The Inflation Attention Threshold and Inflation Surges

Oliver Pfäuti*
The University of Texas at Austin

March 2026

Link to most recent version

Abstract

The recent inflation surge brought inflation back on people's minds. I quantify when and how much attention to inflation changes and derive the macroeconomic implications of these attention changes. I estimate an attention threshold at an inflation rate of 4%, that attention doubles when inflation exceeds this threshold, and that supply shocks have stronger and more persistent effects on inflation in times of high attention. Developing a model featuring the attention threshold, I show that attention changes offer a joint explanation for inflation surges, inflation's co-movement with inflation expectations, and a long last mile of disinflation.

JEL Codes: E3, E4, E5, E7

Keywords: Inattention, Inflation, Inflation Expectations, Monetary Policy

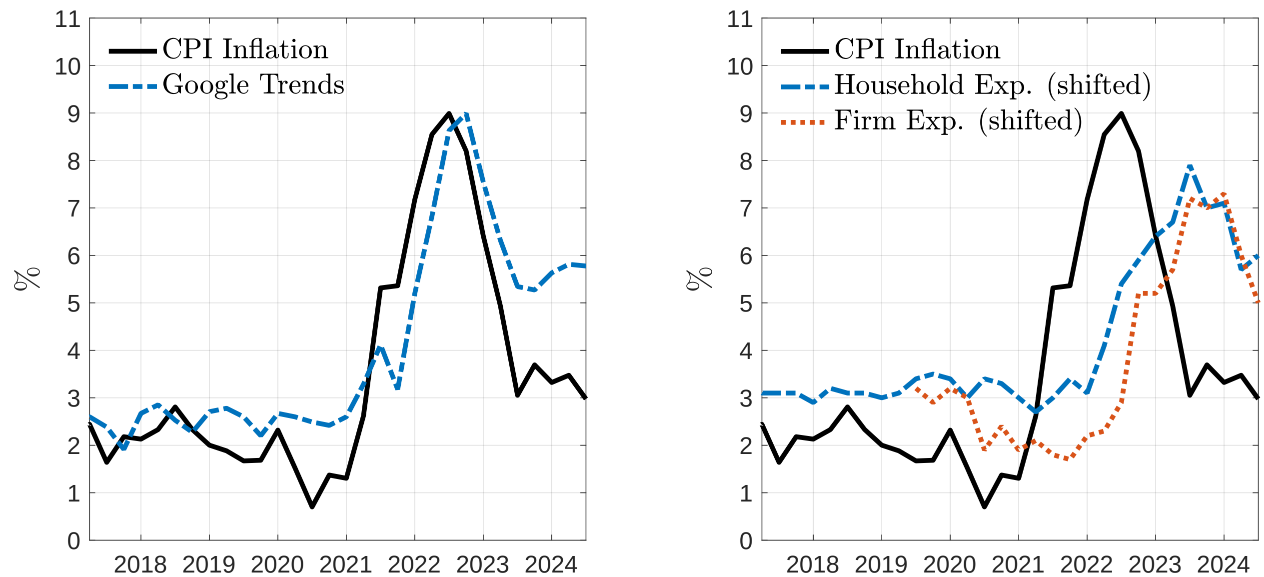
*The University of Texas at Austin, Department of Economics, 2225 Speedway, 78712 Austin TX, pfaeuti.oliver@gmail.com. I thank Hassan Afrouzi, Rudi Bachmann, Luca Benati, Jonathan Benchimol, Pierpaolo Benigno, Jaroslav Borovicka, Oli Coibion, Leland Farmer, Joel Flynn, Francesco Furlanetto, Jose-Elias Gallegos, Laura Gáti, Michael Gelman, Rupal Kamdar, Niklas Kroner, John Leahy, Cyril Monnet, David Munro, Dirk Niepelt, Ricardo Reis, Fabian Seyrich, Hannes Twieling and various seminar and conference participants for helpful comments and suggestions. I thank Manuel Diaz de la Fuente for excellent research assistance. First Version: August 2023.

1 Introduction

Inflation is back. After decades of low and stable inflation, inflation surged in many advanced economies during the recovery phase of the pandemic. Inflation turned out to be higher and more persistent than many expected.[1] With inflation rising, the public's attention to inflation increased as well (panel (a) in Figure 1), and as inflation peaked, households' and firms' inflation expectations also started to pick up (panel (b) in Figure 1). During the build up of the inflation surge, however, inflation expectations substantially underpredicted actual inflation. Only as inflation started to come down, inflation expectations started to overshoot inflation and these expectations remained high while inflation stayed persistently above its target of 2%.

Figure 1: Inflation, attention to inflation, and inflation expectations

(a) Inflation and Google Trends (b) Inflation and inflation expectations



Notes: Panel (a) shows U.S. CPI inflation (black-solid line) together with Google Trends searches of the word 'inflation' in the U.S. (blue-dashed line). Google Trends are normalized to have the same maximum as CPI inflation. Panel (b) shows U.S. CPI inflation (black-solid line) together with one-year-ahead inflation expectations for U.S. households (blue-dashed line) and firms (red-dotted line). Expectations are shifted by one year such that the vertical distance to the black line captures the forecast errors. Household expectations are from the Survey of Consumers from the University of Michigan and firm expectations from the Survey of Firms' Inflation Expectations, provided by the Cleveland Fed.

In this paper, I develop a framework of inflation expectations featuring state-dependent attention levels where these joint dynamics of inflation, expectations, and attention arise endogenously. This model delivers two key implications: (i) regime shifts in the responsiveness of inflation expectations to inflation, and (ii) asymmetric inflation and inflation-expectations dynamics, including a persistent overshooting of expectations and a long "last mile" of disinflation back to target. When inflation exceeds an attention threshold, people pay more attention, expectations become more sensitive to

[1] For example, Federal Reserve Chair Jerome H. Powell said in his 2021 Jackson Hole speech that inflation concerns are "likely to prove temporary" (see https://www.federalreserve.gov/newsevents/speech/powell20210827a.htm).

inflation, and these expectations can reinforce inflationary surges and make them more persistent. When inflation falls back below the threshold, attention declines, expectations become less responsive to disinflationary news, so expectations fall only gradually, and the economy experiences a prolonged last mile. The recent inflation episode in Figure 1 provides a salient example of these dynamics.

To arrive at these results, I propose a tractable model of inflation expectations featuring state-dependent attention reflected in an endogenous updating gain. Attention captures how strongly agents update their short-run inflation expectations in response to forecast errors.[2] In the theoretical framework that motivates my empirical approach, agents decide optimally how much attention to inflation they want to pay. Their prior uncertainty increases as inflation increases and after inflation exceeds a certain threshold, agents decide to increase their attention to a higher level at which attention stays again constant even when inflation increases further. The threshold is endogenous and depends on the underlying fundamentals of the attention problem. The theory thus predicts that attention switches between two regimes, depending on whether inflation is below or above a certain threshold.

The tractability of this framework allows me to bring the model to the data directly and I can estimate the attention threshold and the attention levels in the two regimes jointly using survey data from the University of Michigan's Survey of Consumers for the period 1978 to 2024. I estimate an inflation attention threshold at an annualized inflation rate of about 4%, and that attention is about 0.18 when inflation is below the threshold and practically doubles to 0.35 when inflation is above the threshold. An attention level of 0.35 means that following a 1pp forecast error, agents increase their inflation expectations by 0.35pp. I find that the data suggests that there is only one attention threshold, that the threshold specification is preferred to smooth changes in attention, and that once I control for this threshold, I do not find any evidence that attention increases with inflation *within* regime. These results are robust across different specifications and data sets.

The attention regimes matter for the dynamics of inflation and inflation expectations. Using the high-frequency identification of Känzig (2021), I use oil supply news shocks as a proxy for supply shocks and I find that inflation increases more than twice as much in response to these shocks when they hit in the high-attention regime compared to the low-attention regime. A one standard deviation shock that pushes inflation up by about 20 basis points on average across regimes, leads to a peak increase in inflation of about 40 basis points when the shock hits in the high-attention regime. These effects are persistent as they only diminish after about one to two years. Using regional data and alternative measures of people's attention, I further disentangle the effects of attention from other changes in the economy that may arise from higher inflation, and find that attention indeed leads to a stronger inflation response to supply shocks even when controlling for inflation directly.

Applying my results to the recent inflation surge, I find that the estimated attention threshold implies that the U.S. economy entered the high-attention regime most recently in early 2021. To

[2] For example, Werning (2022) discusses why short-run expectations are more important than long-run expectations for inflation. In Pfäuti (2025), I show that measuring attention to inflation by estimating how agents update their short-run expectations correlates strongly with other measures of attention.

quantify the importance of this increased attention for the recent inflation surge, I first show that the oil supply news shocks explain about 50% of the inflation surge from the time the U.S. economy entered the high-attention regime in early 2021 until the peak in mid-2022. Without the attention increase, however, the implied inflation increase would have only been half as strong.[3] These results suggest that people's higher attention to inflation is not merely a byproduct but an important driver of sudden and persistent inflation surges.

To understand these findings as well as their implications for monetary policy, I develop a New Keynesian model accounting for the attention threshold and the changing levels of attention. I find that in response to an inflationary supply shock that pushes inflation above the attention threshold, inflation keeps on increasing due to the attention increase. As attention increases, agents update their inflation expectations more strongly in response to the increase in inflation. These higher inflation expectations then fuel further inflation increases, leading to even higher inflation expectations. This illustrates how what normally would be a transitory inflation shock can become a very persistent one due to the heightened attention. As the shock dies out, inflation starts to decline after some time. However, inflation may remain elevated for a substantial period of time, because once it falls back below the threshold, people pay less attention to inflation, and hence, only slowly revise their inflation expectations downward. As their prior expectations are now relatively high due to the experienced high inflation period, expectations remain persistently high, leading to a slow decline of actual inflation. The inflation attention threshold and changing degrees of attention hence offer a joint explanation for why inflation surges tend to be hump-shaped, and why inflation expectations initially fall short of actual inflation, but then surpass it once inflation starts to decline (as documented in panel (b) of Figure 1, and consistent with Angeletos et al. (2020b) and Blanco et al. (2025) more broadly). Therefore, changes in attention to inflation can explain why inflation may come down relatively quickly at first, but then remains stubbornly above the central bank's inflation target.

That inflation expectations stay persistently high, even when inflation has already fallen back below the threshold, further gives rise to a heightened risk of another subsequent inflation surge. The higher inflation expectations keep actual inflation higher for longer, and therefore, closer to the attention threshold. Thus, a subsequent inflationary shock is more likely to push inflation back above the threshold and therefore, leading to another episode of persistently high inflation, as for example in the 1970s and 1980s in the U.S.

The attention threshold induces an asymmetry in the dynamics of inflation: attention increases when inflation is particularly high but remains constant when inflation is particularly low. Relative to the model abstracting from the attention threshold, this asymmetry increases the risk of periods of persistently high and volatile inflation rates but leaves the risk of deflation largely unchanged. This inflation asymmetry induced by the attention threshold also offers a potential explanation for why we did not observe long lasting deflationary periods, for example, during the Great Financial

[3] I show that also other shocks, i.e., the *inflation shock* from Angeletos et al. (2020a) as well as monetary policy shocks, become more inflationary in the high-attention regime, indicating that shocks generally become more inflationary when attention to inflation is high.

Crisis (Coibion and Gorodnichenko, 2015b).[4] Using a wide range of alternative models, I show that only the model with the attention threshold can account for the documented empirical patterns of inflation and the asymmetry between periods of high vs low inflation.

This asymmetry also matters for the normative implications of the attention threshold. When monetary policy sets the nominal interest rate following a relatively dovish Taylor rule, the associated central bank losses are significantly larger compared to optimal policy rules or a strict inflation targeting rule. The reason is that under a standard Taylor rule, the economy spends a substantial amount of time in the high-attention regime in which inflation is high and volatile. Being frequently in the high-attention regime also increases the *average level of inflation* due to the asymmetry of the threshold which is costly for the central bank. By following a more hawkish monetary policy rule, the central bank can mitigate these losses.

Related literature. This paper contributes to the literature on inflation expectations formation and, in particular, on the role of limited and time-varying attention for expectations and their macroeconomic implications. A central insight from the macroeconomic literature on the role of expectations is that macroeconomic dynamics depend not only on the level of expectations, but also on how *responsive* expectations are to incoming information: in most of these models, an updating gain governs the responsiveness of expectations (Evans and Honkapohja, 2001; Eusepi and Preston, 2011; Marcet and Nicolini, 2003; Gáti, 2023; Evans and Ramey, 1995; Morris and Shin, 2002; Jørgensen and Lansing, 2024). Within this literature, the paper is most closely related to work on *endogenous gains* and *unanchoring* of inflation expectations. Existing contributions study how the degree of anchoring affects inflation dynamics and optimal policy, either through continuously varying gains (Gáti, 2023) or through discrete changes in anchoring (Carvalho et al., 2023). A key contribution of the present paper is to provide a tractable, attention-based microfoundation for an endogenous gain that changes *discretely* across regimes through an inflation attention threshold, delivering (i) regime shifts in responsiveness and (ii) asymmetric inflation-expectations dynamics, including a long "last mile" of disinflation.

More broadly, the paper relates to models of costly information acquisition and imperfect information updating, including rational inattention and sticky expectations (Sims, 2003, 2010; Reis, 2006a,b; Lorenzoni, 2009; Mackowiak and Wiederholt, 2009; Paciello and Wiederholt, 2014; Afrouzi and Yang, 2021), as well as (potentially state-dependent) behavioral inattention (Gabaix, 2020). In contemporaneous work, Bianchi et al. (2024) develop a model of *smooth* diagnostic expectations in which expectations may overreact more in uncertain times. In contrast, the present paper emphasizes discrete regime shifts in attention and their macroeconomic implications.

Pfäuti (2025) documents that before the recent inflation surge, attention to inflation was at a

[4] In fact, annualized quarter-on-quarter CPI inflation was negative only 7% of the time between 1978 and 2023, whereas it exceeded the attention threshold about 30% of the time during the same period.

historical low.[5] Bracha and Tang (2025) find that when inflation increases, attention to inflation increases as well. Korenok et al. (2023) find, for a large number of countries, that people's attention to inflation increases with inflation only after inflation exceeds a certain threshold. Using Google search data for the period from 2004 to 2022, they estimate this attention threshold to be at an annualized inflation rate of 3.55% for the U.S. Buelens (2025) uses internet search behavior of households in the Euro Area to show that this form of attention to inflation positively depends on the level of inflation. Cavallo et al. (2017) show that survey respondents in high-inflation environments (Argentina) respond less to information about inflation than in low-inflation environments (United States), which is consistent with higher attention to inflation in times of high inflation. Weber et al. (2025) confirm, using a range of randomized control trials spanning over several years and different countries, that attention of households and firms is indeed higher in times of high inflation. Kroner (2024) focuses on financial markets and shows that attention—measured as asset price responses to inflation news—is higher in times of high inflation. My key innovation relative to these papers is that I provide estimates of the attention threshold and the attention levels in the two regimes in a way that directly maps into otherwise standard macroeconomic models, and I quantify the macroeconomic implications of these attention changes.

I further contribute to the literature on the state dependency of shocks (Auerbach and Gorodnichenko, 2012a,b; Caggiano et al., 2014; Ramey and Zubairy, 2018; Tenreyro and Thwaites, 2016; Aastveit et al., 2023; Jo and Zubairy, 2025). The two papers most closely related are Ascari and Haber (2022) and Joussier et al. (2023). Ascari and Haber (2022) show that the inflationary effects of monetary policy shocks are larger in times of high trend inflation. Joussier et al. (2023) find that French firms pass through energy price shocks more to their prices in times of high inflation.[6] I complement this work by highlighting a specific expectations channel—state-dependent attention to inflation—through which both supply and monetary shocks propagate more strongly when inflation is high, and by showing that the same mechanism can generate an asymmetric return to target that lengthens the "last mile" of disinflation.

This paper also contributes to the ongoing debate about the drivers of the recent inflation surge and offers an alternative and complementary perspective by highlighting the role of people's heightened attention during that period. The existing literature emphasizes supply-side factors and

[5] For the years before the recent inflation surge, Candia et al. (2024, 2023) and Coibion et al. (2023) show that U.S. firms as well as households are usually poorly informed about and inattentive to inflation and monetary policy (see also Weber et al. (2022) for a recent review). Goldstein (2023) also finds that inattention varies over time. Different forms of changing attention are considered, e.g., in Coibion and Gorodnichenko (2015a) who show state-dependence in the degree of information rigidity (focusing on GDP growth), or in Kim and Binder (2023) who examine how repeat participation in surveys that ask about inflation expectations may lead to higher attention to inflation, or in Flynn and Sastry (2022) who show—by using a textual proxy for firms' attention toward macroeconomic conditions—that attention is counter-cyclical, or Gallegos (2023) who shows that firms' less sluggish inflation expectations after the Great inflation period offer a potential explanation for the decrease in the persistence of inflation as found, e.g., in Benati (2008). York (2023) exploits differences in survey dates to examine how inflation expectations respond to macroeconomic news releases.

[6] Based on the present paper and Korenok et al. (2023), Joussier et al. (2023) use Google Trends data to proxy for attention to energy prices and energy cost and show that the pass through of energy shocks to prices in France is stronger in times of higher attention to energy prices.

bottlenecks (Shapiro, 2024; Bernanke and Blanchard, 2025; Amiti et al., 2023; Gagliardone and Gertler, 2023), fiscal forces and beliefs about the fiscal framework (Bianchi and Melosi, 2022; Bianchi et al., 2023), nonlinearities and labor-market mechanisms (Cerrato and Gitti, 2022; Gitti, 2024; Benigno and Eggertsson, 2023; Ball et al., 2022; Lorenzoni and Werning, 2023; Crump et al., 2024; Erceg et al., 2024), and changes in price adjustment behavior (Cavallo et al., 2024; Blanco et al., 2024). In contrast to these papers, I focus on the role of changes in attention to inflation. Attention serves as a complementary propagation mechanism rather than as an exogenous impulse, and offers an explanation for the observed interplay between inflation and inflation expectations (panel (b) in Figure 1) as well as for the long last mile of disinflation.[7]

Outline. In Section 2, I develop the theoretical model of state-dependent attention and estimate the inflation attention threshold and attention across regimes. In Section 3, I show that the attention regime matters for the inflationary effects of shocks, and quantify the role of attention for these effects. I introduce the New Keynesian model with limited attention in Section 4, and the model's positive results in Section 5. In Section 6, I discuss the threshold's normative implications and Section 7 concludes.

2 Attention and the Inflation Attention Threshold

In this section, I present a model of inflation expectations formation with limited and state-dependent attention. Attention governs how strongly agents update their inflation expectations when receiving new information. When inflation is low, agents update their inflation expectations relatively little and place more weight on their prior beliefs, whereas they update their expectations more strongly when inflation is high. I then bring this model to the data to estimate how and when attention to inflation changes. I relegate detailed derivations and microfoundations to Appendix A.

2.1 Attention to inflation

In order to estimate how and when attention to inflation changes, I extend the method I develop in Pfäuti (2025). The agent believes that inflation in the next period, π', depends on inflation today, π, as follows:

$$\pi' = \tilde{\rho}_\pi \pi + \nu,$$

where $\nu \sim i.i.N.(0, \sigma_\nu^2)$, and $\tilde{\rho}_\pi \in [0, 1]$ denotes the perceived persistence of inflation.[8] Inflation in the current period is unobservable, so before forming an expectation about future inflation, the agent

[7] See Binder and Kamdar (2022) for a narrative history of how inflation and inflation expectations evolved jointly during several recessions and periods of high inflation since the Great Depression.

[8] That inflation expectations solely depend on perceived inflation is supported by a large empirical literature documenting that perceived inflation is indeed the key predictor for inflation expectations (see, e.g., Jonung (1981) and Weber et al. (2022) for households and Candia et al. (2023) for firms). That said, however, I also allow for a multivariate expectations formation process in the estimation later.

needs to form expectations about today's inflation. Given this perceived law of motion, the full-information forecast is given by $\tilde{\rho}_\pi \pi$. But because π is not perfectly observable, the actual forecast will deviate from this full-information forecast. These deviations are costly and the agent can pay more attention to reduce these deviations. Thus, the agent's choice is not only about how to form her expectations given certain information, but about how to choose and process this information optimally, while taking into account how this will later affect her forecast.

The agent's prior is assumed to be Gaussian; $\pi \sim N\left(\pi^{prior}, \sigma_\pi^2\right)$ with π^{prior} denoting the prior mean and σ_π^2 the prior uncertainty. I assume that the agent's prior uncertainty, σ_π^2, increases smoothly (strictly monotonically) with the level of inflation. This assumption reflects the strong positive empirical correlation between the inflation level and its volatility. In my baseline sample, for example, the correlation between the average level of inflation and its volatility, computed within five-year rolling windows, is 0.76. The agent knows σ_π^2, but she cannot back out the inflation rate from observing σ_π^2.[9]

The signals the agent obtains are Gaussian:

$$s = \pi + \varepsilon,$$

with noise $\varepsilon \sim i.i.N.(0, \sigma_\varepsilon^2)$. Attention is inversely related to the noise variance σ_ε^2 which is chosen optimally. Hence, paying more or less attention is akin to choosing whether to acquire more or less precise signals.

