concavity follows from $\frac{\partial^2 \beta^{NR}}{\partial \delta^2} = -2\psi^2 < 0$ for $\psi > 0$. From the first order condition:

$$\frac{\partial \beta^{NR}}{\partial \delta} = 0 \quad \Rightarrow \quad \delta^{\max} = -\frac{\rho}{\psi} < 0. \tag{88}$$

iii. For $\delta > 0$ it is straightforward to see that

$$\frac{\partial \beta^{NR}}{\partial \delta} = -2\psi(\rho + \delta\psi) < 0, \tag{89}$$

which implies $\beta^{NR}(\rho,\delta;\psi) - \beta^{NR}(\rho,0;\psi) = -2\psi\rho\delta - \psi^2\delta^2 < 0$.

iv. For $\delta < 0$, (ii.) implies

$$\frac{\partial \beta^{NR}}{\partial \delta} \begin{cases} < 0, & \text{for } \delta > \delta^{max}, \\ > 0, & \text{for } \delta < \delta^{max}. \end{cases} \tag{90}$$

Moreover, for $\rho > 0$

$$\beta^{NR}(\rho,\delta;\psi) = \beta^{NR}(\rho,0;\psi) \quad \Rightarrow \quad \delta^0 = -\frac{2\rho}{\psi} \tag{91}$$

so $\beta^{NR}(\rho,\delta;\psi) > \beta^{NR}(\rho,0;\psi)$ for all $\delta \in (\delta^0, 0)$. For $\rho = 0$, $\delta^{\max} = \delta_0 = 0$. In addition, $\delta^0 = 2\delta^{\max}$, so $\delta^0 < \delta^{\max} < 0$.

v. It follows from:

$$\frac{\partial \delta^{\max}}{\partial \rho} = -\frac{1}{\psi} < 0 \tag{92}$$

$$\frac{\partial \delta^0}{\partial \rho} = -\frac{2}{\psi} < 0, \tag{93}$$

so both thresholds are strictly decreasing in ρ for the given parameters. □

B New Keynesian model: Alternative specification

In the main specification, ψ_x is set to 1, for simplicity, and to isolate the mechanism of interest in illustrating the main consequences of embedding the forecasting bias implied by the experiment in the New Keynesian setting. This section relaxes this assumption. There are three main differences: (i) in the biased version, the degree of extrapolation from the past is, by assumption, the same for both inflation and output expectations (i.e., $\psi_x = \psi_\pi = \psi$), (ii) in the rational expectations benchmark both inflation and output expectations are formed rationally, to consistently compare it with the biased version, and (iii) the extrapolation parameter is solved from $\psi = f(\rho(\psi), \delta)$. Figures (B.1-B.3) repeat the complete analysis of Section 5. This specification yields the same δ values across the three cases as in the main version, but a somewhat lower values of both ψ and ρ. Regardless, the main conclusions remain valid.

B.1 Case 1: $\phi_\pi = 1.01$

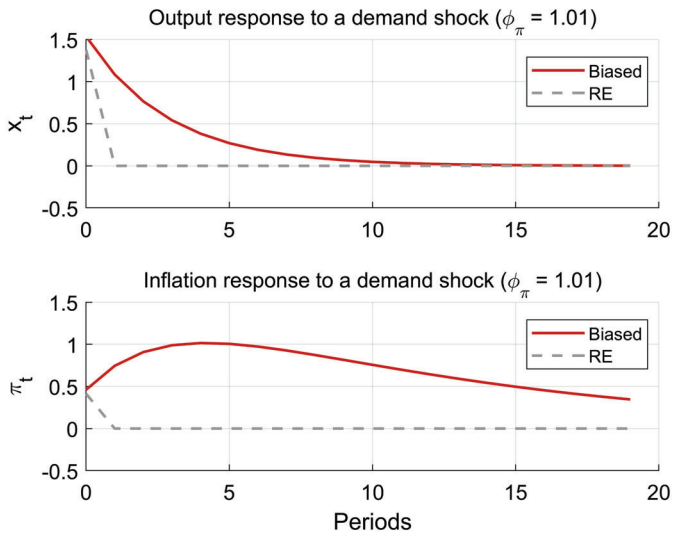
The first case, in Figure (B.1), shows the impulse response of inflation and output gap to a demand shock with a Taylor coefficient that barely satisfies the Taylor principle. As in the original specification, this translates to zero feedback ($\delta = 0$), but persistence and the extrapolation parameter are slightly smaller, at $\rho = 0.70$ and $\psi = 0.92$ respectively.