Given the Gaussian prior and signal, the optimal forecast, π^e, takes the form:

$$\pi^e = \tilde{\rho}_\pi \left(1 - \gamma_\pi\right) \pi^{prior} + \tilde{\rho}_\pi \gamma_\pi s, \tag{1}$$

where $\gamma_\pi \in [0,1]$ measures the agent's attention to inflation, and I refer to it as attention or the updating gain, interchangeably.[10] Note that γ_π is akin to a Kalman gain where this gain is chosen optimally as it reflects the agent's attention choice. Hence, the results in this section can be viewed as an attention-based microfoundation for models of endogenous or time-varying updating gains and unanchoring (as, e.g., in Gáti (2023)). When the agent is inattentive, she obtains relatively noisy signals and thus, puts little weight on these signals, reflected in a small γ_π. Therefore, lower attention, i.e., a lower γ_π, implies that the agent updates her expectations to a given signal s less strongly and instead puts more weight on her prior beliefs π^{prior}.

Equation (1) describes how agents form their inflation expectations for a *given* level of attention

[9]Note that σ_π^2 is the agent's subjective uncertainty, which is part of the agent's information set, whereas the realization of current inflation, π, is not perfectly observed. As in rational-inattention models more broadly, I take the agent's perceived distribution of inflation (including its second moment) as given when attention is chosen. This abstracts from learning or uncertainty about uncertainty itself. Consistent with my model, Li et al. (2025) show that inflation uncertainty increased substantially during the recent inflation surge.

[10]The updating gain γ_π captures the signal-to-noise ratio and can be written as $\gamma_\pi = 1 - \frac{\sigma_{\pi|s}^2}{\sigma_\pi^2} \in [0,1]$, where σ_π^2 denotes the agent's prior uncertainty about inflation, and $\sigma_{\pi|s}^2$ the posterior uncertainty given the signal s.

γ_π. Attention, however, is a choice and the agent chooses γ_π optimally, given the current inflation environment and the costs and benefits of paying attention. As I detail in Appendix A, I assume that the agent obtains a certain level of attention for free, which I denote $\gamma_{\pi,L}$, where L denotes that this is a relatively Low level of attention. Increasing attention beyond this level allows the agent to predict inflation more accurately and hence, the agent makes smaller mistakes. Increasing her attention, however, is costly and the agent solves this trade off optimally. In contrast to most models of rational inattention (see, e.g., Maćkowiak et al., 2023, for a recent review), the agent in my model can choose between paying either a fixed cost, in which case her attention jumps to a Higher level of attention, $\gamma_{\pi,H} > \gamma_{\pi,L}$, or paying a standard mutual information cognitive cost and adjust her attention to inflation smoothly.[11]

As inflation increases, the agent's prior uncertainty increases. Therefore, the cost of being inattentive increases with the level of inflation and the agent becomes more likely to increase her attention. Under conditions I derive in Appendix A, the agent's optimal attention choice takes on the behavior shown in Figure 2.[12] The figure shows the optimal attention level γ_π^* on the vertical axis at different levels of inflation on the horizontal axis.

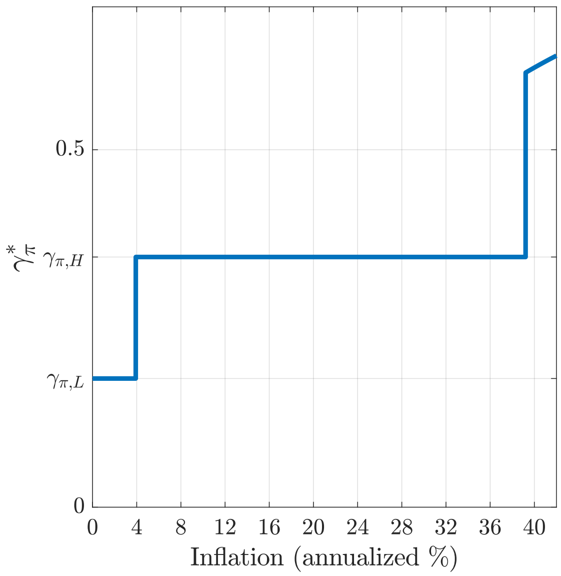
The figure illustrates how agents optimally allocate their attention to inflation. At relatively low levels of inflation, agents do not pay more attention than what they obtain for free. The reason is that their prior uncertainty is low in these times and hence, the benefit of increasing their attention is less than the costs of doing so. Once inflation exceeds a threshold, $\bar{\pi}$, the agent pays the fixed cost and her attention jumps to $\gamma_{\pi,H}$. Attention then stays constant at $\gamma_{\pi,H}$ even at higher levels of inflation. Up until a second threshold at which attention jumps once more after which it increases smoothly with inflation. This second jump mirrors the agent's decision to pay the mutual information cost instead of the fixed cost. The reason is that with the fixed cost, the agent's attention is stuck at $\gamma_{\pi,H}$ and hence, the agent's losses from not being fully attentive increase as inflation increases. Once these losses are large enough, the agent decides to pay the mutual information cost which allows her to increase attention even further. The empirical results in the following section suggest, however, that this last part has not been relevant in the U.S. since the beginning of my sample. Nevertheless, it illustrates that attention may increase even further if inflation reaches these very high levels.

The state-dependent attention gain implies an asymmetry in the agent's expectation formation: if inflation and prior expectations are low, attention is low and expectations only update slowly in response to new signals which means that expectations remain relatively low. If inflation is high, attention rises and expectations are updated strongly. Hence, expectations may increase quickly when inflation itself is high. If inflation declines after a period of high inflation and people become

[11] Mutual-information costs have an axiomatic foundation, but adding a fixed-cost attention option means the overall cost structure is best viewed as a hybrid menu of information technologies and thus need not satisfy all axioms that characterize pure mutual-information costs. This allows me to solve for attention and the attention threshold in closed form, and Section 2.2 shows that the data favor the threshold specification.

[12] In Appendix A, I provide all the details necessary to derive Figure 2 and characterize the threshold analytically. One important condition that needs to hold is that the fixed cost the agent needs to pay for an attention level of $\gamma_{\pi,H}$ must be relatively low compared to the mutual information cost.

Figure 2: Optimal attention



Notes: This figure shows optimal attention to inflation, γ^*_π, as a function of the level of inflation π. The attention levels $\gamma_{\pi,L}$ and $\gamma_{\pi,H}$ are at 0.18 and 0.35 to be consistent with the empirical findings in Table 1.

less attentive, updating becomes slower again. However, if people's prior beliefs are higher after an inflationary episode, the slower updating implies that inflation expectations remain stubbornly higher.

Dynamic model. To bring the attention model to the data, I now introduce dynamics. The agent believes that inflation π follows:

$$\pi_t = (1 - \tilde{\rho}_\pi)\underline{\pi} + \tilde{\rho}_\pi \pi_{t-1} + \nu_t,$$

where $\underline{\pi}$ is the agent's long-run belief about inflation. I assume that the error term ν_t is normally distributed with mean zero and variance σ^2_ν. The agent receives a signal about inflation of the form:

$$s_t = \pi_t + \varepsilon_t,$$

where the noise ε_t is normally distributed with variance σ^2_ε. As in the static model, higher attention is reflected in less noise, i.e., a lower σ^2_ε.

Figure 2 suggests that there are two attention regimes: one with low attention and low inflation, and one with high attention and high inflation. As I show in the following, this two-regime specification is the specification that the data prefers, relative to specifications with more than two regimes or smooth changes in attention. I use subscripts r to denote (potentially) regime-dependent parameters. Note, however, that my estimation approach does not impose that attention has to be higher when inflation is higher.

I assume that agents always use the steady state Kalman filter in each regime $r \in \{L, H\}$, so that optimal updating in a given regime r is given by:

$$\pi^e_{t+1|t} = (1-\tilde{\rho}_{\pi,r})\underline{\pi}_r + \tilde{\rho}_{\pi,r}\pi^e_{t|t-1} + \tilde{\rho}_{\pi,r}\gamma_{\pi,r}\left(\pi_t - \pi^e_{t|t-1}\right) + \eta_t, \tag{2}$$

where $\gamma_{\pi,r}$ captures how attentive the agent is.[13] Explicitly accounting for the attention threshold $\bar{\pi}$, agents form their one-period ahead inflation expectations according to:

$$\pi^e_{t+1|t} = \begin{cases} (1-\tilde{\rho}_{\pi,L})\underline{\pi}_L + \tilde{\rho}_{\pi,L}\pi^e_{t|t-1} + \tilde{\rho}_{\pi,L}\gamma_{\pi,L}\left(\pi_t - \pi^e_{t|t-1}\right) + \eta_t, & \text{when } \pi_{t-1} < \bar{\pi} \\ (1-\tilde{\rho}_{\pi,H})\underline{\pi}_H + \tilde{\rho}_{\pi,H}\pi^e_{t|t-1} + \tilde{\rho}_{\pi,H}\gamma_{\pi,H}\left(\pi_t - \pi^e_{t|t-1}\right) + \eta_t, & \text{when } \pi_{t-1} \geq \bar{\pi}. \end{cases} \tag{3}$$

Here, $\pi^e_{t+1|t}$ denotes expectations about inflation in period $t+1$ based on time-t information. Note that I allow the long-run mean belief, $\underline{\pi}_r$, the perceived persistence, $\tilde{\rho}_{\pi,r}$, and attention, $\gamma_{\pi,r}$, to differ across regimes.[14] I discuss alternative specifications and robustness checks at the end of the section.

2.2 Estimating attention and the attention threshold

Data. As my baseline measure of inflation expectations, I rely on the Survey of Consumers from the University of Michigan. I use average and median household inflation expectations for the period 1978M1-2024M5. For the period 2013-2024, I also use individual inflation expectations from the Survey of Consumer Expectations from the New York Fed. Even though I focus on inflation and inflation expectations over one quarter, I use monthly data to increase the number of observations. As a robustness check, however, I also consider expectations at quarterly frequency. One advantage of using quarterly observations is that the Survey of Consumers provides mean expectations going back to 1960Q2. Both surveys ask consumers for their price growth expectations one-year ahead: $\pi^e_{t+12|t}$. I transform them into one-quarter-ahead forecasts as follows: $\pi^e_{t+3|t} \equiv \frac{\pi^e_{t+12|t}}{4}$. Computing one-quarter-ahead expectations allows me to compare the results directly to the model which is calibrated at quarterly frequency. For actual inflation, I use the monthly CPI inflation rate from the FRED database and to be consistent with the model, I focus on quarter-on-quarter inflation: $\pi_t \equiv \frac{P_t - P_{t-3}}{P_{t-3}}$. As I show in Appendix B, the results are qualitatively similar when using one-year-ahead expectations and year-on-year inflation as the actual measure of inflation.

Estimating attention and the attention threshold. In order to estimate the two attention levels $\gamma_{\pi,r}$ for $r \in \{L, H\}$, as well as the attention threshold $\bar{\pi}$, I estimate the following threshold

[13] Assuming that agents always use the steady state Kalman filter is a standard assumption in the rational inattention literature and basically means that the agent receives all her signals before forming her expectations (see, e.g., Mackowiak and Wiederholt (2009); Maćkowiak et al. (2018, 2023)). This leaves conditional second moments time-invariant and thus, the optimal level of attention constant. An important recent paper that relaxes this assumption is Afrouzi and Yang (2021). However, they focus on the linear-quadratic-Gaussian case which in my case is not applicable due to the non-linearity introduced by the attention threshold.

[14] Note that $\eta_t = \tilde{\rho}_{\pi,r}\gamma_{\pi,r}\varepsilon_t$. Through the lens of the rational inattention model, ε_t, and thus, η_t, are uncorrelated with the independent variables.

regression:

$$\pi^e_{t+3|t} = \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\left(\beta_{0,L} + \beta_{1,L}\pi^e_{t|t-3} + \beta_{2,L}\left(\pi_t - \pi^e_{t|t-3}\right)\right) + \left(1 - \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\right)\left(\beta_{0,H} + \beta_{1,H}\pi^e_{t|t-3} + \beta_{2,H}\left(\pi_t - \pi^e_{t|t-3}\right)\right) + \epsilon_t, \tag{4}$$

where $\beta_{0,r} = (1 - \tilde{\rho}_{\pi,r})\underline{\pi}_r$ denotes the intercept in regime $r \in \{L, H\}$, $\beta_{1,r} = \tilde{\rho}_{\pi,r}$ the perceived persistence, and $\frac{\beta_{2,r}}{\beta_{1,r}} = \gamma_{\pi,r}$ the attention level in regime r. I therefore not only allow the attention parameters to differ across regimes, but also the perceived persistence of inflation as well as agents' long-run beliefs about inflation. $\mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}$ is the indicator function that equals one when inflation in the previous month π_{t-1} was below the threshold $\bar{\pi}$ and zero otherwise. The threshold value is then estimated jointly with the regression coefficients by minimizing the sum of squared residuals obtained for all possible thresholds (see, e.g., Gonzalo and Pitarakis (2002) and Hansen (2011)). Note that I do not impose that the attention level in the regime in which inflation is above the threshold needs to be higher than in the regime with inflation below the threshold.

Estimation results. Table 1 shows the estimation results. For the baseline specification, I estimate an attention threshold of $\hat{\bar{\pi}} = 3.91\%$. Attention in the regime in which inflation is below this threshold is $\hat{\gamma}_{\pi,L} = 0.18$, and attention in the high-inflation regime is equal to $\hat{\gamma}_{\pi,H} = 0.35$. Thus, attention in the high-inflation regime is higher than in the low-inflation regime, and I therefore refer to the regime with inflation above the threshold as the *high-attention regime*. The threshold value of 3.91% implies that the U.S. economy spent about 32% of the time between 1978 and 2024 in the high-attention regime. Most recently, inflation exceeded this threshold in early 2021. Google searches also started to increase around that time (see panel (a) in Figure 1 in the Introduction). These findings are consistent with the theoretical predictions in Figure 2.

To test for multiple thresholds, I use the Bayesian Information Criterion (BIC) to select the number of thresholds, and I find that the data prefer the specification with only one threshold.[15]

Table 1: Estimated attention levels and the attention threshold

	Threshold $\hat{\bar{\pi}}$	Low Att. $\hat{\gamma}_{\pi,L}$	High Att. $\hat{\gamma}_{\pi,H}$	p-val. $H_0 : \gamma_{\pi,L} = \gamma_{\pi,H}$
Mean expectations	3.91%	0.18	0.35	0.000
s.e.		(0.018)	(0.042)	

Notes: This table shows the results from regression (4), where $\hat{\bar{\pi}}$ denotes the estimated threshold, $\hat{\gamma}_{\pi,L}$ and $\hat{\gamma}_{\pi,H}$ the estimated attention levels when inflation is below or above the threshold, respectively. The last column shows the p-value for the null hypothesis that the two attention levels are equal. Standard errors for the attention estimates are computed using the delta method, based on Newey–West heteroskedasticity- and autocorrelation-robust covariance estimates with 12 lags.

[15] For example, the BIC increases from -534.12 to -524.70 when going from one to two thresholds and to -512.03 when going to three thresholds. The BIC is given by $T\ln(SSR/T) + k\ln(T)$, where k is the number of parameters, T the number of periods, and SSR the sum of squared residuals. As a robustness check, I also consider the Hannan-Quinn IC and find that also the HQIC selects the model with one threshold only.

2.3 Robustness

I now discuss the robustness of these results. I summarize the additional tests here briefly and relegate a detailed discussion of all these results to Appendix B.

Median expectations. The results are robust to using median expectations instead of average expectations. The estimated threshold value is somewhat higher, and the attention gains lower. The lower attention gains are consistent with the fact that median expectations tend to be less volatile than mean expectations.

Quarterly frequency. When I run regression (4) at quarterly frequency instead of monthly frequency, I obtain a similar though slightly lower threshold value, and attention estimates that are very close to my baseline estimates.

Alternative threshold-defining variables. I show that using current inflation or inflation volatility instead of last month's inflation rate as the threshold-defining variable leads to similar results. I also show, however, that the BIC prefers the level, not the volatility, as the threshold-defining variable.

Other surveys. I also consider individual household expectations from the New York Fed Survey of Consumer expectations, as well as firm expectations from the Survey of Firms' Inflation Expectations from the Cleveland Fed. Even though the samples for these surveys are much shorter, the findings are consistent with the baseline results in Table 1. Professional forecasters, on the other hand, behave quite differently than households and firms.

Regional variation. The Survey of Consumers also provides the regions consumers are in, which allows me to leverage regional variation to test for the robustness of my results. In particular, I use the four census regions West, North East, North Central (or Midwest) and South. Doing so, and controlling for region-specific fixed effects leaves my baseline estimates largely unchanged.[16]

Alternative attention measure. I consider an alternative measure of attention to inflation, following Bracha and Tang (2025). I extend their estimation until 2024 and estimate attention both at the aggregate as well as at the regional level. I find that my results are robust to this alternative attention measure by showing that the implied threshold is similar to the one estimated in Table 1, and that the differences in attention across regimes are significant. I further show that the data prefer having exactly one threshold, and that controlling additionally for inflation does not affect

[16]For this, I use the following series from the FRED database: Consumer Price Index for All Urban Consumers: All Items in South (CUUR0300SA0), Consumer Price Index for All Urban Consumers: All Items in Northeast (CUUR0100SA0), Consumer Price Index for All Urban Consumers: All Items in West (CUUR0400SA0), Consumer Price Index for All Urban Consumers: All Items in Midwest (CUUR0200SA0).

these results and that the coefficients related to inflation are insignificant, supporting the two-regime specification.

A multivariate system of inflation expectations. The agent's perceived law of motion of inflation could in principle also depend on other variables instead of being an AR(1) process. To test whether this affects my results or not, I allow inflation expectations to also depend on unemployment expectations. I find that the coefficients related to unemployment are not significantly different from 0 and the coefficients related to inflation are largely unaffected.

Markov-switching model. An alternative approach to estimate regime-dependent attention gains is estimating a Markov-switching model where I do not need to take a stance on what defines the regimes. The results I obtain from following that approach are consistent with the ones reported in Table 1.

Smooth changes in attention. When I allow for additional terms in regression (4) capturing potential additional effects of past inflation or inflation volatility, I find that the data (according to the Bayesian Information Criterion) prefers my baseline specification. Additionally, I find that my baseline estimates are largely unchanged and that the additional coefficients are economically very small.

Attention changes within regime. I estimate a time series of attention gains and regress these attention gains on the level of inflation as well as on a dummy variable capturing the threshold. The dummy coefficient is highly statistically significant, whereas the other coefficients are not. These results further support the view that attention follows the threshold-like behavior, consistent with the theoretical predictions in Figure 2.

Long-run expectations. As I detail in Appendix B, I can also use long-run expectations to estimate the attention gains as well as the attention threshold. The findings I obtain when following that approach are very close to my baseline estimates in Table 1. Similarly, controlling for long-run expectations in regression (4) leaves my baseline results largely unchanged.

3 Inflation Dynamics Across Attention Regimes

In this section, I show that negative supply shocks have more inflationary and longer-lasting effects when attention is high, and that the increase in attention in early 2021 roughly doubled the inflationary effects of supply shocks in the subsequent inflation surge.

Supply shocks. As my empirical measure of supply shocks, I use the oil supply news shocks from Känzig (2021) for the period 1975M1-2022M12. In a first step, oil surprises are identified by using institutional features of the Organization of the Petroleum Exporting Countries (OPEC) and high-frequency data on variation in oil futures prices around OPEC announcements. In a second step, the resulting surprises are then used as an external instrument in an oil VAR, to identify a structural oil supply news shock. In the following, I show the responses to a negative shock that pushes up oil prices and lowers oil production. I refer to these shocks as oil news shocks, cost-push shocks or supply shocks interchangeably.

Attention regimes. Given the results in Section 2, I define the high-attention regime to be periods in which inflation in the previous month exceeded 3.91%. As an alternative indicator for the attention regimes, I also use Google Trends data, with the drawback that this data is only available since 2004. However, it will allow me to disentangle attention from other changes in the economy that tend to arise when inflation is higher (e.g., changes in the frequency of price adjustments, see, e.g., Blanco et al. (2024)). The Google data is normalized such that the month with the most Google searches of the word "inflation" is assigned a value of 100 and all the other months are expressed relative to that month. I assign months to the high-attention regime when Google searches of inflation in that month exceed the 75^{th} percentile.[17] As a robustness check, I also use the attention measure proposed by Bracha and Tang (2025) as an alternative.

Table 2: Shock properties across regimes

Regime	Mean	Standard deviation
Inflation as regime-defining variable		
High	-0.017	0.554
Low	0.008	0.550
Google Trends as regime-defining variable		
High	-0.007	0.685
Low	0.021	0.622

Notes: This table shows the mean and the standard deviation for the oil supply news shocks across the two attention regimes. The upper part of the table shows the case where the previous month's inflation is the threshold-defining variable and the lower part the case where Google Trends are used to identify the two regimes.

Table 2 lists the mean and the standard deviation of the shocks for the high-attention and low-attention regime, separately. The upper part of the table shows these statistics for the case in which the regimes are defined based on whether the previous month's inflation rate was below or above the 3.91% threshold, and the lower part of the table for the case where the regimes are defined based on Google Trends data. We see that in both cases, the shock series have similar first and second moments across regimes. When using Google Trends as the regime-defining variable, the volatility of the supply shocks is slightly higher in times of high attention. However, the following empirical results remain robust even when excluding shocks that are larger than one standard deviation. Thus,

[17] When I compute the pre-2020 (or also pre-2021) mean $\hat{\mu}_{google}$ and standard deviation $\hat{\sigma}_{google}$, and then estimate the 95th percentile under the assumption of an underlying normal distribution as $\hat{\mu}_{google} + 1.645 \cdot \hat{\sigma}_{google}$, the resulting value is practically identical to the 75th percentile of the full sample.

the following results are unlikely to be driven by differences in the shock series across regimes.[18]

The role of the attention regime for the propagation of supply shocks to inflation. To examine whether the effects of negative supply shocks on inflation differ across regimes, I estimate the local projection (Jordà, 2005):

$$\pi_{t+j} = \mathbb{1}_H \left(\alpha_j^H + \beta_j^H \varepsilon_t\right) + (1 - \mathbb{1}_H)\left(\alpha_j^L + \beta_j^L \varepsilon_t\right) + \Gamma' X_t + \epsilon_{t+j}, \tag{5}$$

where π_{t+j} denotes inflation at time $t + j$, $\mathbb{1}_H$ is an indicator function that equals one when the economy is in the high-attention regime at the time of the shock and 0 else, ε_t denotes the shock at time t and X_t are controls. In my baseline estimation, I use four lags of the shock, four lags of the unemployment rate, of inflation and of inflation expectations as controls. In Appendix C, I show that the results remain robust when using other controls.