In the rational expectations benchmark, both inflation and output gap return to the equilibrium immediately after the initial impact of the shock. With biased expectations, although the initial effect is the same, convergence in both variables is much slower, with inflation not reaching the equilibrium even after 20 periods.

B.2 Case 2: $\phi_\pi = 1.5$

In the second case, a more aggressive monetary policy characterized by $\phi_\pi = 1.5$ leads to the same negative feedback at $\delta = -0.33$, with persistence $\rho = 0.39$, implying $\psi = 0.57$. Figure (B.2) shows the same impulse response functions of the output gap and inflation to a demand shock. The grey dashed line represents the responses in the rational benchmark, while the other two lines depict the responses in the model with biased expectations. The 'Adjustment' line reflects the response with $\psi = 0.57$ in expectations, and 'No adjustment'

Figure B.1: **IRFs with zero feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.01$, over 20 periods. Solid red lines depict the biased-expectations model with a common extrapolation weight $\psi_x = \psi_\pi = 0.92$; dashed grey lines depict the rational-expectations benchmark in which both inflation and output expectations are formed rationally.

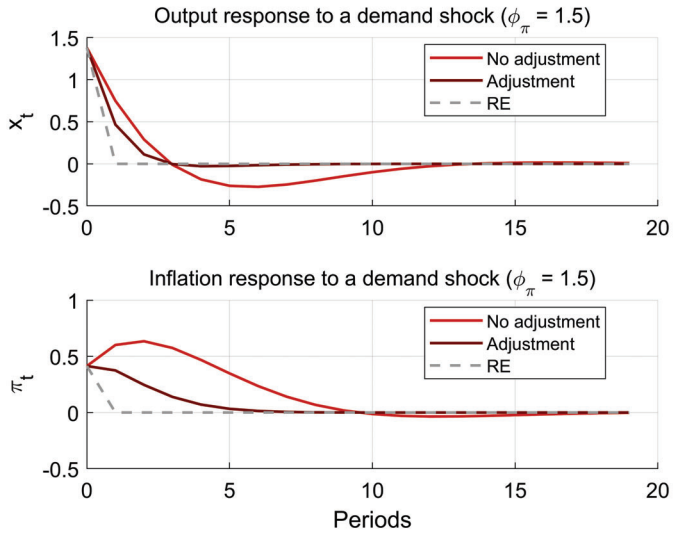
the responses under the assumption that agents use the same weight in their forecasts as if δ remains at zero.

The interpretation is the same as under the main specification. A more aggressive policy rule reduces persistence and extrapolation, and accelerates convergence in both variables. If expectations do not adjust, stronger negative feedback leads to more persistence in inflation and oscillations around the rational expectations path.

B.3 Case 3: $\phi_\pi^{RE} = 1.5$ and $\phi_\pi^B = \{1.5, 2, 2.5\}$

In the third case, Figure (B.3) compares impulse response functions of output and inflation in reaction to a demand shock for $\phi_\pi^B = \{1.5, 2, 2.5\}$ in the biased version, to the responses under rational expectations with $\phi_\pi^{RE} = 1.5$. Increasing ϕ_π^B implies a stronger negative feedback, lower persistence and less extrapolation in expectations. The only difference in interpretation with respect to the original specification is in inflation not reaching the exact or approximate convergence path as under rational expectations even with a much higher Taylor rule coefficient and with adjustment in expectations, considering that rational expectations return to the equilibrium immediately after the