Figure 3 shows the estimation results of regression (5). Panel (a) depicts the inflation response to a negative oil news shock in the high-attention regime, panel (b) in the low-attention regime, panel (c) shows the average effect, and panel (d) the difference between the effects in the high-attention regime and the low-attention regime. In all cases, I consider a one-standard deviation shock. The dark shaded areas depict the 68% confidence bands and the light-shaded areas the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Inflation increases about twice as much in the high-attention regime compared to the low-attention regime. These differences are quite persistent and statistically significant at the 10% significance level in the first six months. Thus, when the economy is hit by a one-standard deviation supply shock when people's attention to inflation is high, inflation increases on average by about 40 basis points. In contrast, when attention to inflation is low at the time of the shock, inflation only increases by about 20 basis points. Figure 14 in Appendix C shows that the results are very similar when using Google Trends as the regime-defining variable.

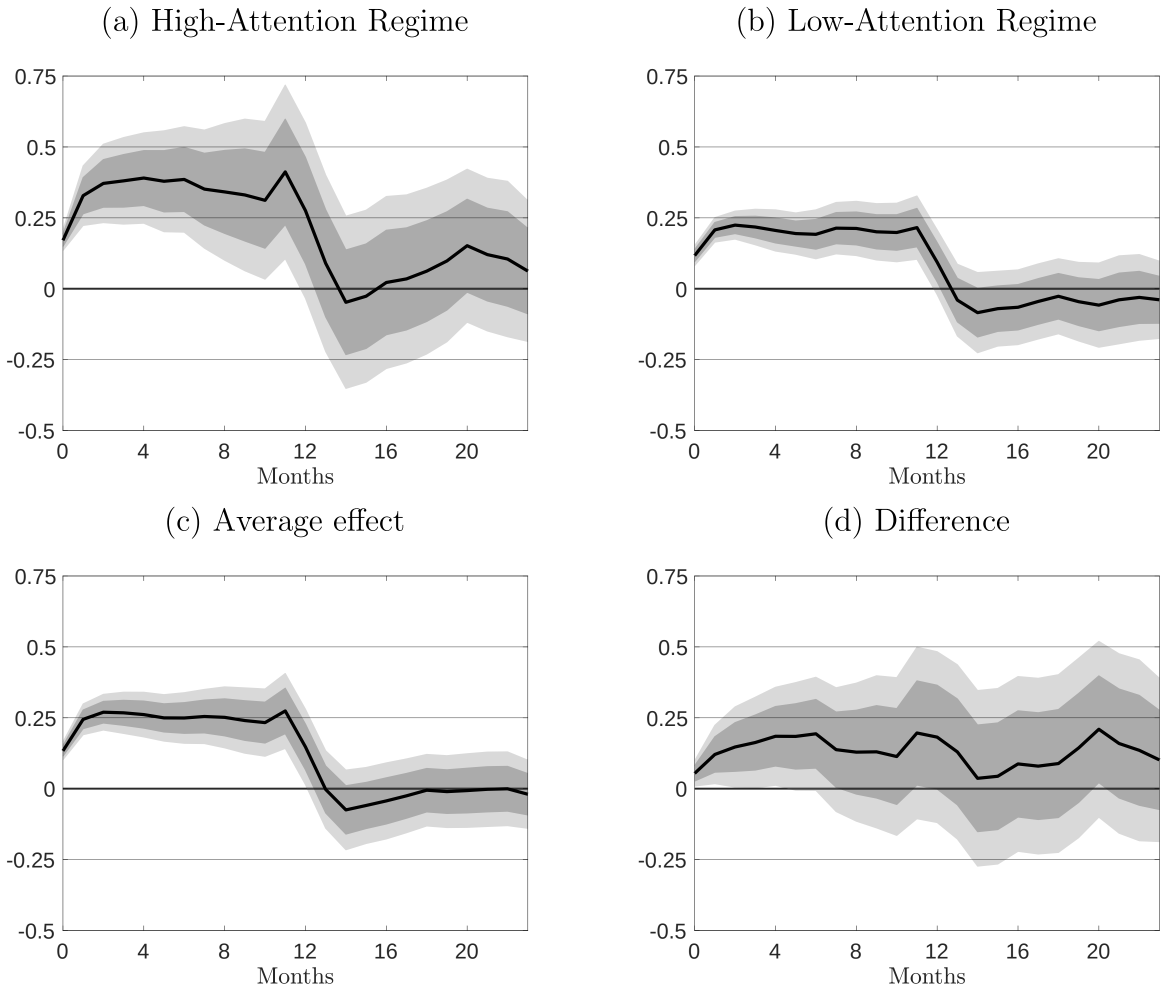
Regional variation. As in Section 2, I now also use regional variation of the four census regions. In particular, I use region-specific CPI inflation as the dependent variable. Additionally, I control for region fixed effects as well as four lags of region-specific inflation, plus interactions of those with the shock, to control for local shocks that may drive inflation independently of attention.

Figure 4 shows the results. We see that the results are very similar to my baseline specification even though the overall effects are slightly smaller and the confidence bands for the high-attention regime wider.

The regional data also helps me to disentangle attention from other phenomena that arise with higher inflation. To do so, I take two approaches. In the first one, I use Google Trends as the attention-

[18] Excluding large shocks also helps to address the issues that may arise with state-dependent local projections when the state is endogenous (Gonçalves et al., 2024).

Figure 3: Inflation response to an oil supply news shock

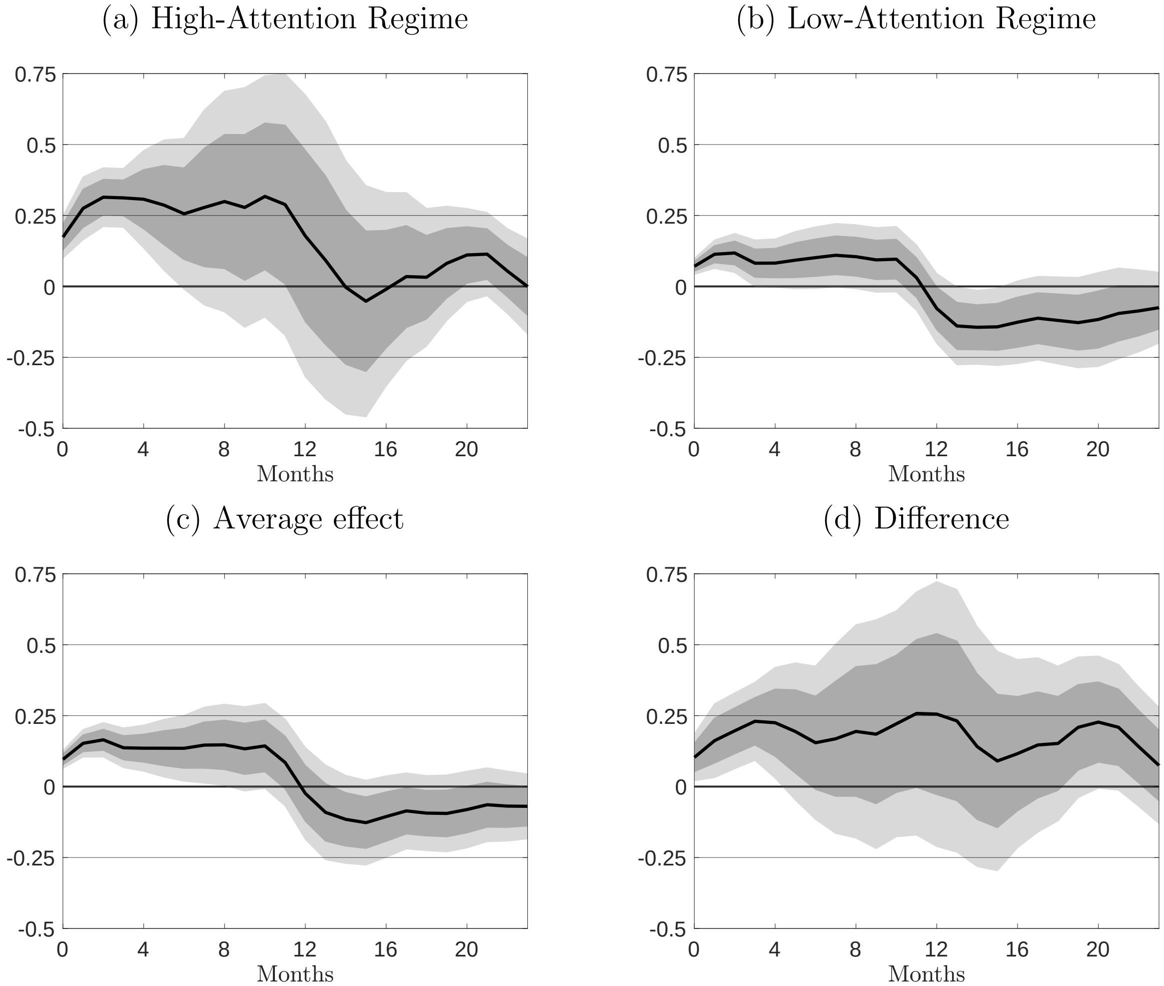


Notes: This figure shows the inflation response to an oil supply news shock (equation (5)) in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

regime defining variable for the four different regions.[19] To control for other phenomena apart from attention that may arise with inflation, I further include an interaction term of the shock with lagged inflation, and I include region fixed effects. Panels (a)-(c) in Figure 5 show the results. In particular, we see that adverse oil supply news shocks have a strong and persistent effect on inflation when attention is high, whereas the effects in times of low attention are substantially weaker. In addition, the interaction term with inflation is practically zero at all horizons. These results therefore suggest that attention to inflation renders supply shocks more inflationary even when controlling for inflation itself. The results in panels (a) and (b) also suggest that the inflation response is *qualitatively* different across the two regimes. While inflation peaks on impact in the low-attention regime, it shows more of a hump-shaped response in the high-attention regime. Following Coibion et al. (2017), I test the

[19] I take the average within each month across all states within a particular census region in order to aggregate Google Trends at the census region level.

Figure 4: Inflation response to an oil supply shock using regional variation



Notes: This figure shows the inflation response to an oil supply news shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)) when using regional variation and region-specific CPI inflation. The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to heteroskedasticity as well as serial and cross-sectional correlation.

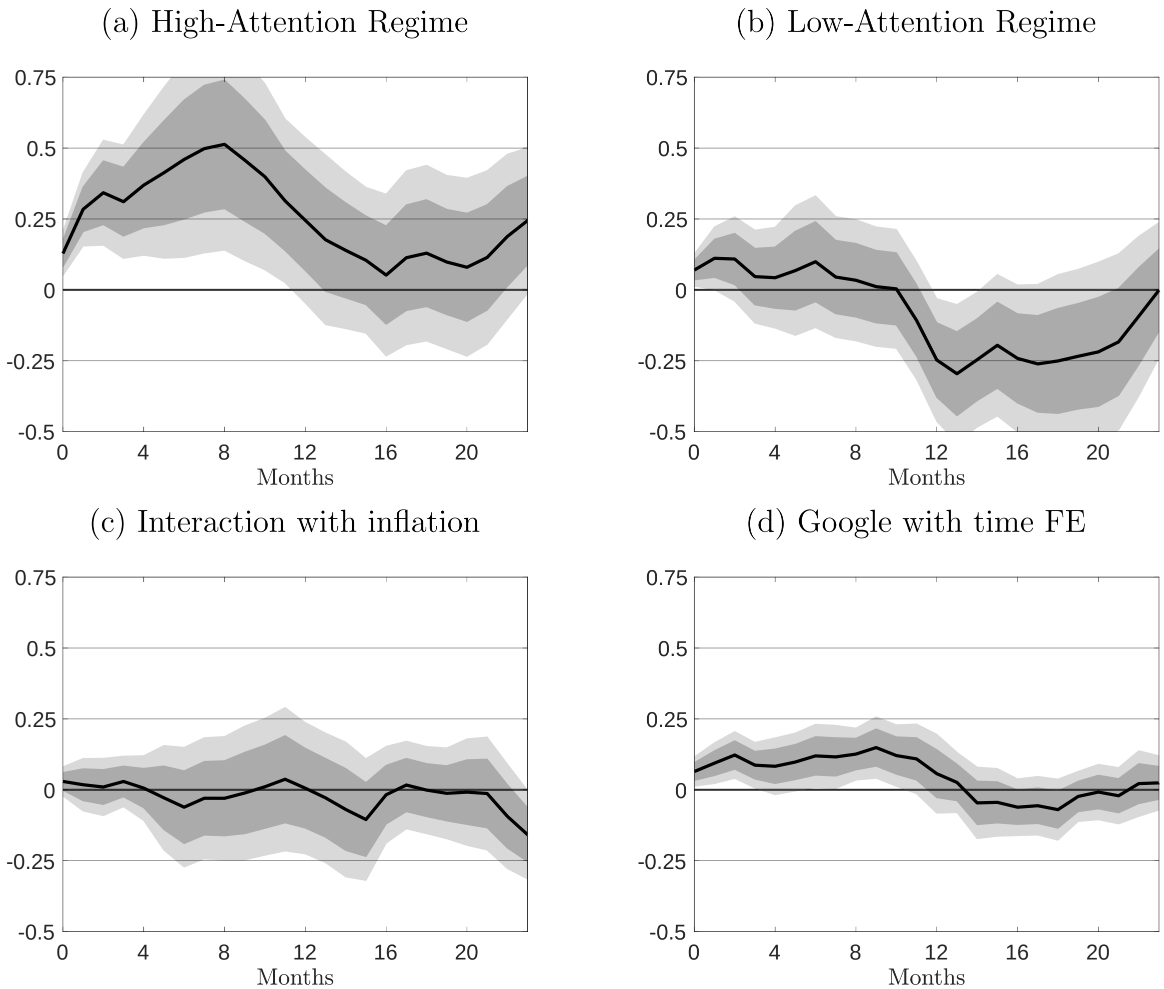
joint null hypothesis that the difference in impulse responses between the high- and low-attention regimes is zero at all horizons h = 0, ..., 24, and reject it with an associated p-value of 0.008.

In the second approach, I control for time and region fixed effects and for regional-specific inflation, as well as an interaction term of the shock with regional inflation in order to control for other local shocks that may affect inflation other than attention.[20] However, since the regimes based on Google Trends overlap strongly across regions—and are therefore largely absorbed by time fixed effects—I instead include an interaction term of last month's Google Trends searches with the shock rather than focusing on regimes. Panel (d) in Figure 5 shows that this interaction term is highly significant in the first 10 months (the coefficient is normalized to show the additional effect of a one standard deviation shock when Google Trends is one standard deviation above its mean). These results indicate that the different inflation responses across attention regimes are unlikely to be solely driven by aggregate

[20] I include four lags of Google trends, regional inflation, and regional inflation interacted with the shock.

phenomena that may confound the results, such as different policy responses across regimes, a general increase in the frequency of price adjustments, or an increase in the slope of the aggregate Phillips curve. Furthermore, controlling for regional fixed effects, inflation and inflation interacted with the shock, the findings in panel (d) are unlikely to be driven by local characteristics or local shocks and changes that affect inflation through other channels than attention. Instead, the significant effects arising from more Google searches of inflation indicate that higher attention indeed drives at least partly the different inflation responses observed across regimes.

Figure 5: Disentangling attention and inflation



Notes: This figure shows the inflation response to an oil supply news shock in the high-attention regime (panel (a)) and the low-attention regime (panel (b)) when using regional CPI data and when controlling for the interaction of the shock with last-month's inflation (which is shown in panel (c)). The attention regimes are based on Google Trends. Panel (d) shows the interaction term of Google Trends with the shock (for a one standard deviation shock and Google Trends being one standard deviation above its mean) when controlling for regional CPI inflation and time and region fixed effects. Standard errors are robust with respect to heteroskedasticity as well as serial and cross-sectional correlation.

Other shocks. So far, I focused on the implications of supply shocks on inflation. In Appendix C, I show that other shocks become more inflationary in the high-attention regime as well. In Figure 12, I show that inflation responds twice as much to the *main inflation shock* from Angeletos et al. (2020a) and that these differences are highly persistent. In Figure 13, I show that inflation also responds more strongly in the high-attention regime to monetary policy shocks identified using a high-frequency identification (the shocks are taken from Jarociński and Karadi (2020) and are purged from the information effects of monetary policy statements). These findings suggest that higher attention to inflation not only renders supply shocks more inflationary, but shocks more generally become more inflationary. The predictions of the model derived later will be consistent with these findings.

Robustness. In Appendix C, I discuss a number of robustness checks. In particular, I show that my results are not driven by different oil price responses (Figures 15 and 16) or a higher sensitivity of macroeconomic variables in general (Figures 17 and 18). The results are very similar when excluding the Covid period (Figure 19), or when excluding the Great Inflation period (Figure 20). Similarly, the results remain robust when using different control variables (Figures 21 and 22), or when focusing on the cumulative changes in the price level rather than the cumulative changes in inflation compared to the initial inflation rate (Figure 23). I also use the attention measure based on Bracha and Tang (2025) and show that the results remain robust to using this alternative measure (Figure 24). I also consider two alternative approaches to construct Google Trends at the regional level (Figure 25). Overall, the results in this section show that inflation responds 2-3 times as strongly to inflationary supply shocks when the public's attention to inflation is high compared to periods in which it is low.

The role of attention in the recent inflation surge. In Appendix C.3, I use the empirical estimates of this section to show that (i) the oil supply news shocks considered in this section, and (ii) the increase in people's attention in early 2021 in the United States contributed in a quantitatively important way to the subsequent inflation surge in 2021 and 2022. In particular, I estimate that the oil supply news shocks explain about half of the inflation dynamics between 2021 and 2022, but that without the attention increase, these effects would have only been about half as pronounced.

4 A Monetary Model with the Attention Threshold

To better understand the underlying mechanisms of the empirical results of the previous section, as well as the macroeconomic implications of the attention threshold more generally, I now introduce the attention threshold and changing degrees of people's attention to inflation in an otherwise standard New Keynesian model (Woodford, 2003; Galí, 2015).

4.1 Households

There is a representative household obtaining utility from consumption and disutility from working, with lifetime utility:

$$\tilde{E}_0 \sum_{t=0}^{\infty} \beta^t Z_t \left[\log\left(C_t\right) - \Xi H_t\right], \tag{6}$$

where C_t is consumption of the final good, H_t is hours worked, β is the household's time discount factor, and $\tilde{E}_t$ denotes the household's subjective expectations operator based on information available in period t. Z_t are exogenous preference shocks. The parameter Ξ is the utility weight on hours worked.

Households maximize their lifetime utility subject to the flow budget constraints:

$$C_t + B_t = W_t H_t + \frac{1+R_{t-1}}{1+\pi_t} B_{t-1} + \frac{T_t^H}{P_t}, \qquad \text{for all } t, \tag{7}$$

where B_t is the real value of government bonds, W_t the real wage, π_t is the net inflation rate, and R_t the nominal interest rate. T_t^H comprises lump-sum taxes and transfers from the government as well as firm profits.

Maximizing (6) subject to (7) yields the Euler equation:

$$Z_t C_t^{-1} = \beta(1+R_t)\tilde{E}_t\left[Z_{t+1}C_{t+1}^{-1}\frac{1}{1+\pi_{t+1}}\right], \tag{8}$$

and the labor-leisure condition:

$$W_t = \Xi C_t. \tag{9}$$

The labor-leisure equation (9) shows that consumption fluctuations are solely driven by aggregate wage fluctuations.

4.2 Firms

The firm sector is held by risk-neutral managers that discount future profits by β and they have a mass of zero, such that their consumption is 0 and all their profits go to the households, as in Bayer et al. (forthcoming).

Final good producer. There is a representative final good producer that aggregates the intermediate goods $Y_t(j)$ to a final good Y_t, according to:

$$Y_t = \left(\int_0^1 Y_t(j)^{\frac{\epsilon-1}{\epsilon}} dj\right)^{\frac{\epsilon}{\epsilon-1}}, \tag{10}$$

with $\epsilon > 1$. Nominal profits are given by $P_t\left(\int_0^1 Y_t(j)^{\frac{\epsilon-1}{\epsilon}} dj\right)^{\frac{\epsilon}{\epsilon-1}} - \int_0^1 P_t(j)Y_t(j)dj$, and profit maximization gives rise to the demand for each variety j:

$$Y_t(j) = \left(\frac{P_t(j)}{P_t}\right)^{-\epsilon} Y_t. \tag{11}$$

Thus, demand for variety j is a function of its relative price, the price elasticity of demand ϵ and aggregate output Y_t. The aggregate price level is given by:

$$P_t = \left(\int_0^1 P_t(j)^{1-\epsilon} dj\right)^{\frac{1}{1-\epsilon}}. \tag{12}$$

Intermediate producers. Intermediate producer of variety j produces output $Y_t(j)$ using labor $H_t(j)$ as its only input:

$$Y_t(j) = H_t(j). \tag{13}$$

All intermediate producers pay the same wage W_t and a sales tax (or subsidy) τ_t, which in steady state is set such that profits in steady state are 0.[21] These taxes are given back to firms in a lump-sum fashion, denoted $T_t^F(j)$. Taxes are assumed to be constant in the efficient economy, i.e., absent price rigidities, but fluctuate around their steady state in the economy with price rigidities in order to give rise to exogenous cost-push shocks.