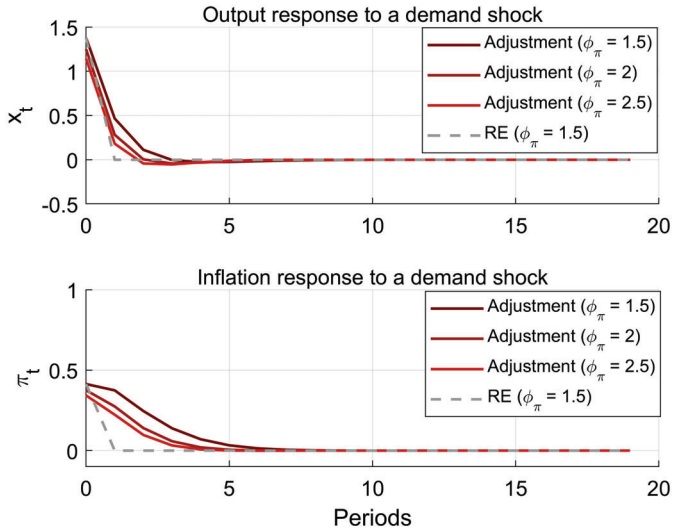
Figure B.2: **IRFs with negative feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.01$, over 20 periods. The grey dashed line depicts the rational-expectations benchmark in which both inflation and output expectations are formed rationally. The solid red line ('No adjustment') depicts the biased-expectations model with a common extrapolation weight $\psi_x = \psi_\pi = 0.92$; the dark red line ('Adjustment') depicts the biased-expectations model with $\psi_x = \psi_\pi = 0.57$.

impact of the shock.

Figure B.3: **IRFs with negative feedback**



Note: This figure shows impulse response functions of the output gap (x_t) and inflation (π_t) to a one-time demand shock under the Taylor rule coefficient $\phi_\pi = 1.01$, over 20 periods. The grey dashed line represents the response in the rational-expectations benchmark with $\phi_\pi = 1.5$. The three solid red lines ("Adjustment") depict the biased-expectations model for $\phi_\pi \in \{1.5, 2, 2.5\}$, using in each case the corresponding equilibrium extrapolation weight in expectations.

C Experimental instructions

Below are the instructions given to participants in the experiment. In *gray* is additional information available in the *Feedback* condition. Otherwise, the instructions are identical between the two.

C.1 Experimental environment

Your task is to predict future values of a random process over 45 consecutive periods. In every period, you will predict the value of the process for the following period.

You will be part of a group of six people randomly selected from the participants in the room today. You will remain in the same group for the duration of the experiment. In every period, once each of you submits your prediction, the computer will calculate the average prediction of the group for that period.

The value of the process in any given period depends on two elements:

1. *Average group prediction for the following period*
    - *The higher your group's average prediction for the following period, the lower the actual value of the process in the current period. Similarly, the lower the group's prediction for the following period, the higher the actual value today.*
    - *This means that there is a negative relationship between your predictions for the following period and the actual value today. It also means that your prediction and the prediction of others in your group for the following period will affect the value of the process in the current period.*
2. *Random term, that is not perfectly predictable.*

Throughout the experiment, except in the first period, you will see all previous values of the process and your individual predictions. *You will not be able to see the predictions of the others in your group or the average predictions of the entire group. You will also not be able to see the value of the random term.*

Note that in each period of the experiment, you will see the actual value of the process only up to the previous period but will be asked to predict its value for the following period.

When making your prediction, you will not be able to see the value of the process in the current period.

C.2 Payments

Your compensation will depend on the accuracy of your predictions: the closer your prediction to the actual values, the higher your prediction score and your payment. Your prediction score will depend on your prediction error, which is the absolute distance between your prediction and the actual value. You can earn a maximum of 100 points for a perfect prediction, and from there, points decrease as the prediction error increases. The table below shows the points you can earn for some examples of prediction errors.

Prediction error and scores

Prediction error	0	1	2	3	4	7	9
Points per period	100	50	33.33	25	20	12.5	10

For example, if you predict the value of the process in the following period to be 5, and it turns out to be 5, your prediction error is zero, and you earn 100 points. If, instead, you guessed it to be 4, your prediction error is equal to 1, and you earn 50 points. You can input values with up to two decimals. You will have up to 1 minute and 30 seconds in every period to input your prediction.

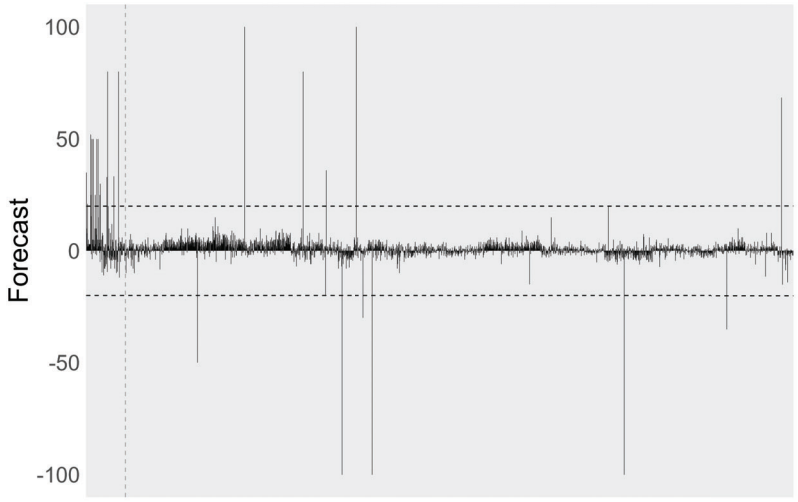
Your final score will be the sum of all the points you earned over the 45 periods in the experiment. For your final payment, you will receive a 3 EUR participation fee and, additionally, 0.75 EUR for every 100 points (1.5 EUR for 200 points, and so on).

D Outliers

In the first period of the experiment, participants do not observe any previous realizations of the process they are forecasting. An alternative approach could provide participants with simulated values of the process for several periods before the experiment begins. However, when feedback from expectations is present, the realized values depend on participants' aggregate forecasts, making it impossible to accurately predefine the actual values of the process. To simulate the data, one would need to make an assumption about how aggregate expectations are formed, which could potentially bias the participants' understanding of what the process *should* look like if influenced by expectations in one particular way.

To avoid this issue, participants start with an empty chart and have to guess the starting values, resulting in noisy initial predictions. Nevertheless, they quickly adapt to the incentive scheme and stabilize their predictions by the fifth period. Figures D.1 and D.2 show overlapping individual predictions from all participants across all periods of the experiment for the *Feedback* and *Baseline* conditions, respectively. The grey dashed vertical line indicates the fifth period. All forecasts from the first five periods are discarded from the dataset as part of the learning phase.

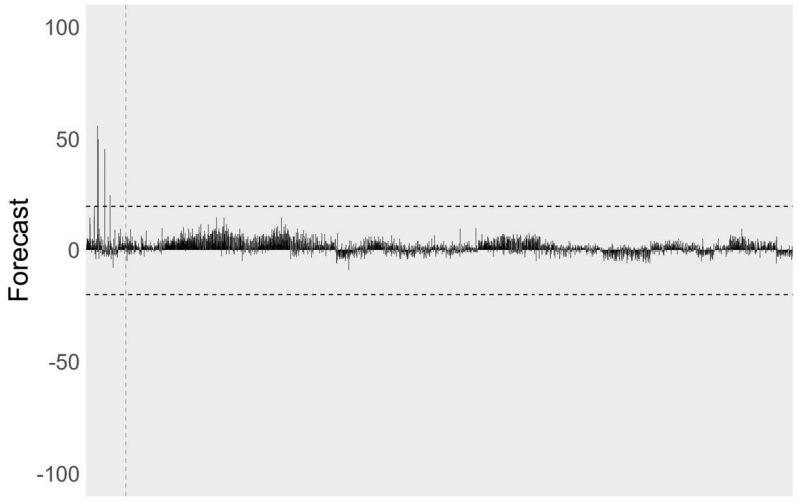
Figure D.1: **Individual forecasts in *Feedback* treatments**



Note: This figure shows overlaid individual forecasts of all participants in both *Feedback* treatments over all 45 periods of the experiment. The vertical dashed line marks the fifth period. The two horizontal dashed lines mark the forecast values of -20 and 20.

Furthermore, Figure D.1 highlights a few extreme predictions that do not appear in the *Baseline* condition shown in Figure D.2. These outliers likely arise from either input errors or intentional deviations from the imposed incentives. Some participants may have experimented with how an unreasonably large forecast could impact group dynamics. Since these values are unlikely to reflect genuine predictions but could significantly affect the results, they have been excluded from the dataset (a total of 11 observations).

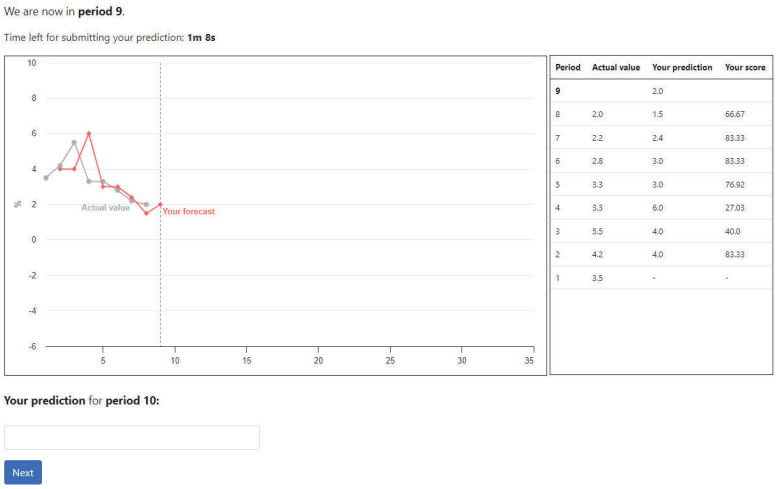
Figure D.2: **Individual forecasts in *Baseline* treatments**