Each intermediate firm has two managers: one is responsible for the firm's forecasts and the other one sets the price of firm j given these forecasts, similar to the setup in, e.g., Adam and Padula (2011). I first discuss the problem of the price setter and discuss the forecaster's problem later.

When adjusting the price, the firm is subject to a Rotemberg (1982) price-adjustment friction. Cavallo et al. (2024) and Blanco et al. (2024) argue that the frequency of price adjustments has increased during the recent inflation surge (see also Hall (2023)). While attention and the frequency of price adjustment are empirically both positively correlated with inflation, my results using alternative measures of attention to inflation (such as Google Trends or the measure based on Bracha and Tang (2025))—in which I control for inflation and time fixed effects—suggest that changes in price-setting behavior alone cannot explain the whole inflation dynamics and that attention indeed played an important role. In Appendix E.2, I show that a model with constant attention but state-dependent price adjustment costs cannot account for the inflation patterns arising from changing attention levels.

The per-period profits (in real terms) are given by:

$$(1-\tau_t)P_t(j)\left(\frac{P_t(j)}{P_t}\right)^{-\epsilon}\frac{Y_t}{P_t} - W_t H_t(j) - \frac{\psi}{2}\left(\frac{P_t(j)}{P_{t-1}(j)} - 1\right)^2 Y_t + T_t^F(j), \tag{14}$$

where $\psi \geq 0$ captures the price-adjustment cost parameter. The price setter sets the price to maximize profits:

$$\Omega_0(j) = \tilde{E}_0^j \sum_{t=0}^{\infty} \beta^t \left[(1-\tau_t)P_t(j)\left(\frac{P_t(j)}{P_t}\right)^{-\epsilon}\frac{Y_t}{P_t} - MC_t H_t(j) - \frac{\psi}{2}\left(\frac{P_t(j)}{P_{t-1}(j)} - 1\right)^2 Y_t + T_t^F(j)\right],$$

where $MC_t = W_t$ denotes the real marginal cost which is the same for every firm. Using the production function to substitute for $H_t(j)$ and the demand for firm j's product from the final goods

[21] Therefore, we have $1 - \tau = \frac{\epsilon}{\epsilon - 1}$ in steady state.

producer, the corresponding first order condition is then given by:

$$(1-\tau_t)(\epsilon-1)P_t(j)^{-\epsilon}\frac{Y_t}{P_t^{1-\epsilon}} = \epsilon MC_t\left(\frac{P_t(j)}{P_t}\right)^{-\epsilon-1}\frac{Y_t}{P_t} - \psi\left(\frac{P_t(j)}{P_{t-1}(j)}-1\right)\frac{Y_t}{P_{t-1}(j)} + \beta\psi\tilde{E}_t^j\left[\left(\frac{P_{t+1}(j)}{P_t(j)}-1\right)\frac{P_{t+1}(j)}{P_t(j)}\frac{Y_{t+1}}{P_t(j)}\right].$$

Defining $T_t \equiv 1-\tau_t$, it follows that after a linearization around the zero-inflation steady state, firm j sets its price according to:

$$p_t(j) = \frac{1}{\psi+\epsilon}\left[\psi p_{t-1}(j) + \epsilon\left(mc_t - t_t + p_t\right) + \beta\psi\tilde{E}_t^j\pi_{t+1}^j\right], \tag{15}$$

where small letters denote log deviations of the respective variables from their steady state values (see Appendix D for all derivations). Therefore, prices may only differ across firms j due to differences in forecasts $\tilde{E}_t^j\pi_{t+1}^j$ or differences in past prices $p_{t-1}(j)$.

Government. The government imposes a sales tax τ_t on sales of intermediate goods, issues government bonds, and pays lump-sum taxes and transfers $\mathcal{T}_t^H$ to households and T_t^F to firms. The real government budget constraint is given by:

$$B_t = B_{t-1}\frac{1+R_{t-1}}{1+\pi_t} + \frac{\mathcal{T}_t^H}{P_t} - \tau Y_t + T_t^f.$$

The monetary authority sets the nominal interest rate according to a (linearized) Taylor rule:

$$i_t = \rho_i i_{t-1} + (1-\rho_i)\left(\phi_\pi\pi_t + \phi_x x_t\right), \tag{16}$$

where i_t denotes the nominal interest rate in deviations from its steady state, ρ_i captures interest-rate smoothing, ϕ_π and ϕ_x pin down the response coefficients with respect to inflation and the output gap, respectively. I discuss other rules for monetary policy in Section 6.

4.3 Subjective expectations under limited attention

The subjective expectations of households and firm managers are modeled in such a way that there is a direct mapping from the empirical results of Section 2 to the model. I solve the model under the assumption that the agents in a given regime do not foresee that their attention will change when inflation exceeds or falls below the threshold. Even though the attention model in Section 2.1 features a second threshold (see Figure 2), I focus here on the case with only one threshold, given that the empirical evidence in Section 2 suggests that the second threshold has not been relevant in the post 1970s U.S. experience.

Households. Households have to form expectations about inflation and real wages, as consumption fluctuations are solely driven by aggregate wage fluctuations.[22] Households believe that wages and inflation both follow an AR(1):[23]

$$\pi_t = (1-\tilde{\rho}_\pi)\underline{\pi} + \tilde{\rho}_\pi \pi_{t-1} + \nu_{\pi,t}$$
$$w_t = (1-\tilde{\rho}_w)\underline{w} + \tilde{\rho}_w w_{t-1} + \nu_{w,t},$$

where $\nu_{\pi,t}$ and $\nu_{w,t}$ are normally distributed with mean zero, and independent from each other. I discuss the case in which households hold rational expectations about real wages later. As a baseline, I focus on $\tilde{\rho}_\pi = \tilde{\rho}_w = 1$ which implies that long-run expectations align with the actual long-run variables and I do not need to take a stance on what the long-run beliefs $\underline{\pi}$ and $\underline{w}$ are. I discuss the case with $\tilde{\rho}_\pi < 1$ in Appendix E.2. Agents observe the aggregate shocks τ_t and Z_t perfectly.

At the time the household forms her expectations about future wages and inflation, she does not perfectly observe their current realizations, capturing that the household is not at all times perfectly monitoring the (real) wage incomes of the members of the household and that inflation is not perfectly observable in real time. Instead, the household only sees noisy signals of the form:

$$s_{\pi,t} = \pi_t + \varepsilon_{\pi,r,t}$$
$$s_{w,t} = w_t + \varepsilon_{w,t},$$

with normally distributed noise terms $\varepsilon_{\pi,r,t}$ and $\varepsilon_{w,t}$. I assume that signals are public signals and every agent in the economy receives the same signals. This implies that all agents have the same information and hence, hold the same expectations. Households treat these signals as separate, meaning that households do not infer from the signal about real wages any information about the price level and inflation (or vice versa).[24]

As detailed in Appendix D, potential output, i.e., output under flexible prices, is constant and consumption in log-deviations from steady state is equal to the output gap x_t in equilibrium (and equal to wages in deviations from steady state), $c_t = w_t = x_t$.[25] It follows that if we assume initial values $c^e_{0|-1} = w^e_{0|-1} = x^e_{0|-1}$, we have $c^e_{t+1|t} = w^e_{t+1|t} = x^e_{t+1|t}$ for all t. This holds true, even if the household does not know that consumption and wages are equal to the output gap in equilibrium.

[22] I assume that bonds adjust residually such that the budget constraint always holds (a standard assumption in the rational inattention literature, see, e.g., the discussion on p. 242 in Maćkowiak et al. (2023)). Households do not update their information set at the end of the period when bond holdings adjust.

[23] The perceived law of motion and the actual law of motion coincide in general equilibrium only in the special case of full attention and no endogenous persistence and choosing the perceived persistencies to be equal to the persistence of the exogenous shock.

[24] Thus, the relevant information set of households to form expectations consists of the history of those two signals and they treat them separately when forming their expectations about real wages and inflation. An alternative is that agents have rational expectations about the wage and they solely obtain signals about inflation. Again, households do not use information about other variables to infer anything about inflation or the price level. I show in Appendix E.2 that the results are very similar in that case.

[25] Note that the price adjustment costs do neither affect the steady state nor the linearized resource constraint when linearized around the zero-inflation steady state, such that $y_t = c_t$.

Expectations about the output gap, x_t, then evolve as follows:

$$x^e_{t+1|t} = x^e_{t|t-1} + \gamma_x \left(x_t - x^e_{t|t-1}\right), \tag{17}$$

where $x^e_{t+1|t}$ denotes the agent's expectations of the one-period-ahead output gap. The parameter γ_x denotes attention to the output gap, based on the agent's subjective model of how the output gap evolves. A higher γ_x denotes a higher attention level. If $\gamma_x = 0$, the agent is completely inattentive and just sticks to her prior belief $x^e_{t|t-1}$, whereas $\gamma_x = 1$ captures the case of full attention in which case the agent believes $x^e_{t+1|t} = x_t$, which is the full-information belief of someone who believes that the output gap follows a random walk. As I discuss in more detail in the calibration section later on, I do not find any differences in attention to unemployment changes (which I use as a proxy for the output gap) in the data. I therefore impose that γ_x does not change across regimes in the baseline case. I discuss the cases with regime-dependent γ_x and with rational output gap expectations later.

Inflation expectations follow the law of motion derived in equation (3) for $\tilde{\rho}_\pi = 1$, so that they are given by:

$$\pi^e_{t+1|t} = \begin{cases} \pi^e_{t|t-1} + \gamma_{\pi,L}\left(\pi_t - \pi^e_{t|t-1}\right), & \text{when } \pi_{t-1} < \bar{\pi} \\ \pi^e_{t|t-1} + \gamma_{\pi,H}\left(\pi_t - \pi^e_{t|t-1}\right), & \text{when } \pi_{t-1} \geq \bar{\pi}, \end{cases} \tag{18}$$

where $\gamma_{\pi,r}$ captures the optimal level of attention to inflation in regime $r \in \{L, H\}$.[26]

Firm managers. Since there are no idiosyncratic shocks, I assume that the forecasting manager of firm j uses expectations about aggregate inflation to form her expectations about firm j's future price change, i.e., $\tilde{E}^j_t \pi^j_{t+1} = \tilde{E}^j_t \pi_{t+1}$, and that they form expectations about inflation in the same way as households. Consistent with these assumptions, Yotzov et al. (2024) show that firms' expected own-price growth is strongly positively correlated with changes in CPI inflation, and McClure et al. (2025) show that managers and non-managers hold similar average inflation and unemployment expectations and respond similarly to information treatments.[27] As for households, all forecasters receive the same public signal about current inflation, from which it follows that $\tilde{E}^j_t \pi^j_{t+1} = \tilde{E}^j_t \pi_{t+1} = \tilde{E}_t \pi_{t+1} \equiv \pi^e_{t+1|t}$.

Assuming that firms all start out with the same initial price, i.e., $p_{-1}(j) = p_{-1}$ for all $j \in [0, 1]$, it follows that all firms set the same price (as they do under rational expectations). Thus, firm-specific inflation is indeed equal to aggregate inflation, which confirms the forecaster's belief that the two are equal, and hence, the forecaster has no reason to deviate from these beliefs.

[26] Without loss of generality, I do not explicitly model noise shocks. Accounting for noise shocks would introduce an additional exogenous shock, but does not affect my analysis of how supply and demand shocks are transmitted.

[27] McClure et al. (2025) further show that managers' expectations indeed affect their economic decisions. Relatedly, Link et al. (2025) elicit firm managers' attention to economic topics and they show that firm managers that pay more attention to inflation update their inflation expectations more strongly (consistent with my measure of attention) and they increase their prices more or plan to increase their prices more in times of high inflation compared to firm managers that do not pay attention to inflation.

Given these assumptions, equation (15) can be written as:

$$\pi_t = \frac{\epsilon}{\psi}\left(mc_t - t_t\right) + \beta\pi^e_{t+1|t}. \tag{19}$$

From the labor-leisure equation and the production function, we have $mc_t = y_t$, and since potential output is constant we have $y_t = x_t$. Defining cost-push shocks as $u_t \equiv -\frac{\epsilon}{\psi}t_t$ and $\kappa \equiv \frac{\epsilon}{\psi}$, we arrive at the linearized New Keynesian Phillips Curve:

$$\pi_t = \beta\pi^e_{t+1|t} + \kappa x_t + u_t. \tag{20}$$

4.4 Equilibrium

The equilibrium is a *restricted perceptions equilibrium* (Sargent, 1991; Branch, 2006): private agents' perceived law of motion for inflation and real wages is an AR(1) process which differs from the actual law of motion of these variables.[28] The model can be summarized by three equilibrium equations when expressed in log-deviations from the zero-inflation steady state (see Appendix D):

$$\pi_t = \beta\pi^e_{t+1|t} + \kappa x_t + u_t, \tag{21}$$

$$x_t = x^e_{t+1|t} - \left(i_t - \pi^e_{t+1|t} - r^*_t\right), \tag{22}$$

$$i_t = \rho_i i_{t-1} + (1-\rho_i)\left(\phi_\pi \pi_t + \phi_x x_t\right), \tag{23}$$

as well as the two laws of motion for output gap expectations and inflation expectations, equations (17) and (18). Equations (21)-(23) are identical to the 3-equation representation of the standard New Keynesian model (see Galí (2015)) except for the expectations formation of agents.

Equation (21) is the New Keynesian Phillips curve, representing the supply side of the economy, and equation (22) denotes the aggregate Euler (or IS) equation, which together with monetary policy (equation (23)) pins down aggregate demand. The natural interest rate r^*_t is the real rate that prevails in the economy with fully flexible prices, and solely depends on the exogenous shocks Z_t. It follows an AR(1) process with persistence $\rho_r \in [0, 1]$ and innovations $\varepsilon^r \sim i.i.N.(0, \sigma^2_r)$, independent of ε^u. I will refer to shocks to r^*_t as demand shocks. The nominal interest rate i_t and the natural rate are both expressed in absolute deviations of their respective steady state values, $\underline{i}$ and $\underline{r}^*$, with $\underline{i} = \underline{r}^*$, as the model is linearized around the zero-inflation steady state.

4.5 Calibration

Most of the calibration is standard. I set the discount factor β to target a steady state natural rate of 1% (annualized), and the slope of the Phillips curve κ to 0.057. The Taylor rule coefficients are

[28] In the special case of full attention, the restricted perceptions equilibrium (RPE) may coincide with the rational expectations equilibrium if $\rho_i = 0$ and the perceived persistence of real wages and inflation coincide with the exogenous persistence of the shocks. In this case, the RPE is equivalent to a *consistent expectations equilibrium* (Hommes and Sorger, 1998). See Moll (2024) for a recent overview of these different equilibrium concepts and how they relate to each other.

set to $\rho_i = 0.7$, $\phi_\pi = 2$ and $\phi_x = 0.125$.[29]

I assume that both shocks follow an AR(1) process with persistence 0.8, and I set $\sigma_u = \sigma_r = 0.3\%$. I set $\gamma_{\pi,L} = 0.18$, $\gamma_{\pi,H} = 0.35$, and the threshold to $\bar{\pi} = 3.91\%$ (annualized), as estimated in Section 2.[30]

To calibrate the attention parameter with respect to the output gap γ_x, I follow the same procedure as for inflation but focus on expectations about unemployment changes (see Appendix D.1). This results in $\gamma_x = 0.25$ for both regimes, and hence, I impose that γ_x does not change across regimes and set it to $\gamma_x = 0.25$.

As a sanity check of the proposed empirical approach to estimate the attention threshold and the degrees of attention in a setup in which inflation and inflation expectations are determined jointly and different types of shocks are hitting the economy, I estimate the threshold regression (4) on model-simulated data and find that the regression results are exactly equal to the calibrated parameters.

5 Inflation Surges

In this section, I show how the model with two inflation-attention regimes can jointly generate persistent heightened inflation periods, a long last mile of disinflation, forecast error dynamics mirroring the ones recently observed in the United States, and lead to an asymmetry in the dynamics of inflation. In Appendix E.1, I show how the model can be represented in a standard AS-AD framework in which crossing the attention threshold leads to a steepening of both the AS and the AD curve. Thus, the attention threshold gives rise to a *dynamic* non-linearity. Additionally, the heightened attention amplifies the inflationary effects of shocks. In Appendix E.1, I also provide empirical evidence for the changes in the slopes of the AS and AD curve.

5.1 Crossing the attention threshold

Figure 6 shows the inflation dynamics following a cost-push shock that pushes inflation in the first period above the attention threshold $\bar{\pi}$ with a persistence of 0.8 (the red-dashed line shows the inflation response, and the black-dotted line the attention threshold), as well as following a shock that does not push inflation above the threshold (blue-dashed-dotted line).

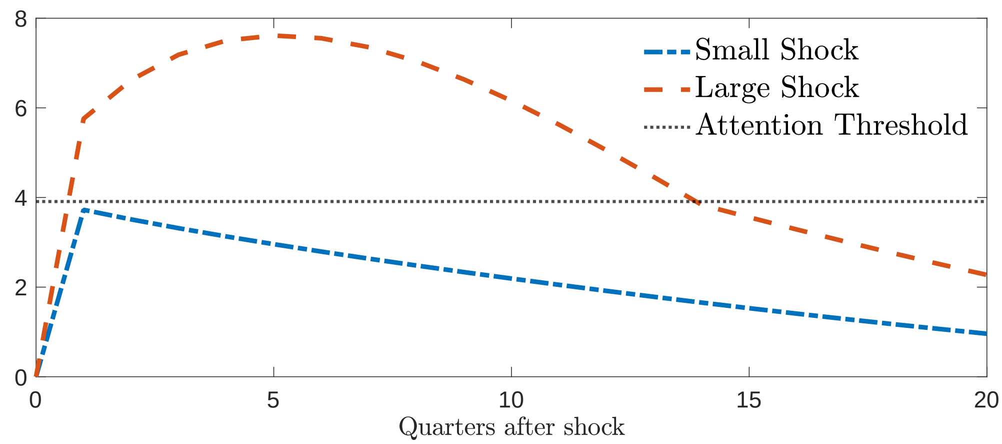
The two cases exhibit fundamentally different inflation dynamics. When attention remains low, which happens in response to the relatively small shock, inflation peaks on impact and then gradually

[29] The model is determinate under this specification for the monetary policy rule. Numerically, I find that the model does not satisfy the Blanchard-Kahn conditions any longer when attention reaches a level of 0.82 (while keeping the remaining parameters fixed). When considering the calibration used to produce Figure 2, I find that this level of attention is reached at an inflation rate of 80%, indicating that the economy may become indeterminate at such a high level of inflation.

[30] Instead of directly calibrating the threshold, I could instead calibrate the different cost parameters, as outlined in Appendix A. As these cost parameters only matter for the threshold, however, I take the more direct route by directly calibrating the threshold to the empirically-estimated one.

returns to its initial value of zero. After the larger shock, however, inflation keeps on increasing for about five periods before it peaks and then starts to decrease thereafter. While this decrease happens relatively fast at first, it slows down once inflation falls back below the attention threshold, and thus, inflation remains elevated quite persistently. These self-reinforcing dynamics in the first few periods would not be present in the case of deflationary shocks because attention to inflation stays constant in that scenario. Therefore, the model with the attention threshold offers a simultaneous explanation for the recent inflation surge as well as the 'missing deflation puzzle' in the aftermath of the Great Financial Crisis (see, e.g., Coibion and Gorodnichenko (2015b)). Additionally, the different inflation dynamics also mirror the ones discussed in Figure 5, namely that inflation peaks on impact when attention is low whereas it is hump shaped in times of high attention.[31]

Figure 6: Inflation dynamics



Notes: The blue-dashed-dotted line shows the (annualized) inflation response to a cost-push shock that does not push inflation above the attention threshold, and the red-dashed line shows the inflation dynamics for a cost-push shock that is large enough to push inflation above the threshold.

To understand how inflation surges may arise in the model with the attention threshold, both regime switches that take place are key. The first regime switch occurs because the shock impulse is large enough to push inflation above the attention threshold. Thus, in the second period, agents are more attentive to inflation and hence, increase their inflation expectations more strongly in response to the forecast errors they make. For a given nominal rate, households now perceive the real rate to be lower and thus, increase their consumption in response. The attention regime change also matters for the supply side. Firms increase their inflation expectations more strongly which leads them to increase their prices more strongly. On top of that, the equilibrium inflation response to the cost-push shock in a given state of the economy is higher in the high-attention regime. Hence, inflation keeps on increasing, further fueling higher inflation expectations, leading to additional inflation increases.