Note: This figure shows overlaid individual forecasts of all participants in both *Baseline* treatments over all 45 periods of the experiment. The vertical dashed line marks the fifth period. The two horizontal dashed lines mark the forecast values of -20 and 20.

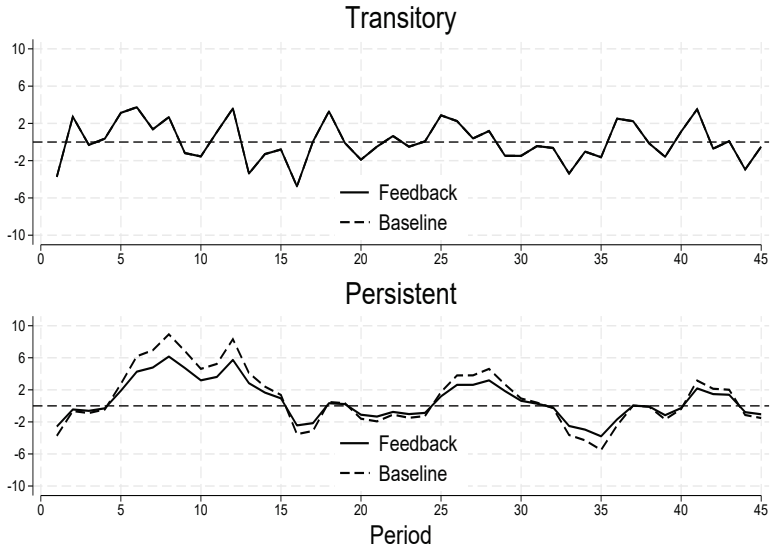
E Figures

Figure E.1: **Experimental interface**



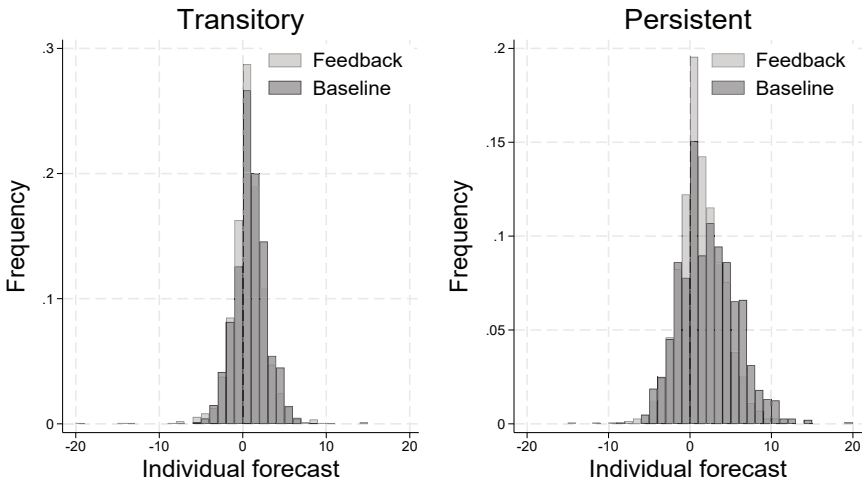
Note: This figure shows the interface participants faced in the experiment. In the chart on the left, they observe the actual realizations of the process and their own individual forecasts over all periods. In the table on the right, they observe the same information, along with their forecasting score for each period. The very top indicates the current period and shows the timer. Participants were encouraged to submit their predictions in 90 seconds in each period. After 90 seconds, a pop-up window would remind them to input their predictions, but they had as much time as they needed.