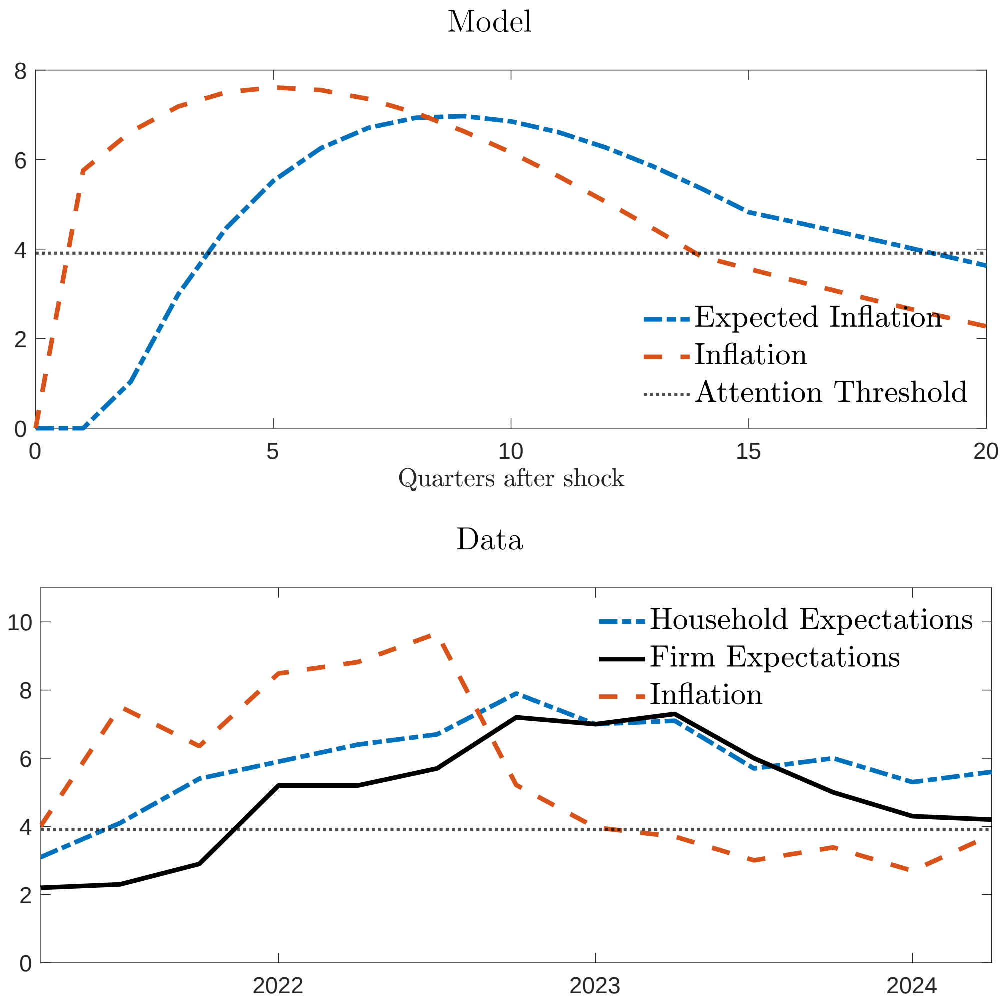
As the shock slowly dies out, inflation eventually starts to decline. This initial period of disinflation happens relatively quickly. The reason is that once inflation actually starts to come down, households and firm managers are still highly attentive and hence, revise their expectations downwards strongly.

[31]Figure 31 shows the response of inflation when the shock hits in times of high-attention and illustrates that the inflation response is indeed hump shaped in that case.

Thus, actual inflation falls relatively quickly.

However, once inflation falls back below the threshold $\bar{\pi}$—and the second regime switch occurs—the speed of disinflation decreases. When the economy enters the low-attention regime, agents decrease their attention to inflation. Thus, they mainly stick to their prior beliefs and only slowly update their expectations. Their priors are now relatively high because of the high inflation period that agents just lived through, and therefore, their inflation expectations remain persistently high which hinders actual inflation from decreasing quickly.[32]

Figure 7: Inflation and inflation expectations dynamics



Notes: The upper panel shows the dynamics of inflation and inflation expectations after a cost-push shock that pushes inflation above the attention threshold. The lower panel plots quarter-on-quarter (annualized) CPI inflation and average inflation expectations for the period 2021Q1-2024Q1, as well as the estimated attention threshold at 3.91%. In both panels, the inflation expectations are shifted such that the vertical distance between the two shows the respective forecast errors.

To clearly see the interplay of inflation and inflation expectations, the upper panel in Figure 7 plots expected and actual inflation, jointly. Inflation expectations are shifted such that the vertical

[32]Lebow and Peneva (2024) also show that the perceptions of inflation among households came down more slowly than actual inflation.

distance between the two lines captures the forecast errors. Initially, expected inflation does not quite catch up with actual inflation, leading to positive forecast errors. After some time—around 8 quarters after the shock—when inflation is already decreasing, expected inflation surpasses actual inflation. Hence, forecast errors become negative.

The recent inflationary period provides a salient example of these patterns, as documented in panel (b) in Figure 1 in the Introduction. The lower panel in Figure 7 shows the same time series but for quarter-on-quarter inflation to make them consistent with the model (Figure 26 in Appendix C shows these patterns in the data conditional on an oil supply news shock).[33] The figure shows annualized quarter-on-quarter CPI inflation (red-dashed line) and average inflation expectations from the Michigan Survey (blue-dashed-dotted line) and from the Survey of Firms' Inflation Expectations (black-solid line) for the period from 2021Q1 until 2024Q1. Consistent with the model, inflation peaks about a year and a half after exceeding the threshold. During this inflation surge, inflation expectations lagged actual inflation. As inflation peaked and began to decline, however, inflation expectations started to surpass actual inflation after about two years after inflation exceeded the attention threshold. Consistent with the model, inflation peaks at a higher value than inflation expectations. The recent inflation surge is just one example of a high-attention period. The model's dynamics are also consistent with the empirical findings in Blanco et al. (2025) who document, for a number of countries and different periods of inflation surges, that (i) inflation stays persistently high after the initial surge and (ii) that short-run inflation expectations initially fall short of actual inflation.

That inflation expectations stay persistently high may lead to a prolonged period of inflation stubbornly above the central bank's target even though the initial disinflationary phase happens relatively quickly. Put differently, once attention to inflation decreases again, the higher prior expectations after the inflation surge may render the 'last mile' of inflation back to target more arduous. In Appendix E.2, I show that the attention decrease can explain a substantial fraction of the long last mile of disinflation observed in the U.S. after the recent inflation surge.

Stubbornly high inflation and inflation expectations also give rise to a heightened risk of another subsequent inflation surge. The higher inflation expectations keep actual inflation higher for longer, and therefore, closer to the attention threshold. Thus, a subsequent inflationary shock is more likely to push inflation back above the threshold and, therefore, lead to another episode of persistently high inflation. Hence, endogenous changes in the private sector's attention to inflation may partially account for some of the empirically-documented time variation in trend inflation (see, e.g., Stock and Watson (2016)).

In Appendix E.2, I show that these inflation dynamics are unique to the model with the attention threshold. In particular, the models under full-information rational expectations (FIRE) or in which attention remains low predict that inflation peaks on impact and then comes down relatively quickly. Additionally, models with FIRE could not account for the observed dynamics of forecast errors

[33] Angeletos et al. (2020b) show that this initial undershooting followed by a delayed overshooting of expectations is a general pattern in response to shocks in the data.

(among other failures of this model in accounting for the recent inflation dynamics). A model variant in which attention to inflation is always high, on the other hand, can produce the initial hump-shaped response but then predicts a fast decline back to target, i.e., this model would not explain why we see such a long last mile of disinflation. Furthermore, a model in which attention is always high would have predicted a very pronounced deflationary period after the Great Financial Crisis. Thus, accounting for the observed attention changes is necessary to rationalize the recent inflationary episode.

In Appendix E.2, I further discuss (i) the inflation dynamics following a demand shock (Figure 32), (ii) what the model predicts when attention to the output gap decreases in times of high attention to inflation (Figure 33), and (iii) how the inflation dynamics look like when consumption expectations are formed under FIRE. In cases (i) and (ii), the inflation surge is somewhat more persistent but otherwise very similar to the baseline case of the discussed cost-push shock with constant attention to the output gap, whereas in (iii) the model produces slightly less endogenous persistence. I also consider the possibility that attention stays elevated for some time after an inflation surge and show that my results are largely unaffected when doing so.

5.2 Asymmetry

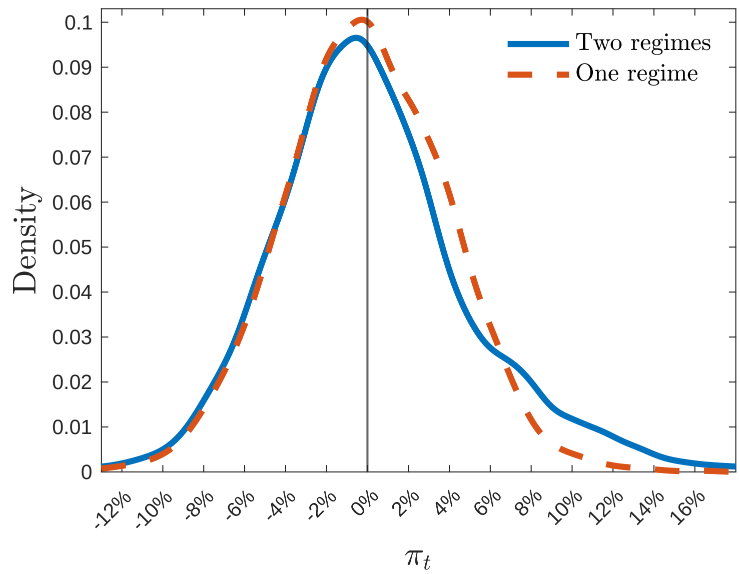
The attention threshold induces an asymmetry in the dynamics of inflation. When inflation exceeds the threshold, attention increases, leading to a period of persistently high inflation. If, however, inflation is particularly low, attention remains unchanged. For example, when a shock pushes inflation from 0% to -4%, attention does not change. Therefore, we would not see long-lived periods of pronounced deflation after such a shock, consistent with the fact that the U.S. economy did not experience any long lasting deflationary periods since the second World War. These asymmetries matter for the overall properties of inflation. In particular, ignoring the attention threshold results in underpredicting the risk of persistently high inflation rates.

Figure 8 plots the distribution of inflation for the model with the attention threshold (blue-solid line) and for the limited-attention model without the threshold (red-dashed line). These figures are obtained by simulating the model for 10,000 periods by hitting it with random normally-distributed cost-push and demand shocks.

The model with the attention threshold has a substantially thicker right tail than the one without the attention threshold, as is the case empirically. The probability that inflation exceeds 8% (annualized) is about 7.9% in the model with the threshold, and thus very close to its empirical counterpart of 7.3% whereas the model without the threshold predicts inflation to exceed 8% only 2.2% of the time. When it comes to periods of deflation, the two models exhibit very similar properties. For example, both models predict inflation to fall below -8% in only 2-3% of the periods, compared to 0.6% empirically, and the median in both models is very close to 0.[34] Thus,

[34]The mean is above 0 in the model with the threshold, whereas it is practically 0 in the model without the threshold. I discuss this in more detail in Section 6 where I derive the normative implications of these results.

Figure 8: The attention threshold and inflation asymmetry



Notes: This figure shows the distribution of inflation (annualized) for the model with the attention threshold (solid-blue line) and the one without the attention threshold (dashed-red line).

inflation in the model without the attention threshold is symmetric around its steady state value of 0, whereas it features a substantially higher risk of periods of high inflation in the model with the threshold. Inflation is also more volatile when attention to inflation is higher.

6 Monetary Policy Implications

I now characterize the normative implications of the inflation attention threshold and the corresponding changes in attention. To compare different policy rules, I define the central bank's loss function as:

$$\frac{1}{2}E_0\sum_{t=0}^{\infty}\beta^t\left[\pi_t^2+\Upsilon x_t^2\right], \tag{24}$$

where Υ is the relative weight of the output gap, which I set to $\Upsilon = 0.007$ as in Adam and Billi (2006). The policymaker wants to minimize (24) subject to the private sector's optimality conditions, characterized by equations (21) and (22). But I assume that the planner wrongly believes that the private sector holds rational expectations. The problem from the policymaker's perspective is then identical to the one in Clarida et al. (1999) who show that the optimal policy is given by:

$$\pi_t + \tfrac{\Upsilon}{\kappa}\left(x_t - x_{t-1}\right) = 0$$

if the policymaker can fully commit to its policy. If the policymaker cannot commit, the optimal policy rule under discretion is given by:

$$\pi_t + \tfrac{\Upsilon}{\kappa}x_t = 0.$$

In the following, I compare the implications of these two policy rules to the Taylor rule (23) as well as to a Taylor rule without interest rate smoothing and no response to the output gap, i.e., to (23) where $\rho_i = \phi_x = 0$, as well as to a strict-inflation targeting regime in which inflation is kept at zero at all times. Table 3 summarizes these different policy rules.

Table 3: Monetary policy rules

Nr.	Name	Equation
(1)	Taylor rule with smoothing	$i_t = \rho_i i_{t-1} + (1-\rho_i)(\phi_\pi \pi_t + \phi_x x_t)$
(2)	Taylor rule without smoothing	$i_t = \phi_\pi \pi_t$
(3)	Optimal RE commitment policy	$\pi_t + \frac{\Upsilon}{\kappa}(x_t - x_{t-1}) = 0$
(4)	Optimal RE discretionary policy	$\pi_t + \frac{\Upsilon}{\kappa} x_t = 0$
(5)	Strict inflation targeting	$\pi_t = 0$

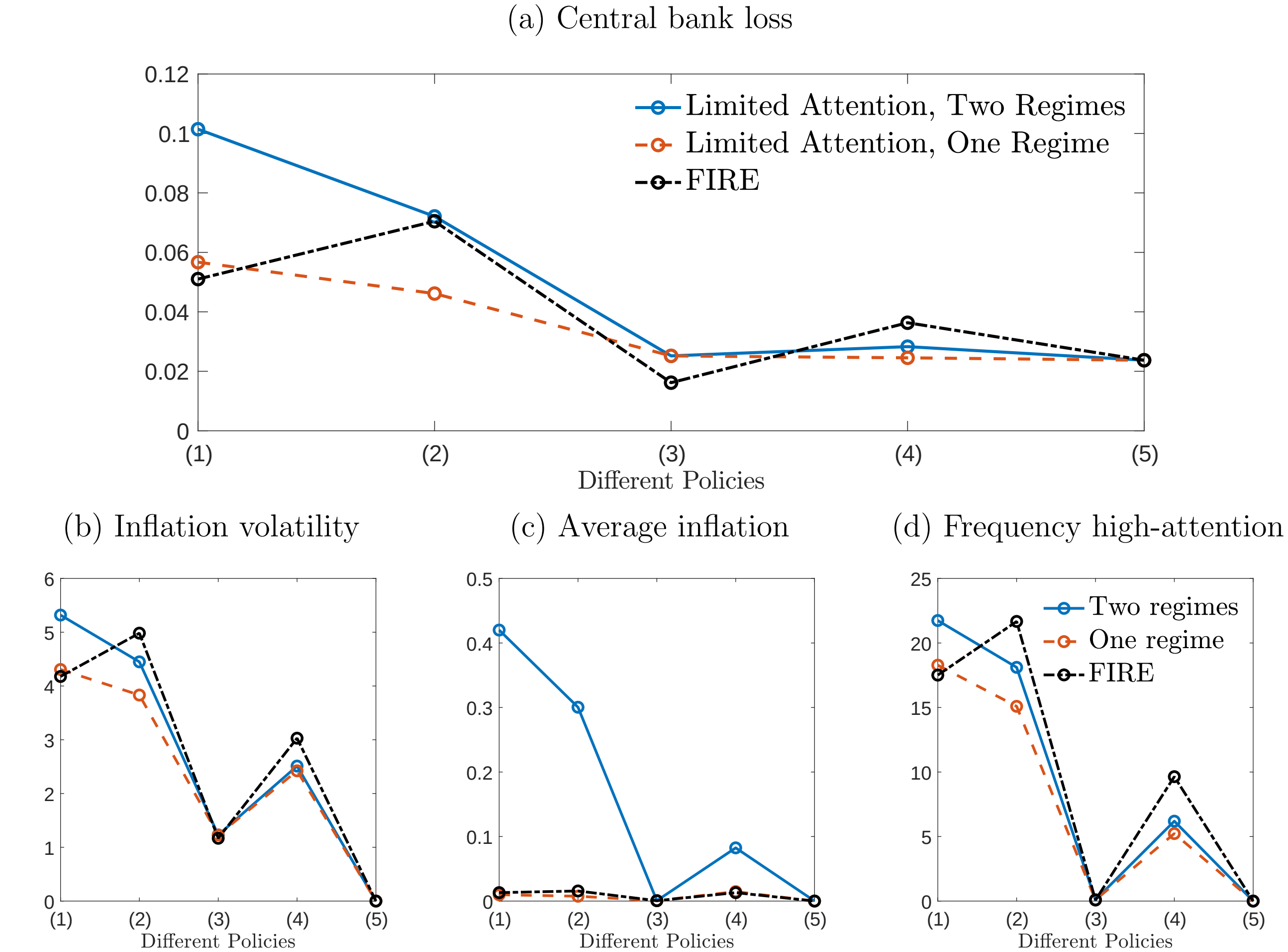
I then simulate the economy for 10,000 periods for each of the different policy rules. Panel (a) in Figure 9 plots the central bank's losses (24) for the model with the attention threshold (blue-solid line), with limited attention but absent the attention threshold (red-dashed line) and the model under FIRE (black-dashed-dotted line) for the 5 different policy rules. Panel (b) shows the respective inflation volatilities, panel (c) the average level of inflation, and panel (d) the frequency of how often inflation is above the threshold of 3.91%.

There are two main takeaways from Figure 9. First, following simple Taylor rules (policy rules (1) and (2)) leads to substantially larger central bank losses compared to optimal policy rules or a strict inflation targeting rule in the model with the attention threshold, especially for the case with interest-rate smoothing.[35] The reason is that the economy spends a substantial amount of time in the high-attention regime in which inflation is high and volatile (see panel (d) in Figure 9). Due to the asymmetry of the attention threshold the average level of inflation is higher when the economy is in the high-attention regime frequently (panel (c)). This is in stark contrast to the model without the threshold or the one with fully-informed rational agents, where inflation fluctuates symmetrically around zero. Interest rate smoothing introduces additional persistence, such that the periods in the high-attention regime last longer and thus, the average level of inflation as well as its volatility increase.

The second main takeaway is that these losses can be mitigated when monetary policy follows one of the rules (3)-(5). In these cases, inflation is relatively stable and fluctuates almost symmetrically around 0, as the economy very rarely stays in the high-attention regime. In the case of the strict-inflation targeting regime (5) the inflation volatility is exactly 0. However, in that case, the output gap is more volatile (not shown) such that the overall losses are very similar to the ones under the other policy rules from (3) and (4).

[35] Gáti (2023) shows that in her model with varying degrees of long-run-expectations anchoring, losses of Taylor rules with a low inflation-response coefficient may go to infinity, as in this case the model becomes unstable and inflation (and the output gap and interest rates) becomes explosive. Similarly, Benchimol and Bounader (2023) find that under inattention to prices, central bank losses under Taylor rules are higher than with other policy rules.

Figure 9: Implications of different monetary policy rules



Notes: This figure shows the central bank loss (using equation (24) and shown in panel (a)), inflation volatility (panel (b)), average inflation (panel (c)), and the frequency of being in the high-attention regime (panel (d)) for the five different monetary policy rules from Table 3.

Overall, Figure 9 illustrates that following simple Taylor rules, especially ones with interest-rate smoothing and relatively low inflation-response coefficients, can lead to large central bank losses when the public's attention to inflation increases during times of high inflation. Policies that are more hawkish and induce much smaller inflation fluctuations, in contrast, can mitigate the potentially detrimental effects of sudden increases in attention much more effectively, or even prevent these episodes from happening completely.

7 Conclusion

The recent inflation surge brought inflation back on people's minds. In this paper, I quantify the inflation attention threshold after which people start to pay more attention to inflation. I estimate this attention threshold to be at an inflation rate of about 4% and that attention doubles from the low-attention regime to the high-attention regime. Supply shocks become twice as inflationary in the high-attention regime.

A New Keynesian model that accounts for the inflation attention threshold can replicate the empirical findings and produces inflation and inflation expectations dynamics that are consistent with the ones documented empirically and recently observed in the U.S. As inflation exceeds the threshold, the increase in people's attention leads to an endogenous amplification of the initial inflation surge. After inflation peaks, inflation comes back down relatively quickly initially, but once attention to inflation decreases again, the higher prior expectations after the inflation surge may render the 'last mile' of inflation back to target more arduous. Accounting for varying attention levels also matters for the model's normative predictions. I show that following simple Taylor rules leads to substantially larger central bank losses compared to more hawkish rules.

The state-dependent attention mechanism studied here may interact with other forms of state dependence, such as in the frequency of price adjustment, changing supply chain networks, or the conduct of monetary policy, and exploring these interactions presents a promising avenue for future work. Similarly, characterizing the fully optimal monetary policy when people's attention to inflation is subject to sudden changes, and studying how the conduct of monetary policy feeds back into people's attention is a fruitful direction for future research.

References

Aastveit, K. A., H. C. Bjørnland, and J. L. Cross (2023): "Inflation expectations and the pass-through of oil prices," *Review of Economics and Statistics*, 105, 733–743.

Adam, K. and R. M. Billi (2006): "Optimal monetary policy under commitment with a zero bound on nominal interest rates," *Journal of Money, Credit and Banking*, 1877–1905.

Adam, K. and M. Padula (2011): "Inflation dynamics and subjective expectations in the United States," *Economic Inquiry*, 49, 13–25.

Afrouzi, H. and C. Yang (2021): "Dynamic inattention, the Phillips curve, and forward guidance," *Working Paper*.

Amiti, M., S. Heise, F. Karahan, and A. Şahin (2023): "Inflation Strikes Back: The Role of Import Competition and the Labor Market," *NBER Macroeconomics Annual, Volume 38*.

Andre, P., C. Pizzinelli, C. Roth, and J. Wohlfart (2022): "Subjective models of the macroeconomy: Evidence from experts and representative samples," *The Review of Economic Studies*, 89, 2958–2991.