Figure E.2: **Simulations of the process across treatment groups**



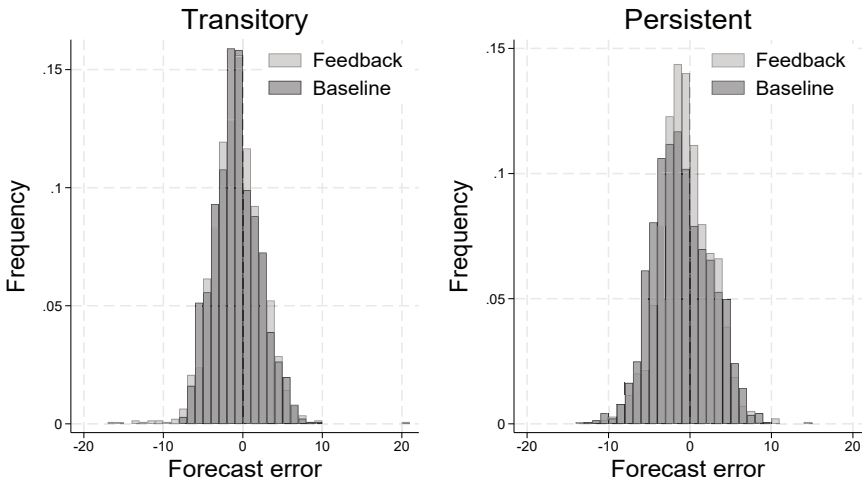
Note: This figure shows the simulated values of the process participants are predicting in the experiment, across all four treatment groups varying the presence of feedback (*Baseline* with $\delta = 0$ or *Feedback* with $\delta = -0.5$) and persistence (*Transitory* with $\rho = 0$ or *Persistent* with $\rho = 0.9$). In *Baseline*, the process is an AR(1) with the same sequence of shocks, varying only the persistence parameter. In *Feedback*, for the simulations in this figure, expectations are assumed to be formed rationally. In the experiment, participants face the same sequence of shocks as in the *Baseline*, but the realizations are affected by their aggregate predictions. In the *Transitory* case, the processes overlap perfectly.

Figure E.3: **Histogram of individual forecasts**



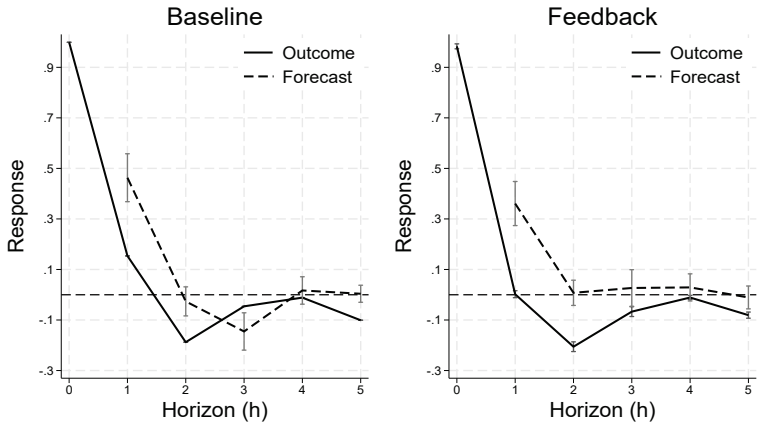
Note: This figure shows the histogram of individual forecasts across all treatment groups, after removing 10 outliers in the Feedback condition, as detailed in Section D. The left (right) panel shows forecasts in Baseline and Feedback conditions in the Transitory (Persistent) group.

Figure E.4: **Histogram of forecast errors**

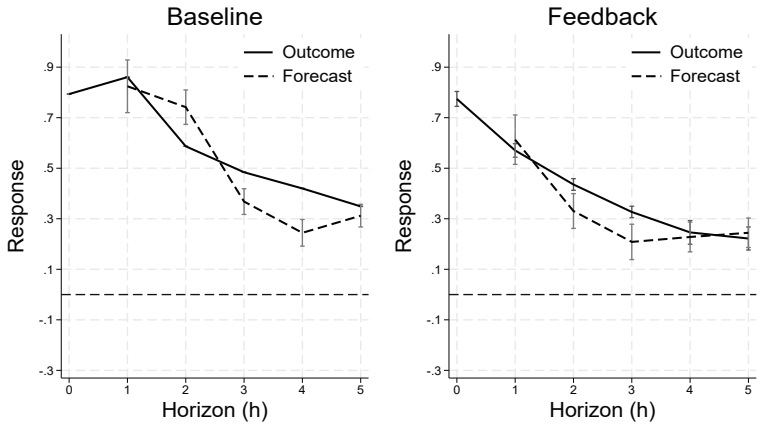


Note: This figure shows the histogram of individual forecast errors ($f_t^i y_{t+1} - y_{t+1}$) across all treatment groups, across all treatment groups, after removing 10 outliers in individual forecasts in the Feedback condition, as detailed in Section D. The left (right) panel shows forecast errors in Baseline and Feedback conditions in the Transitory (Persistent) group.