Angeletos, G.-M., F. Collard, and H. Dellas (2020a): "Business-cycle anatomy," *American Economic Review*, 110, 3030–3070.

Angeletos, G.-M., Z. Huo, and K. Sastry (2020b): "Imperfect expectations: Theory and evidence," in *NBER Macroeconomics Annual 2020, volume 35*, University of Chicago Press.

Aruoba, B. (2020): "Term structures of inflation expectations and real interest rates," *Journal of Business & Economic Statistics*, 38, 542–553.

Aruoba, S. B. and T. Drechsel (2024): "Identifying monetary policy shocks: A natural language approach," *Working paper*.

Ascari, G. and T. Haber (2022): "Non-linearities, state-dependent prices and the transmission mechanism of monetary policy," *The Economic Journal*, 132, 37–57.

Auerbach, A. J. and Y. Gorodnichenko (2012a): "Fiscal multipliers in recession and expansion," in *Fiscal policy after the financial crisis*, University of Chicago Press, 63–98.

——— (2012b): "Measuring the output responses to fiscal policy," *American Economic Journal: Economic Policy*, 4, 1–27.

Ball, L. M., D. Leigh, and P. Mishra (2022): "Understanding US inflation during the covid era," *Brookings Papers on Economic Activity*.

Barnichon, R. and G. Mesters (2020): "Identifying modern macro equations with old shocks," *The Quarterly Journal of Economics*, 135, 2255–2298.

——— (2021): "The phillips multiplier," *Journal of Monetary Economics*, 117, 689–705.

Bayer, C., B. Born, and R. Luetticke (forthcoming): "Shocks, frictions, and inequality in US business cycles," *American Economic Review*.

Benati, L. (2008): "Investigating inflation persistence across monetary regimes," *The Quarterly Journal of Economics*, 123, 1005–1060.

Benchimol, J. and L. Bounader (2023): "Optimal monetary policy under bounded rationality," *Journal of Financial Stability*, 67, 101151.

Benigno, P. and G. B. Eggertsson (2023): "It's baaack: The surge in inflation in the 2020s and the return of the non-linear phillips curve," *Working paper*.

Bernanke, B. and O. Blanchard (2025): "What caused the US pandemic-era inflation?" *American Economic Journal: Macroeconomics*, 17, 1–35.

Bhandari, A., J. Borovička, and P. Ho (2025): "Survey data and subjective beliefs in business cycle models," *Review of Economic Studies*, 92, 1375–1437.

Bianchi, F., R. Faccini, and L. Melosi (2023): "A Fiscal Theory of Persistent Inflation," *The Quarterly Journal of Economics*.

Bianchi, F., C. L. Ilut, and H. Saijo (2024): "Smooth Diagnostic Expectations," *Working paper*.

Bianchi, F. and L. Melosi (2022): "Inflation as a fiscal limit," *2022 Jackson Hole Symposium Volume*.

Binder, C. and R. Kamdar (2022): "Expected and realized inflation in historical perspective," *Journal of Economic Perspectives*, 36, 131–155.

Blanco, A., C. Boar, C. J. Jones, and V. Midrigan (2024): "The Inflation Accelerator," *Working paper*.

Blanco, A., P. Ottonello, and T. Ranošová (2025): "The dynamics of large inflation surges," *Review of Economics and Statistics*, 1–31.

Bracha, A. and J. Tang (2025): "Inflation levels and (in) attention," *Review of Economic Studies*, 92, 1564–1594.

Branch, W. A. (2006): "Restricted perceptions equilibria and learning in macroeconomics," *Post Walrasian Macroeconomics: Beyond the Dynamic Stochastic General Equilibrium Model. Cambridge University Press, New York*, 135–160.

Buelens, C. (2025): "Googling 'inflation': Household inflation attention across the euro area," *European Journal of Political Economy*, 89, 102702.

Caggiano, G., E. Castelnuovo, and N. Groshenny (2014): "Uncertainty shocks and unemployment dynamics in US recessions," *Journal of Monetary Economics*, 67, 78–92.

Candia, B., O. Coibion, and Y. Gorodnichenko (2024): "The Inflation Expectations of US Firms: Evidence from a new survey," *Journal of Monetary Economics*, 145, 103569.

Candia, B., M. Weber, Y. Gorodnichenko, and O. Coibion (2023): "Perceived and Expected Rates of Inflation of US Firms," *AEA Papers and Proceedings*, 113, 52–55.

Canova, F. (2007): "G-7 inflation forecasts: Random walk, Phillips curve or what else?" *Macroeconomic Dynamics*, 11, 1.

Carlson, J. A. and M. Parkin (1975): "Inflation expectations," *Economica*, 42, 123–138.

Carvalho, C., S. Eusepi, E. Moench, and B. Preston (2023): "Anchored inflation expectations," *American Economic Journal: Macroeconomics*, 15, 1–47.

Cavallo, A., G. Cruces, and R. Perez-Truglia (2017): "Inflation expectations, learning, and supermarket prices: Evidence from survey experiments," *American Economic Journal: Macroeconomics*, 9, 1–35.

Cavallo, A., F. Lippi, and K. Miyahara (2024): "Large shocks travel fast," *American Economic Review: Insights*, 6, 558–574.

Cerrato, A. and G. Gitti (2022): "Inflation since covid: Demand or supply," *Working paper.*

Clarida, R., J. Gali, and M. Gertler (1999): "The science of monetary policy: a new Keynesian perspective," *Journal of economic literature*, 37, 1661–1707.

Coibion, O. (2012): "Are the effects of monetary policy shocks big or small?" *American Economic Journal: Macroeconomics*, 4, 1–32.

Coibion, O. and Y. Gorodnichenko (2012): "What can survey forecasts tell us about information rigidities?" *Journal of Political Economy*, 120, 116–159.

——— (2015a): "Information rigidity and the expectations formation process: A simple framework and new facts," *American Economic Review*, 105, 2644–78.

——— (2015b): "Is the Phillips curve alive and well after all? Inflation expectations and the missing disinflation," *American Economic Journal: Macroeconomics*, 7, 197–232.

——— (2025): "Inflation, Expectations and Monetary Policy: What Have We Learned and to What End?" .

Coibion, O., Y. Gorodnichenko, E. S. I. Knotek, and R. Schoenle (2023): “Average Inflation Targeting and Household Expectations,” *Journal of Political Economy: Macroeconomics*.

Coibion, O., Y. Gorodnichenko, L. Kueng, and J. Silvia (2017): “Innocent Bystanders? Monetary policy and inequality,” *Journal of Monetary Economics*, 88, 70–89.

Crump, R. K., S. Eusepi, M. Giannoni, and A. Şahin (2024): “The unemployment–inflation trade-off revisited: The Phillips curve in COVID times,” *Journal of Monetary Economics*, 145, 103580.

Erceg, C., J. Lindé, and M. Trabandt (2024): “Monetary Policy and Inflation Scares,” .

Eusepi, S. and B. Preston (2011): “Expectations, learning, and business cycle fluctuations,” *American Economic Review*, 101, 2844–72.

Evans, G. W. and S. Honkapohja (2001): *Learning and Expectations in Macroeconomics*, Cambridge, MA: MIT Press.

Evans, G. W. and G. Ramey (1995): “Expectation calculation, hyperinflation and currency collapse,” *The new macroeconomics: Imperfect markets and policy effectiveness*, 307.

Faust, J. and J. H. Wright (2013): “Forecasting inflation,” in *Handbook of economic forecasting*, Elsevier, vol. 2, 2–56.

Fisher, J. D., L. Melosi, and S. Rast (2025): “Long-run inflation expectations,” *Working Paper*.

Flynn, J. P. and K. Sastry (2022): “Attention cycles,” *Working Paper*.

Fulton, C. and K. Hubrich (2021): “Forecasting US inflation in real time,” *Econometrics*, 9, 36.

Gabaix, X. (2020): “A Behavioral New Keynesian Model,” *American Economic Review*, 110, 2271–2327.

Gagliardone, L. and M. Gertler (2023): “Oil Prices, Monetary Policy and Inflation Surges,” *Working paper*.

Galí, J. (2015): *Monetary policy, inflation, and the business cycle: an introduction to the new Keynesian framework and its applications*, Princeton University Press.

Gallegos, J. E. (2023): “Inflation persistence, noisy information and the Phillips curve,” *Working paper*.

Gáti, L. (2023): “Monetary Policy & Anchored Expectations An Endogenous Gain Learning Model,” *Journal of Monetary Economics*.

Gemmi, L. and R. Valchev (2023): “Biased surveys,” .

Gitti, G. (2024): “Nonlinearities in the Regional Phillips Curve with Labor Market Tightness,” *Working paper*.

Goldstein, N. (2023): “Tracking inattention,” *Journal of the European Economic Association*.

Gonçalves, S., A. M. Herrera, L. Kilian, and E. Pesavento (2024): "State-dependent local projections," *Journal of Econometrics*, 105702.

Gonzalo, J. and J.-Y. Pitarakis (2002): "Estimation and model selection based inference in single and multiple threshold models," *Journal of econometrics*, 110, 319–352.

Hall, R. E. (2023): "A major shock makes prices more flexible and may result in a burst of inflation or deflation," *Working paper.*

Hamilton, J. D. (1994): *Time series analysis*, Princeton university press.

Hansen, B. E. (2011): "Threshold autoregression in economics," *Statistics and its Interface*, 4, 123–127.

Hilscher, J., A. Raviv, and R. Reis (2025): "How likely is an inflation disaster?" *Review of Financial Studies (forthcoming).*

Hommes, C. and G. Sorger (1998): "Consistent expectations equilibria," *Macroeconomic Dynamics*, 2, 287–321.

Jarociński, M. and P. Karadi (2020): "Deconstructing monetary policy surprises—the role of information shocks," *American Economic Journal: Macroeconomics*, 12, 1–43.

Jo, Y. J. and S. Zubairy (2025): "State-dependent government spending multipliers: downward nominal wage rigidity and sources of business cycle fluctuations," *American Economic Journal: Macroeconomics*, 17, 379–413.

Jonung, L. (1981): "Perceived and expected rates of inflation in Sweden," *The American Economic Review*, 71, 961–968.

Jordà, Ò. (2005): "Estimation and inference of impulse responses by local projections," *American economic review*, 95, 161–182.

Jørgensen, P. L. and K. J. Lansing (2024): "Anchored inflation expectations and the slope of the Phillips curve," *European Economic Review (forthcoming).*

Joussier, R. L., J. Martin, and I. Mejean (2023): "Cost pass-through and the rise of inflation," *Working paper.*

Kahneman, D. (2011): *Thinking, fast and slow*, Farrar, Straus and Giroux.

Kamdar, R. and W. Ray (2024): "Attention-Driven Sentiment and the Business Cycle," *Working Paper.*

Känzig, D. R. (2021): "The macroeconomic effects of oil supply news: Evidence from OPEC announcements," *American Economic Review*, 111, 1092–1125.

Kim, G. and C. Binder (2023): "Learning-through-survey in inflation expectations," *American Economic Journal: Macroeconomics*, 15, 254–278.

Korenok, O. and D. Munro (2024): "The Rockets and Feathers of Inflation Attention," *Working Paper.*

Korenok, O., D. Munro, and J. Chen (2023): "Inflation and attention thresholds," *Review of Economics and Statistics*, 1–28.

Kroner, N. (2024): "Inflation and Attention: Evidence from the Market Reaction to Macro Announcements," *Working paper.*

Lebow, D. and E. Peneva (2024): "Inflation Perceptions During the Covid Pandemic and Recovery," *Working paper.*

Li, T. H., J. M. Londono, and S. Ma (2025): "The Global Transmission of Inflation Uncertainty," .

Link, S., A. Peichl, O. Pfäuti, C. Roth, and J. Wohlfart (2025): "Attention to the Macroeconomy," *Working Paper.*

Lorenzoni, G. (2009): "A theory of demand shocks," *American economic review*, 99, 2050–2084.

Lorenzoni, G. and I. Werning (2023): "Wage-price spirals," *Brookings Papers on Economic Activity*, 2023, 317–393.

Macaulay, A. (2022): "Shock transmission and the sources of heterogeneous expectations," *Working Paper.*

Maćkowiak, B., F. Matějka, and M. Wiederholt (2018): "Dynamic rational inattention: Analytical results," *Journal of Economic Theory*, 176, 650–692.

——— (2023): "Rational inattention: A review," *Journal of Economic Literature*, 61, 226–273.

Mackowiak, B. and M. Wiederholt (2009): "Optimal sticky prices under rational inattention," *American Economic Review*, 99, 769–803.

Marcet, A. and J. P. Nicolini (2003): "Recurrent hyperinflations and learning," *American Economic Review*, 93, 1476–1498.

Matějka, F. and A. McKay (2015): "Rational inattention to discrete choices: A new foundation for the multinomial logit model," *American Economic Review*, 105, 272–98.

McClure, E. M., V. Yaremko, O. Coibion, and Y. Gorodnichenko (2025): "The macroeconomic expectations of us managers," *Journal of Money, Credit and Banking*, 57, 683–716.

Mitra, A. (2023): "Imperfect Information and Slow Recoveries in the Labor Market," *Working paper.*

Moll, B. (2024): "The Trouble with Rational Expectations in Heterogeneous Agent Models: A Challenge for Macroeconomics," *Working paper.*

Morris, S. and H. S. Shin (2002): "Social value of public information," *american economic review*, 92, 1521–1534.

Newey, W. K. and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation," *Econometrica*, 55, 703–708.

Paciello, L. and M. Wiederholt (2014): "Exogenous information, endogenous information, and optimal monetary policy," *Review of Economic Studies*, 81, 356–388.

Pfäuti, O. (2025): "Inflation–who cares? Monetary Policy in Times of Low Attention," *Journal of Money, Credit and Banking*, 57, Issue 5.

Pfäuti, O. and F. Seyrich (2023): "A behavioral heterogeneous agent New Keynesian model," *Working paper.*

Ramey, V. A. and S. Zubairy (2018): "Government spending multipliers in good times and in bad: evidence from US historical data," *Journal of political economy*, 126, 850–901.

Reis, R. (2006a): "Inattentive consumers," *Journal of monetary Economics*, 53, 1761–1800.

——— (2006b): "Inattentive producers," *The Review of Economic Studies*, 73, 793–821.

——— (2022a): *The burst of high inflation in 2021-22: How and why did we get here?*, Bank for International Settlements, Monetary and Economic Department.

——— (2022b): "Losing the inflation anchor," *Brookings Papers on Economic Activity*, 2021, 307–379.

Rotemberg, J. J. (1982): "Sticky prices in the United States," *Journal of political economy*, 90, 1187–1211.

Sargent, T. J. (1991): "Equilibrium with signal extraction from endogenous variables," *Journal of Economic Dynamics and Control*, 15, 245–273.

Shapiro, A. H. (2024): "Decomposing Supply-and Demand-Driven Inflation," *Journal of Money, Credit and Banking.*

Sims, C. A. (2003): "Implications of rational inattention," *Journal of monetary Economics*, 50, 665–690.

——— (2010): "Rational inattention and monetary economics," in *Handbook of monetary economics*, Elsevier, vol. 3, 155–181.

Stock, J. H. and M. W. Watson (2016): "Core inflation and trend inflation," *Review of Economics and Statistics*, 98, 770–784.

Tenreyro, S. and G. Thwaites (2016): "Pushing on a string: US monetary policy is less powerful in recessions," *American Economic Journal: Macroeconomics*, 8, 43–74.

Vellekoop, N. and M. Wiederholt (2019): "Inflation expectations and choices of households," *Working Paper.*

Weber, M., B. Candia, H. Afrouzi, T. Ropele, R. Lluberas, S. Frache, B. Meyer, S. Kumar, Y. Gorodnichenko, D. Georgarakos, et al. (2025): "Tell Me Something I Don't Already Know: Learning in Low-and High-Inflation Settings," *Econometrica*, 93, 229–264.

Weber, M., F. D'Acunto, Y. Gorodnichenko, and O. Coibion (2022): "The subjective inflation expectations of households and firms: Measurement, determinants, and implications," *Journal of Economic Perspectives*, 36, 157–184.

Werning, I. (2022): "Expectations and the Rate of Inflation," *Working Paper*.

Woodford, M. (2003): *Interest and Prices*, Princeton University Press.

York, J. (2023): "Do Household Inflation Expectations Respond to Macroeconomic Data Releases?" *Working Paper*.

Yotzov, I., N. Bloom, P. Bunn, P. Mizen, and G. Thwaites (2024): "The Speed of Firm Response to Inflation," *Working Paper*.

Online Appendix

A Appendix to Section 2.1

A.1 Details and microfoundations

In Section 2.1, I show how attention may jump when inflation crosses a certain threshold. In this Section, I provide the mathematical details necessary to arrive at these results. The model builds on Pfäuti (2025).[36] The key novelties are, first, that agents get a certain level of attention for free, which I denote $\gamma_{\pi,L}$, where L denotes that this is a relatively Low level of attention. As discussed in Section 2.1, attention γ_π is negatively related to $\sigma^2_{\pi|s}/\sigma^2_\pi$, where σ^2_π denotes the agent's prior uncertainty about inflation π, and $\sigma^2_{\pi|s}$ the posterior uncertainty given the signal s. Second, if the agent wants to increase her attention to inflation above $\gamma_{\pi,L}$, she has two options to do so. First, the agent can choose to pay a fixed cost Λ in which case she can increase her attention to a fixed and known level of attention, $\gamma_{\pi,H}$, where the subscript H indicates that this is a Higher level of attention than the one the agent gets for free. The second option to increase attention above the free level is by paying a standard mutual information cognitive cost and adjusting her attention to inflation smoothly. Conceptually, the fixed cost captures a lumpy one-off action, such as googling "inflation" to obtain the headline result, which delivers a discrete jump in attention to the predetermined level $\gamma_{\pi,H}$. In contrast, the mutual-information cost captures the continuous cognitive effort of processing many small bits of information, such as noticing and processing individual price changes while grocery shopping, which may lead to a smooth increase in attention. When paying the mutual information cost, the agent does not need to pay the fixed cost.

The agent believes that inflation in the next period, π', depends on inflation today, π, as follows:[37]

$$\pi' = \tilde{\rho}_\pi \pi + \nu,$$

where $\nu \sim i.i.N.(0, \sigma^2_\nu)$, and $\tilde{\rho}_\pi \in [0,1]$ denotes the perceived persistence of inflation. Given this perceived law of motion, the full-information forecast π^{e*} is $\pi^{e*} \equiv \tilde{\rho}_\pi \pi$. But because π is not perfectly observable, the actual forecast will deviate from this full-information forecast. These deviations are costly and the agent can pay more attention in order to reduce these deviations. Thus, the agent's choice is not only about how to form her expectations given certain information, but about how to choose and process this information optimally, while taking into account how this will later affect

[36] The method proposed in Pfäuti (2025) builds on Maćkowiak et al. (2023) and Vellekoop and Wiederholt (2019), see also Weber et al. (2025) for a similar model but with the goal of explaining their RCT findings.

[37] Agents' subjective model (or their perceived law of motion) does not necessarily need to be consistent with the actual law of motion (see Andre et al. (2022) and Macaulay (2022) for empirical evidence that the subjective models agents hold may not necessarily be consistent with the actual behavior of the economy or with experts' subjective models). Empirically, AR(1) models for inflation tend to perform quite well (see, e.g., Canova (2007) or Faust and Wright (2013)). Fulton and Hubrich (2021) show that AR(1) models are especially hard to beat in forecasting inflation in real time.

her forecast. That is, she chooses the form of the signal s she receives about current inflation. But processing information is costly, and hence, it cannot be optimal to acquire different signals that lead to identical forecasts. Due to this one-to-one relation of signal and forecast, I will directly let the agent choose the joint distribution of π^e and π, denoted $f(\pi^e, \pi)$, instead of working with the signal. I use $g(\pi)$ to denote the agent's prior, which is assumed to be Gaussian; $\pi \sim N\left(\pi^{prior}, \sigma_\pi^2\right)$. To link the agent's attention choice to the level of inflation, I assume that the agent's prior uncertainty, σ_π^2, increases smoothly (strictly monotonically) with the level of inflation. The agent knows her own subjective prior uncertainty σ_π^2, but she cannot back out the inflation rate π from observing σ_π^2.