Figure E.5: **Dynamic response of outcomes and forecasts (Transitory)**

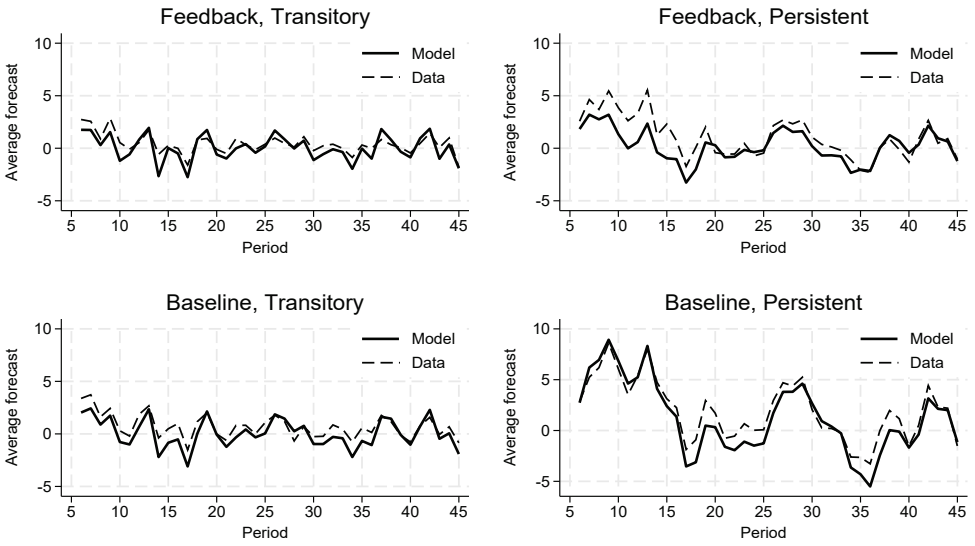


Dynamic response of outcomes and forecasts (Persistent)



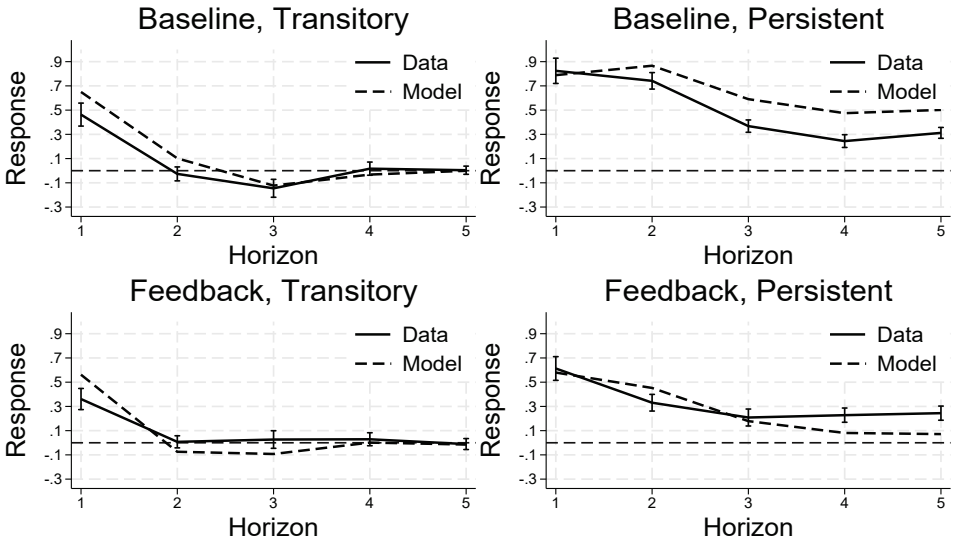
Note: The figure compares estimates of γ_{1h} and γ_{2h} from horizon-h local-projection impulse responses of outcomes and individual forecasts to exogenous shocks, as specified in equations (5) and (6), run separately by treatment group. The top (bottom) panel shows the Transitory (Persistent) group. The left (right) panels show Baseline (Feedback) group. Solid lines denote responses of outcomes and dashed lines responses of forecasts. Points show the estimated response of y_{t+h} or $f_t^i y_{t+1}$ to a one-unit innovation in ϵ. Responses for forecasts start at horizon $h = 1$ because participants in the experiment observe realizations only up to period $t - 1$. Vertical bars are 95% confidence intervals. Standard errors are clustered at the individual level.

Figure E.6: **Average forecasts in model simulations and data**



Note: This figure compares average forecasts across all treatment groups between the data and the model. In the model, the average forecast is simulated using the same sequence of shocks that participants faced in the lab.

Figure E.7: **Response of f^i_{t+h} to ϵ_t: Data and model**



Note: The figure plots estimates of γ_{2h} from horizon-h local-projection impulse response regressions, as specified in equation (6), run separately by treatment group, and compares it to similar estimates using model simulated data. The left (right) panels shows the Transitory (Persistent) group; the top (bottom) panels show the Baseline (Feedback) group. Solid lines represent responses in the experimental data and dashed lines responses in the model simulated data. All series from the model are simulated using the same sequence of shocks that participants faced in the lab. Points show the estimated response of $f^i_{t+h}y_{t+h+1}$ to a one-unit innovation in ϵ_t. Vertical bars in the experimental data estimates are 95% confidence intervals.

F Tables

Table F.1: **Comparison of overreaction estimates with Afrouzi et al. (2023)**

	Transitory		**Persistent**	
	Lab ($\rho = 0$)	Online ($\rho = 0$)	Lab ($\rho = 0.9$)	Online ($\rho = 0.8$)
y_{t-1}	-0.65	-0.52	-0.24	-0.10
	(0.05)	(0.05)	(0.04)	(0.02)
N	1330	1280	1368	1120

Note: This table compares estimates of the overreaction coefficient defined as the correlation between individual forecast errors ($y_{t+1} - f_t^i y_{t+1}$) and the last observation available to the participants, y_{t-1}, as defined in equation (3) between the data collected in the lab and the data from the online experiment of Afrouzi et al. (2023). The overreaction is estimated only for the *Baseline* condition. The data in the lab is collected for persistence values of $\rho = 0$ and $\rho = 0.9$. Afrouzi et al. (2023) collected data for $\rho = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. The closest value for the comparison to the $\rho = 0.9$ lab case is $\rho = 0.8$. Standard errors, in parentheses, are clustered at the individual level.

Table F.2: **Overreaction estimates by treatment (Average data)**

	Transitory		**Persistent**	
	Baseline	Feedback	Baseline	Feedback
y_{t-1}	-0.65	-0.49	-0.24	-0.26
	(0.17)	(0.07)	(0.13)	(0.06)
Intercept	-0.72	-0.67	-0.84	-0.50
	(0.34)	(0.16)	(0.49)	(0.17)
N	38	227	38	223

Note: The table shows estimates of the correlation between forecast errors defined as $y_{t+1} - F_t y_{t+1}$ and the last observation available to the participants, y_{t-1}, separately for each of the four treatment groups. $F_t y_{t+1} \equiv \frac{1}{N}\sum_{i=1}^{N} f_t^i y_{t+1}$ is the aggregate forecast defined as the mean of all individual forecasts within a group. In *Baseline*, the average is computed over all participants in each persistence group. In *Feedback*, the averages are computes within each of the groups of six participants representing one experimental economy.

Table F.3: **Estimates of extrapolation weights**

	Transitory		**Persistent**	
	Baseline	Feedback	Baseline	Feedback
y_{t-1}	0.46	0.34	0.90	0.66
	(0.06)	(0.02)	(0.05)	(0.12)
N	39	234	39	234

Note: This table shows estimates of extrapolation weights $\hat{\psi}$, representing how strongly aggregate forecasts in the experiment $F_t y_{t+1}$ rely on the last available observation, y_{t-1}, estimated as $F_t y_{t+1} = \psi y_{t-1} + \epsilon_t$ across all four treatment groups. The aggregate forecast $F_t y_{t+1} \equiv \frac{1}{N} \sum_{i=1}^{N} f_t^i y_{t+1}$ is defined as the mean of all individual forecasts within a group. In *Baseline*, the average is computed over all participants in each persistence group. In *Feedback*, the averages are computes within each of the groups of six participants representing one experimental economy.