Let $U(\pi^e, \pi)$ denote the negative of the utility loss that is incurred when the agent's forecast deviates from the forecast under full information. The agent receives a minimum amount of attention for free, resulting in a fixed ratio of posterior uncertainty to prior uncertainty $\sigma^2_{\pi|s,L}/\sigma^2_\pi$. The agent can reduce her posterior uncertainty further—put differently, the agent can decide to pay costly attention to inflation—in two possible ways. Either by paying the mutual information cost $C(f(\pi^e, \pi))$ and choosing the posterior uncertainty $\sigma^2_{\pi|s} \in [0, \sigma^2_{\pi|s,L}]$ optimally. The standard assumption in the literature (see, e.g., Maćkowiak et al. (2023)) is that $C(f(\pi^e, \pi))$ is linear in mutual information $I(\pi; \pi^e)$, i.e., the expected reduction in entropy of π due to knowledge of π^e:

$$I(\pi; \pi^e) = \left(H(\pi) - E\left[H(\pi|\pi^e)\right]\right),$$

where $H(\pi) = -\int g(\pi) log(g(\pi)) d\pi$ is the entropy of π, and $H(\pi|\pi^e)$ is the conditional entropy. Given that the agent gets attention $\gamma_{\pi,L}$ for free, the agent only pays the cost when paying additional attention:

$$C(f(\pi^e, \pi)) = \frac{1}{\lambda}\left[I(\pi; \pi^e) - I_{free}\right],$$

where I_{free} denotes the cost the agent would have to pay to achieve the free level of attention $\gamma_{\pi,L}$, and $\frac{1}{\lambda}$ denotes the parameter that governs the cost of information. Alternatively, the agent can pay the fixed cost Λ in which case she reduces her posterior uncertainty to a fixed ratio of $\sigma^2_{\pi|s,H}/\sigma^2_\pi < \sigma^2_{\pi|s,L}/\sigma^2_\pi$.[38] Taken together, the agent's problem can be written as follows:

$$\max_{f(\pi^e,\pi)} \int U(\pi^e, \pi) f(\pi^e, \pi) d\pi d\pi^e - C(f(\pi^e, \pi)) \cdot \mathbf{1}\{\gamma_\pi > \gamma_{\pi,L} \wedge \gamma_\pi \neq \gamma_{\pi,H}\} - \Lambda \cdot \mathbf{1}\{\gamma_\pi = \gamma_{\pi,H}\}$$
$$\text{subject to} \int f(\pi^e, \pi) d\pi^e = g(\pi), \text{ for all } \pi.$$

The indicator functions $\mathbf{1}\{\cdot\}$ capture the alternative attention strategies. If $\mathbf{1}\{\gamma_\pi > \gamma_{\pi,L} \wedge \gamma_\pi \neq \gamma_{\pi,H}\} = \mathbf{1}\{\gamma_\pi = \gamma_{\pi,H}\} = 0$, the agent does not pay any costly attention to inflation. If $\mathbf{1}\{\gamma_\pi = \gamma_{\pi,H}\} = 1$, she pays the fixed cost, and if $\mathbf{1}\{\gamma_\pi > \gamma_{\pi,L} \wedge \gamma_\pi \neq \gamma_{\pi,H}\} = 1$, she pays the mutual

[38]The fixed cost could also be interpreted through the lens of the System 1/System 2 framework (Kahneman, 2011). In the low-attention regime, agents operate in the fast, heuristic System 1 mode. The fixed cost Λ could then be interpreted as capturing the cognitive cost of switching to the more deliberate System 2 mode—a switch that agents only make when the perceived stakes are high enough. This interpretation is consistent with the experimental evidence discussed in Kahneman (2011) that System 2 engagement is triggered by perceived importance.

information cost. I assume that if the agent decides to pay the mutual information cost, she does not have to pay the fixed cost. This assumption also implies that the agent's optimal attention choice when paying the mutual information cost can be higher or lower than $\gamma_{\pi,H}$. I discuss the case in which the agent always has to pay the fixed cost below. Throughout my analysis, I assume that it is cheaper for the agent to obtain an attention level of $\gamma_{\pi,H}$ by paying the fixed cost rather than the mutual information cost. Otherwise, the fixed cost would be obsolete.

The objective function $U(.)$ is assumed to be quadratic:

$$U(\pi^e, \pi) = -\chi \left(\tilde{\rho}_\pi \pi - \pi^e\right)^2,$$

where χ measures the stakes of making a mistake.

The signals the agent obtains are Gaussian:[39]

$$s = \pi + \varepsilon,$$

with noise $\varepsilon \sim i.i.N.(0, \sigma_\varepsilon^2)$. The noise variance σ_ε^2 is chosen optimally and reflects the agent's attention.

Given the Gaussian prior and signal, the optimal forecast is given by $\pi^e = \tilde{\rho}_\pi E[\pi|s]$, where $E[\pi|s]$ is the conditional expectation, and Bayesian updating implies:

$$\pi^e = \tilde{\rho}_\pi (1 - \gamma_\pi) \pi^{prior} + \tilde{\rho}_\pi \gamma_\pi s, \tag{25}$$

where $\gamma_\pi = 1 - \frac{\sigma_{\pi|s}^2}{\sigma_\pi^2} \in [0, 1]$ measures the agent's attention to inflation, and π^{prior} denotes the prior mean of π.

With Gaussian signals, the cost function $C(f(\pi^e, \pi))$ can be expressed as:

$$C(f(\pi^e, \pi)) = \frac{1}{\lambda}\left[I(\pi; \pi^e) - I_{free}\right] = \frac{1}{\lambda}\left[\frac{1}{2}\log\left(\frac{\sigma_\pi^2}{\sigma_{\pi|s}^2}\right) - \frac{1}{2}\log\left(\frac{\sigma_\pi^2}{\sigma_{\pi|s,L}^2}\right)\right] = \frac{1}{2\lambda}\log\left(\frac{\sigma_{\pi|s,L}^2}{\sigma_{\pi|s}^2}\right), \tag{26}$$

for $\sigma_{\pi|s}^2 < \sigma_{\pi|s,L}^2$.

For this interior solution, we can express the attention choice problem in terms of γ_π as follows:

$$\max_{\gamma_\pi \in [\gamma_{\pi,L}, 1] \wedge \gamma_\pi \neq \gamma_{\pi,H}} \left[-\chi \tilde{\rho}_\pi^2 (1 - \gamma_\pi) \sigma_\pi^2 - \frac{1}{2\lambda} log\left(\frac{1 - \gamma_{\pi,L}}{1 - \gamma_\pi}\right)\right], \tag{27}$$

where I used $\gamma_\pi = 1 - \frac{\sigma_{\pi|s}^2}{\sigma_\pi^2}$ in the expression for $C(f(\pi^e, \pi))$ in equation (26) and in $U(\pi^e, \pi)$.

[39] In a setup with a quadratic objective function, Gaussian prior and a linear mutual information cost, Gaussian signals are the optimal and unique solution (Matějka and McKay (2015) and Maćkowiak et al. (2023)). This also applies to the interior solution in my case, i.e., the case with $\mathbf{1}\{\gamma_\pi > \gamma_{\pi,L} \& \gamma_\pi \neq \gamma_{\pi,H}\} = 1$. At the corner solution, where attention is equal to the free level of attention $\gamma_{\pi,L}$ or equal to $\gamma_{\pi,H}$, any signal family that yields the required posterior variance is optimal; thus Gaussian signals are still optimal but not uniquely so at that point. To keep the model tractable, I assume that the signals in these cases are Gaussian, too.

Let γ_π^{MI} denote the solution to the problem in equation (27), featuring the Mutual Information cost. The following Proposition characterizes the optimal attention choice.

Proposition 1 *The optimal attention choice γ_π^* is determined as follows:*

$$\gamma_\pi^* = \begin{cases} \gamma_{\pi,L}, & \text{if } A \geq B \wedge A \geq C \\ \gamma_{\pi,H}, & \text{if } B > A \wedge B \geq C \\ \gamma_\pi^{MI}, & \text{otherwise,} \end{cases}$$

where

$$\gamma_\pi^{MI} = 1 - \frac{1}{2\lambda\chi\tilde{\rho}_\pi^2\sigma_\pi^2} \tag{28}$$

is the solution to problem (27), and

$$\begin{aligned} A &\equiv -\chi\tilde{\rho}_\pi^2(1-\gamma_{\pi,L})\sigma_\pi^2 \\ B &\equiv -\chi\tilde{\rho}_\pi^2(1-\gamma_{\pi,H})\sigma_\pi^2 - \Lambda \\ C &\equiv -\chi\tilde{\rho}_\pi^2(1-\gamma_\pi^{MI})\sigma_\pi^2 - \frac{1}{2\lambda}\log\left(\frac{1-\gamma_{\pi,L}}{1-\gamma_\pi^{MI}}\right), \end{aligned}$$

capture the benefits (minus the attention costs) of the respective attention choices.

Proposition 1 determines the optimal attention choice γ_π^*. To see how the level of inflation affects the agent's optimal attention choice, recall that a higher inflation rate increases the agent's prior uncertainty σ_π^2. Since $\gamma_{\pi,H} > \gamma_{\pi,L}$, we can immediately see that A decreases more strongly with inflation than B because $(1-\gamma_{\pi,L}) > (1-\gamma_{\pi,H})$. Hence, there is an inflation threshold $\bar{\pi}$ at which attention jumps from $\gamma_{\pi,L}$ to $\gamma_{\pi,H}$, unless $C > B$ at $\gamma_\pi^{MI} \in (\gamma_{\pi,L}, \gamma_{\pi,H})$. The sufficient condition for the threshold to emerge is:

$$\chi\tilde{\rho}_\pi^2\sigma_\pi^2\left(\gamma_{\pi,H} - \gamma_\pi^{MI}\right) > \Lambda - \frac{1}{2\lambda}\log\left(\frac{1-\gamma_{\pi,L}}{1-\gamma_\pi^{MI}}\right), \tag{29}$$

for $\gamma_\pi^{MI} \in (\gamma_{\pi,L}, \gamma_{\pi,H})$, which says that increasing one's attention to inflation by paying the fixed cost Λ is relatively cheap compared to paying the mutual information cost. As I will show later, the data favors the threshold-like behavior, indicating that condition (29) holds in the data. Hence, I calibrate the parameters such that the condition holds. If the agent would always have to pay the fixed cost when increasing her attention above $\gamma_{\pi,L}$, i.e., also when paying the mutual information cost, the threshold-like behavior would always emerge. To see this, note that the sufficient condition in this case would be given by:

$$\chi\tilde{\rho}_\pi^2\sigma_\pi^2\left(\gamma_{\pi,H} - \gamma_\pi^{MI}\right) > -\frac{1}{2\lambda}\log\left(\frac{1-\gamma_{\pi,L}}{1-\gamma_\pi^{MI}}\right). \tag{30}$$

Because the right-hand side of equation (30) is negative, this condition can only be violated if $\gamma_\pi^{MI} > \gamma_{\pi,H}$. But because σ_π^2 is a smooth function of π, there has to be an inflation level at which attention jumps discretely from $\gamma_{\pi,L}$ to $\gamma_{\pi,H}$, before it may increase to γ_π^{MI} at a higher inflation level.

If condition (29) holds, we can solve analytically for the threshold $\bar{\pi}$, at which attention jumps from $\gamma_{\pi,L}$ to $\gamma_{\pi,H}$, by setting $A = B$. Solving for the implied prior uncertainty, we obtain the following implicit expression for the threshold:

$$\bar{\pi} = h^{-1}\left(\frac{\Lambda}{(\gamma_{\pi,H} - \gamma_{\pi,L})\chi\tilde{\rho}_\pi^2}\right), \tag{31}$$

where $h(\pi) = \sigma_\pi^2(\pi)$ denotes prior uncertainty as a function of inflation and $h^{-1}\left(\frac{\Lambda}{(\gamma_{\pi,H}-\gamma_{\pi,L})\chi\tilde{\rho}_\pi^2}\right)$ its inverse evaluated at $\frac{\Lambda}{(\gamma_{\pi,H}-\gamma_{\pi,L})\chi\tilde{\rho}_\pi^2}$. Because the agent cannot invert σ_π^2 to recover π, she does not know the exact inflation rate at which the attention jump occurs; the threshold is defined in her information set only in terms of the observable prior uncertainty (observable for the agent), not in terms of the unobserved π.[40] The expression (31) shows that the threshold is higher when the fixed cost Λ is higher, as in this case, increasing attention to $\gamma_{\pi,H}$ is costlier. Thus, the benefit of paying the fixed cost has to be higher, which is the case at a higher inflation rate. The threshold is lower when the increase in attention, $(\gamma_{\pi,H} - \gamma_{\pi,L})$, is larger, when the stakes χ are higher, or when the perceived persistence $\tilde{\rho}_\pi$ is higher.

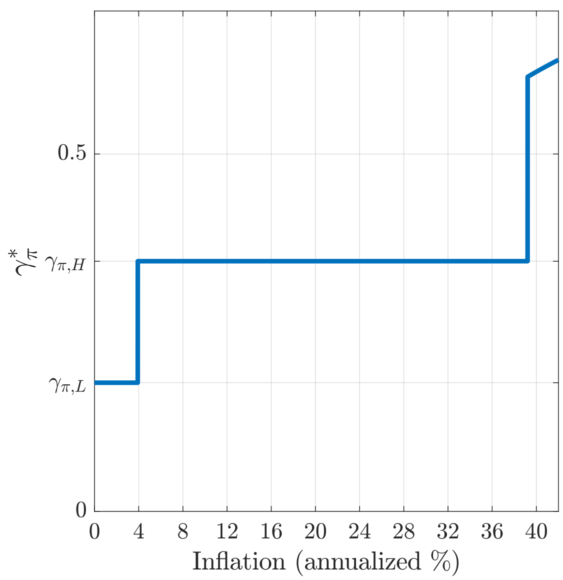
Figure 10 shows γ_π^* graphically as a function of π, for the case in which prior uncertainty is a linear function of inflation, $\sigma_\pi^2 = \underline{\sigma}_\pi^2 + a \cdot \pi$.[41] The Figure illustrates how agents optimally allocate their attention to inflation. At relatively low levels of inflation, agents do not pay more attention than what they obtain for free. The reason is that their prior uncertainty is low in these times and hence, the benefit of increasing their attention is less than the costs of doing so. Once inflation exceeds a threshold, here at $\bar{\pi} = 3.91\%$, the agent pays the fixed cost Λ and her attention jumps to $\gamma_{\pi,H}$. As equation (31) shows, the threshold is endogenous and depends on the underlying parameters, and the functional form assumption of σ_π^2. Attention then stays constant at $\gamma_{\pi,H}$ even at higher levels of inflation. Up until a second threshold at which attention jumps once more from where on it increases smoothly with inflation. This second jump mirrors the agent's decision to pay the mutual information cost instead of the fixed cost. The reason is that with the fixed cost, the agent's attention is stuck at $\gamma_{\pi,H}$ and hence, the agent's losses from not being fully attentive increase as inflation increases. Once these losses are large enough, the agent decides to pay the mutual information cost which allows her to increase attention even further. Where exactly this threshold lies and to what attention level the agent jumps depends on all of the underlying parameters. My empirical results suggest, however, that this last part has not been relevant in the U.S. since the beginning of my sample. Nevertheless,

[40] Note that π is unobservable for the agent, but it is observable for the econometrician. In contrast, the agent observes her prior uncertainty but it is unobservable for the econometrician.

[41] I set $\underline{\sigma}_\pi^2 = 3$, and $a = 0.74$. The rest of the parameters are set as follows: $\gamma_{\pi,L} = 0.18$, $\gamma_{\pi,H} = 0.35$, $\chi = 1$, $\tilde{\rho} = 1$, $\lambda = 0.04$ and $\Lambda = 1$. This results in a threshold $\bar{\pi} = 3.91\%$ at which attention jumps from $\gamma_{\pi,L}$ to $\gamma_{\pi,H}$. These parameters are chosen to be consistent with the empirical results obtained later.

it illustrates that attention may increase even further if inflation reaches these very high levels.

Figure 10: Optimal attention



Notes: This figure illustrates Proposition 1 graphically, i.e., the figure shows optimal attention to inflation, γ^*_π, as a function of the level of inflation π.

The following Appendix A.2 provides an alternative microfoundation in which the fixed cost is a function of inflation and in which the agent needs to pay the fixed cost and the mutual information cost in order to increase her level of attention above the free level of attention. A similar pattern as in Figure 10 emerges.

A.2 An alternative microfoundation

In this section, I provide an alternative microfoundation giving rise to a threshold of attention. Many of the features of this model are the same as the one presented in the previous Section. I will therefore focus on the main differences solely and refer the reader to the previous Section for more details.

As in Section A.1, agents get a certain level of attention for free, which I denote $\gamma_{\pi,L}$. In addition to the standard mutual information cognitive cost the agent has to pay a fixed cost Λ if she decides to pay more attention than $\gamma_{\pi,L}$. The fixed cost depends negatively on the level of inflation, and I assume that $\Lambda(\pi)$ is strictly decreasing in π. The agent only observes the cost $\Lambda(\pi)$ but she cannot infer π from it, that is, she does not know how π maps into $\Lambda(\pi)$. A key difference from the model in Section A.1 is that paying the fixed cost itself does not increase the agent's attention (or, equivalently, her obtained signal's precision). A possible interpretation of this fixed cost is that the agent first needs to find out what the reliable information sources regarding inflation are before processing the information and doing so becomes easier when inflation is higher. I assume that all parameters except $\Lambda(\pi)$ (including prior uncertainty) are constant and hence, do not depend on inflation.

The agent's perceived law of motion of inflation, and the loss function $U(\pi^e, \pi)$, are the same as in Section A.1. Let $C(f(\pi^e, \pi))$ denote the cost of information function. Then, the agent's problem is given by:

$$\max_{f(\pi^e,\pi)} \int U(\pi^e,\pi)f(\pi^e,\pi)d\pi d\pi^e - C(f(\pi^e,\pi)) - \Lambda(\pi)\cdot \mathbf{1}\{\gamma_\pi > \gamma_{\pi,L}\}$$
$$\text{subject to } \int f(\pi^e,\pi)d\pi^e = g(\pi), \text{ for all } \pi,$$

where $g(\pi)$ is the agent's prior, which is assumed to be Gaussian; $\pi \sim N(\pi^{prior}, \sigma_\pi^2)$, and σ_π^2 denotes prior uncertainty. The cost function is the same as in Section A.1. The signals the agent obtains are again Gaussian.[42]

Following the same steps as in Section A.1, we can express the attention problem in terms of γ_π:

$$\max_{\gamma_\pi \in [\gamma_{\pi,L},1]} \left[-\chi\tilde{\rho}_\pi^2(1-\gamma_\pi)\sigma_\pi^2 - \frac{1}{2\lambda} log\left(\frac{1-\gamma_{\pi,L}}{1-\gamma_\pi}\right) - \Lambda(\pi)\cdot \mathbf{1}\{\gamma_\pi > \gamma_{\pi,L}\}\right], \tag{32}$$

where I used $\gamma_\pi = 1 - \frac{\sigma^2_{\pi|s}}{\sigma^2_\pi}$. Solving the problem in (32) and defining auxiliary variables A and B:

$$A \equiv -\chi\tilde{\rho}_\pi^2(1-\gamma_{\pi,L})\sigma_\pi^2$$
$$B \equiv -\chi\tilde{\rho}_\pi^2(1-\gamma_\pi)\sigma_\pi^2 - \frac{1}{2\lambda} log\left(\frac{1-\gamma_{\pi,L}}{1-\gamma_\pi}\right) - \Lambda(\pi).$$

yields the following Proposition that is akin to Proposition 1, characterizing the *optimal* level of attention.

Proposition 2 *The optimal attention choice* γ_π^* *is given by:*

$$\gamma_\pi^* = \begin{cases} \gamma_{\pi,L}, & if\ A \geq B \\ 1 - \frac{1}{2\lambda\chi\tilde{\rho}_\pi^2\sigma_\pi^2}, & if\ A < B. \end{cases} \tag{33}$$

Proposition 2 pins down the optimal attention choice γ_π^*. A higher inflation rate increases B while it leaves A unaffected. Thus, attention jumps discretely from $\gamma_{\pi,L}$ to a higher level given by $\gamma_{\pi,H} = 1 - \frac{1}{2\lambda\chi\tilde{\rho}_\pi^2\sigma_\pi^2}$ when inflation exceeds some threshold $\bar{\pi}$.[43]

Figure 11 shows γ_π^* graphically as a function of π.[44] The figure illustrates that optimal attention to inflation takes on two values, depending on the level of inflation. If inflation is relatively low (below 4% in the Figure), $\gamma_\pi^* = \gamma_{\pi,L}$ as the cost of paying attention is too high relative to the benefits

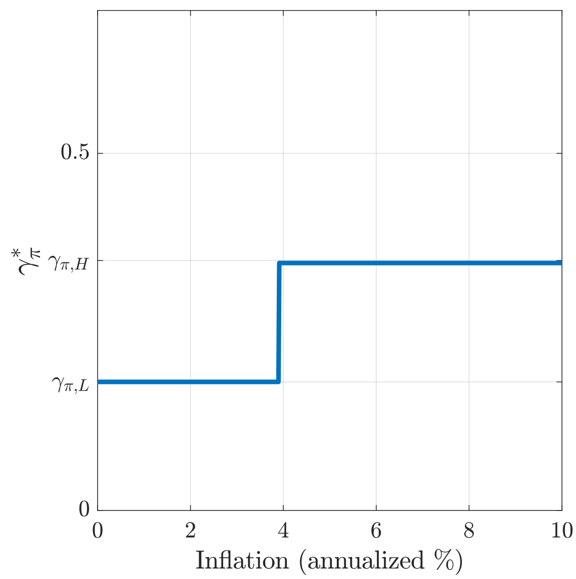
[42]The arguments in Matějka and McKay (2015) and Maćkowiak et al. (2023) regarding the optimality and uniqueness of Gaussian signals also apply to the interior solution in my case, i.e., once the fixed cost is paid. At the corner solution, where attention is equal to the free level of attention $\gamma_{\pi,L}$, any signal family that yields the required posterior variance is costless; thus Gaussian signals are still optimal but not uniquely so at that point. To keep the model tractable, I assume that the free signals are Gaussian.

[43]Proposition 2 and the statements following the Proposition require that $\gamma_{\pi,H} \in [\gamma_{\pi,L}, 1]$. If this would not be the case, the agent would never pay more attention than $\gamma_{\pi,L}$.

[44]I assume a linear functional form for $\Lambda(\pi) = \max\{0, a - b\pi\}$ with $a = 0.09$ and $b = 1/112$. The rest of the parameters is set as follows: $\gamma_{\pi,L} = 0.18$, $\chi = 1$, $\tilde{\rho} = 1$, $\lambda = 0.255$ and $\sigma_\pi^2 = 3$. This results in $\gamma_{\pi,H} = 0.35$. These parameters are chosen to be consistent with the empirical results.

of paying more attention. As inflation passes the threshold at 4%, however, attention immediately jumps up to $\gamma_{\pi,H} = 1 - \frac{1}{2\lambda\chi\tilde{\rho}_\pi^2\sigma_\pi^2}$.

Figure 11: Optimal attention



Notes: This figure illustrates Proposition 2 graphically, i.e., the figure shows optimal attention to inflation, γ_π^*, as a function of the level of inflation π.

How high attention is once it jumps, depends on the cost parameter λ (but not $\Lambda(\pi)$), as well as the stakes χ. If the cost of information is lower or the stakes are higher, $\gamma_{\pi,H}$ is higher. Additionally, $\gamma_{\pi,H}$ depends positively on the perceived persistence $\tilde{\rho}_\pi$ and the prior uncertainty σ_π^2. All of these parameters, as well as the fixed cost $\Lambda(\pi)$, also affect the threshold $\bar{\pi}$. Thus, $\gamma_{\pi,H}$ as well as the threshold are endogenous and are the outcome of the attention choice problem.

B Appendix to Empirical Section 2

In this section, I discuss a wide range of robustness checks and sensitivity analyses related to my main findings in Section 2.

Median expectations. The results for median expectations are similar, as the second row in Table 4 shows (the first row replicates the baseline results for comparison). Attention in both regimes tends to be somewhat lower when using median expectations and the attention threshold higher. Median expectations tend to be less volatile than average expectations, therefore, it is not surprising that the estimated attention levels—capturing the updating gains—are lower for median expectations. Nevertheless, the null hypothesis that the two attention levels across regimes are equal is marginally rejected (with a p-value of 0.058, see last column). Again, I find that the Bayesian information criterion selects having one threshold only.

Table 4: Estimated attention levels and the attention threshold

	Threshold $\hat{\bar{\pi}}$	Low Att. $\hat{\gamma}_{\pi,L}$	High Att. $\hat{\gamma}_{\pi,H}$	p-val. $H_0 : \gamma_{\pi,L} = \gamma_{\pi,H}$
Mean expectations	3.91%	0.18	0.35	0.000
s.e.		(0.018)	(0.042)	
Median expectations	4.44%	0.16	0.22	0.058
s.e.		(0.018)	(0.030)	
Quarterly frequency	3.21%	0.14	0.38	0.013
s.e.		(0.031)	(0.099)	

Notes: This table shows the results from regression (4), where $\hat{\bar{\pi}}$ denotes the estimated threshold, $\hat{\gamma}_{\pi,L}$ and $\hat{\gamma}_{\pi,H}$ the estimated attention levels when inflation is below or above the threshold, respectively. The last column shows the p-value for the null hypothesis that the two attention levels are equal. Standard errors for the attention estimates are computed using the delta method, based on Newey–West heteroskedasticity- and autocorrelation-robust covariance estimates with 12 lags.

Quarterly observations. The last two rows in Table 4 show the results when using observations at quarterly frequency for the period 1960Q2-2023Q2. We see that the estimated threshold is somewhat lower at 3.21%. The estimated attention levels within regime are very similar to the baseline monthly specification and again, the difference in attention across regimes is highly statistically significant.

Alternative threshold-defining variables. When I use average inflation of the last three months as the threshold-defining variable instead of the last month's inflation rate, I estimate a threshold at 3.45% and attention levels of $\gamma_{\pi,L} = 0.18$ and $\gamma_{\pi,H} = 0.33$.

When using the BIC to select whether to use the last-period's inflation rate or the inflation volatility as the threshold-defining variable, the data prefer using the level rather than the volatility. To arrive at this finding, I first estimate a GARCH(1,1) model for quarterly inflation and then use the predicted time-varying volatility as the threshold defining variable. Consistent with my baseline results, I find that the specification with only one threshold is preferred to specifications with multiple thresholds, and the estimated attention parameters are similar: when last-period's volatility was below the threshold, the implied attention parameter is 0.23 and it increases to 0.31 when inflation volatility exceeds the threshold. The threshold is at a volatility of 5.25, indicating that about 17% of the periods fall into the high-attention regime which is lower than the estimated values based on the specification in terms of levels. A key difference of the volatility specification compared to my baseline specification, is that the specification in terms of inflation volatility does not differentiate between periods of pronounced inflation rises and declines. For example, at the outbreak of the Great Financial Crisis, quarterly CPI inflation dropped briefly but quite substantially, leading to a spike in volatility. However, at least households did not seem to pay attention to inflation during that time period when looking at Google Trends data, which may explain why the specification in terms of inflation levels is preferred over the volatility specification.[45]

Alternative surveys. A drawback of the Survey of Consumers is that it surveys households at most twice. Therefore, it is not possible to examine how individual households update their expectations and how they change their updating behavior with the level of inflation. Focusing on mean (and median) expectations helps to overcome this issue but it also ignores potentially important cross-sectional information. Therefore, I also estimate the attention levels using individual consumer inflation expectations from the Survey of Consumer Expectations. When last month's inflation was below the threshold of 3.91%, I estimate an attention level of 0.08. When inflation was above the threshold, attention increases to 0.33, and I reject the null hypothesis that the two estimates are equal (p-value of 0.000). Thus, I obtain similar, though slightly lower, attention parameters than in my baseline estimates even though this survey is only available since 2013.

When using the firm expectations from the Survey of Firms' Inflation Expectations (SoFIE) from the Cleveland Fed—with the caveat that this data is only available at quarterly frequency and only since 2018Q2, resulting in only 25 observations—I estimate that attention of firms has increased from practically 0.07 up to 0.18 as inflation exceeded the threshold of 3.91%.

A multivariate system of inflation expectations. As the perceived law of motion of inflation is an AR(1) process, inflation expectations under limited attention only depend on prior beliefs and inflation itself. In principle, however, it could be possible that inflation expectations also depend on other variables such as unemployment.[46] To take this into account, I therefore now estimate the

[45] Of course, this may look very different for professional forecasters.

[46] Kamdar and Ray (2024), for example, show that consumers tend to expect disinflation when expecting an economic expansion.

regression:

$$\pi^e_{t+1|t} = \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\left[\beta_{0,L} + \beta_{1,L}\pi^e_{t|t-1} + \beta_{2,L}\left(\pi_t - \pi^e_{t|t-1}\right)\right] \tag{34}$$

$$+ \left(1 - \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\right)\left[\beta_{0,H} + \beta_{1,H}\pi^e_{t|t-1} + \beta_{2,H}\left(\pi_t - \pi^e_{t|t-1}\right)\right] \tag{35}$$

$$+ \beta_3 U^e_{t|t-1} + \beta_4\left(U_t - U^e_{t|t-1}\right) + \tilde{\epsilon}_t, \tag{36}$$

where $U^e_{t|t-1}$ denotes expectations about unemployment changes and U_t are unemployment changes (see Appendix D.1 for further details). I find that both, β_3 and β_4, are insignificant ($\hat{\beta}_3 = -0.13$ and $\hat{\beta}_4 = 0.04$ with p-values of 0.332 and 0.337, respectively). The estimated attention-to-inflation parameters are also barely affected. In fact, I estimate attention in the low regime to be equal to 0.18 and 0.34 in the high regime.

Regional variation. The Survey of Consumers also provides the regions consumers are in, which allows me to leverage regional variation to test for the robustness of my results. In particular, I use the four census regions West, North East, North Central (or Midwest) and South. I control for region fixed effects and estimate that attention is 0.22 with a standard error of 0.023 (standard errors are robust with respect to heteroskedasticity as well as serial and cross-sectional correlation) when last-period's inflation is below the threshold of 3.91%. Attention increases to 0.42 (s.e. of 0.051) when inflation exceeds the threshold. The two estimates are highly statistically significantly different from each other (p-value of 0.000). These results indicate that my baseline results remain robust when considering regional variation.[47]

An alternative attention measure. In a recent paper, Bracha and Tang (2025) develop an alternative measure of (in)attention to inflation. They use the Michigan Survey of Consumers. In particular, their measure of inattention uses the share of individuals who say that they "Don't know" when asked for an estimate of current inflation. I apply their approach to an extended sample covering 1982-2024. I do this at quarterly and monthly frequency. Additionally, I also do this at the regional level to increase the sample size and to account for possible regional differences.

I then estimate the following threshold regression:

$$\gamma_{BT,t} = \begin{cases} \alpha_L, & \text{if } \pi_{t-1} < \bar{\pi} \\ \alpha_H, & \text{if } \pi_{t-1} \geq \bar{\pi}, \end{cases} \tag{37}$$

where $\gamma_{BT,t}$ is the Bracha-Tang inattention measure, such that a higher $\gamma_{BT,t}$ implies more *inattention*.

[47] I do not include time fixed effects in this regression. When I do, I find that if inflation is below the threshold of 3.91%, the estimated attention gain is 0.6 (with a standard error of 0.095), and if inflation is above the threshold, attention is 1.0 (with a s.e. of 0.122). The difference is highly statistically significant with a p-value 0.009.

I allow for multiple thresholds in the data and use the BIC to select the number of thresholds. As in my baseline specification, I find that the data prefer the specification with exactly one threshold.

Table 5 shows the results. The first row shows that the threshold is with 3.39% somewhat lower than 3.91%. The differences in inattention are significantly different from 0 as the last column indicates. As the second row shows, when I impose a threshold at 3.91%, I find very similar results. The estimated inattention levels are similar as for the specification in row 1 and the levels are again significantly different across regimes. If I additionally include lagged inflation as an independent variable (not shown in the Table), I find that the results are barely affected and that the coefficients related to lagged inflation are insignificant in both regimes (with associated p-values 0.45 and 0.7, respectively). This last finding provides further support that attention moves between two regimes and stays constant within regime.

The middle and lower part in Table 5 show that in all regions and across both monthly and quarterly frequency, inattention is higher when inflation is below the threshold, and in most cases, the differences are significant.

Markov-switching model. Instead of estimating regression (4), I also estimate a Markov-switching dynamic regression model in which the long-run inflation beliefs $\underline{\pi}$, the perceived persistence $\tilde{\rho}_\pi$ and attention to inflation γ_π are allowed to differ across unobserved states. That implies that I do not need to take a stance on what defines the regimes, in contrast to equation (4) where I assumed that it is last period's inflation rate that determines the regimes (see Chapter 22 of Hamilton (1994) for a textbook treatment).

The implied attention estimates are 0.20 and 0.32 for the two states and thus, close to the estimates from regression (4) of 0.18 and 0.35. When I allow the variance to change with the state, I estimate attention parameters of 0.2 and 0.34. The implied probability of being in the low-attention regime from the Markov-switching model is 70% which aligns well with the fact that annualized quarter-on-quarter CPI inflation was below the threshold 70% of the time.

Smooth changes in attention. My baseline regression imposes that attention to inflation takes on only two values and jumps from one to the other when inflation exceeds the threshold or falls back below it. Another possibility is that attention changes smoothly with inflation or inflation volatility. To test for this, I estimate the two alternative regressions:

$$\pi^e_{t+3|t} = \beta_0 + \beta_1 \pi^e_{t|t-3} + \beta_2 \left(\pi_t - \pi^e_{t|t-3}\right) + \beta_3 \pi_{t-1} \left(\pi_t - \pi^e_{t|t-3}\right) + \varepsilon_t \tag{38}$$

$$\pi^e_{t+3|t} = \delta_0 + \delta_1 \pi^e_{t|t-3} + \delta_2 \left(\pi_t - \pi^e_{t|t-3}\right) + \delta_3 \left(\pi_{t-1} - \pi_{t-2}\right)^2 \left(\pi_t - \pi^e_{t|t-3}\right) + \varepsilon_t, \tag{39}$$

where β_3 captures that attention may change with the past level of inflation and δ_3 that it may change with inflation volatility, here measured as the squared change in past inflation.

Again, I use the Bayesian Information Criterion to select which model fits the data best and I find that the data prefer the threshold specification (4) over the two specifications (38) and (39).

Table 5: Estimated attention levels and the attention threshold

	Threshold $\hat{\bar{\pi}}$	Low Att. $\widehat{\alpha}_L$	High Att. $\widehat{\alpha}_H$	p-val. $H_0 : \alpha_L = \alpha_H$
Estimated threshold	3.39%	0.11	0.08	0.000
s.e.		(0.0038)	(0.0050)	
Imposed threshold	3.91%	0.10	0.08	0.010
s.e.		(0.0059)	(0.0094)	
Monthly frequency	3.91%	0.10	0.08	0.004
s.e.		(0.0056)	(0.0077)	
Regional analysis				
West	3.91%	0.10	0.08	0.277
s.e.		(0.0068)	(0.0129)	
Northcentral	3.91%	0.10	0.07	0.003
s.e.		(0.00624)	(0.0083)	
Northeast	3.91%	0.10	0.07	0.002
s.e.		(0.0071)	(0.0100)	
South	3.91%	0.10	0.08	0.022
s.e.		(0.00683)	(0.0105)	
Regional, monthly				
West	3.91%	0.10	0.08	0.073
s.e.		(0.0065)	(0.0102)	
Northcentral	3.91%	0.10	0.07	0.000
s.e.		(0.0065)	(0.0070)	
Northeast	3.91%	0.10	0.07	0.015
s.e.		(0.0071)	(0.0096)	
South	3.91%	0.10	0.09	0.171
s.e.		(0.0066)	(0.0104)	

Notes: This table shows the results from regression (37), where $\hat{\bar{\pi}}$ denotes the estimated threshold, $\widehat{\alpha}_L$ and $\widehat{\alpha}_H$ the estimated intercepts when inflation is below or above the threshold, respectively. The last column shows the p-value for the null hypothesis that the two coefficients are equal. Standard errors are robust with respect to heteroskedasticity and serial correlation (Newey and West (1987) with four lags, twelve lags for the monthly specification).

Additionally, even though the estimated β_3 and δ_3 are statistically significantly different from 0, they are close to 0 (0.009 and -0.0002, respectively). In the following, I also test whether a combination of such a 'smooth-attention model' with my threshold model would perform better.

Attention changes within regime. In Table 1, we saw that attention increases when inflation exceeds the threshold of 3.91%. But what about changes within regime? To look at this, I estimate a time series of attention. In particular, I estimate the regression:

$$\pi^e_{t+3|t} = \beta_0 + \beta_1 \pi^e_{t|t-3} + \beta_2 \left(\pi_t - \pi^e_{t|t-3}\right) + \epsilon_t,$$

using a rolling-window approach, where each window has a length of five years. I denote the estimated time series of attention parameters by $\gamma_{\pi,t}$, and I compute the window-specific average of the monthly quarter-on-quarter inflation rate, π^{avg}_t. To then test whether attention within regime is higher when inflation is higher, I estimate the following regression:

$$\gamma_{\pi,t} = \delta_0 + \delta_1 \mathbb{1}_{\pi_t^{avg} \geq 3.91} + \delta_2 \pi_{t-1} + \delta_3 \mathbb{1}_{\pi_t^{avg} \geq 3.91} \pi_{t-1} + \varepsilon_t, \tag{40}$$

where $\mathbb{1}_{\pi_t^{avg} \geq 3.91}$ is an indicator that equals one when in period t average inflation over the last twelve months is above 3.91% and zero otherwise. Thus, δ_1 tells us the difference in attention across regimes, δ_2 the effect of last-period's inflation on attention and δ_3 the additional effect of inflation on attention in the high-inflation regime.

Table 6 shows the results. We see that attention is significantly higher in the high-inflation regime, as indicated by the estimate of δ_1. Yet, inflation does not have any additional significant effect on attention when accounting for the threshold, as depicted by the last two columns.

Table 6: Attention changes within regime

	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$
Regression results	0.08**	-0.002	0.001
s.e.	(0.037)	(0.004)	(0.007)

Notes: This table shows the results from regression (40). Standard errors are robust with respect to heteroskedasticity and serial correlation (Newey and West (1987) with 12 lags). Significance levels: *: p-value < 0.1, **: p-value < 0.05, ***: p-value < 0.01.

An alternative way to test for changes of attention in addition to the jumps across regimes is to estimate the regression:

$$\begin{aligned}\pi^e_{t+1|t} =& \mathbb{1}_{\pi_{t-1} \leq \bar{\pi}} \left(\beta_{0,L} + \beta_{1,L} \pi^e_{t|t-1} + \beta_{2,L} \left(\pi_t - \pi^e_{t|t-1}\right)\right) \\ &+ \left(1 - \mathbb{1}_{\pi_{t-1} \leq \bar{\pi}}\right) \left(\beta_{0,H} + \beta_{1,H} \pi^e_{t|t-1} + \beta_{2,H} \left(\pi_t - \pi^e_{t|t-1}\right)\right) \\ &+ \beta_3 \left(\pi_t - \pi^e_{t|t-1}\right) \cdot \pi_{t-1} + \beta_4 \cdot \pi^e_{t|t-1} \cdot \pi_{t-1} + \tilde{\epsilon}_t,\end{aligned}$$

where β_3 and β_4 capture that the perceived persistence and the attention parameter may depend on last-period's inflation in addition to the changes across regimes. I estimate $\hat{\beta}_3 = 0.0037$ (with a standard error of 0.0033 and a p-value of 0.26), and $\hat{\beta}_4 = -0.0031$ (with a standard error of 0.0057 and a p-value of 0.60). When ignoring the effects coming from β_3 and β_4, the attention parameters are $\hat{\gamma}_{\pi,L} = \frac{\beta_{2,L}}{\beta_{1,L}} = 0.20$ and $\hat{\gamma}_{\pi,H} = \frac{\beta_{2,H}}{\beta_{1,H}} = 0.31$. When accounting for the changes in attention through β_3 and β_4 the implied attention parameters are $\hat{\gamma}_{\pi,L} = \frac{\beta_{2,L} + \hat{\beta}_3 \cdot 2}{\beta_{1,L} + \hat{\beta}_4 \cdot 2} = 0.21$ at $\pi_{t-1} = 2\%$ and $\hat{\gamma}_{\pi,H} = \frac{\beta_{2,H} + \hat{\beta}_3 \cdot 5}{\beta_{1,H} + \hat{\beta}_4 \cdot 5} = 0.34$ at $\pi_{t-1} = 5\%$. Overall, these estimated attention parameters are quantitatively close to the ones from my baseline estimation of 0.18 and 0.35. As furthermore the interaction terms are insignificant, these findings suggest that the threshold model of attention captures the data well.

Long-run expectations. I now turn to long-run expectations. In particular, I consider an exercise that also offers an alternative view and test of my baseline results and the model of short-run expectations. To do so, I rely on the Michigan Survey of Consumers and use the average response to the question "*What about the outlook for prices over the next 5 to 10 years?*" as my measure of long-run expectations. This series is consistently available since 1990. Let me denote this long-run forecast $\pi^e_{LR|t}$. I then run the following threshold regression:

$$\begin{aligned}\pi^e_{LR|t} &= \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\left(\beta_{0,L} + \beta_{1,L}\pi^e_{t|t-3} + \beta_{2,L}\left(\pi_t - \pi^e_{t|t-3}\right)\right) \\ &\quad + \left(1 - \mathbb{1}_{\pi_{t-1}\leq\bar{\pi}}\right)\left(\beta_{0,H} + \beta_{1,H}\pi^e_{t|t-3} + \beta_{2,H}\left(\pi_t - \pi^e_{t|t-3}\right)\right) + \epsilon_t,\end{aligned} \tag{41}$$

which is the exact same regression as my baseline regression (4) except that the dependent variable is now the long-run expectation.

Note that the way I solve the New Keynesian model in Section 4 is that I assume that agents do not foresee regime changes, and instead act as if they would remain in the current regime forever. Thus, their long-run forecasts (in a given regime) are given by:

$$\pi^e_{t+h|t} = \underline{\pi}(1 - \tilde{\rho}^h_\pi) + \tilde{\rho}^h_\pi\left(\pi^e_{t|t-1} + \gamma_\pi(\pi_t - \pi^e_{t|t-1})\right) + \eta_t. \tag{42}$$

Therefore, we have $\beta_0 = \underline{\pi}(1 - \tilde{\rho}^h_\pi)$, $\beta_1 = \tilde{\rho}^h_\pi$, and $\beta_2 = \tilde{\rho}^h_\pi\gamma_\pi$. Hence, by computing $\frac{\beta_2}{\beta_1}$, I should obtain similar attention estimates as in my baseline regression, and by extension, also a similar threshold.

Table 7 shows the results. The results are indeed quite similar to my baseline results, consistent with the theory. The estimated threshold is at 4.66%, somewhat higher than my baseline (but close to the estimate when using median expectations, see Table 4 above). The estimated attention parameter in times inflation is below the threshold is 0.18, and it increases to 0.33 when inflation exceeds the threshold.[48] These estimates are very close to my baseline estimates of 0.18 and 0.35.

Table 7: Estimated attention levels and the attention threshold for long-run expectations

	Threshold $\hat{\bar{\pi}}$	Low Att. $\hat{\gamma}_{\pi,L}$	High Att. $\hat{\gamma}_{\pi,H}$
Mean expectations	4.66%	0.18	0.33
s.e.		(0.039)	(0.267)

Notes: This table shows the results from regression (41), where $\hat{\bar{\pi}}$ denotes the estimated threshold, $\hat{\gamma}_{\pi,L}$ and $\hat{\gamma}_{\pi,H}$ the estimated attention levels when inflation is below or above the threshold, respectively. Standard errors are robust with respect to heteroskedasticity.

Long- and short-run expectations are correlated (see, e.g., Aruoba (2020); Fisher et al. (2025)), and hence, long-run expectations could also affect my attention estimates that are based solely on short-run expectations. To control for this, I re-run the threshold regression (4) while additionally controlling for long-run expectations. Table 8 shows the results. The estimated attention threshold is completely unaffected. The estimated attention parameter in the low-attention regime increases slightly from 0.18 to 0.19, and in the high-attention regime it decreases from 0.35 to 0.3. The coefficients on long-run expectations are 0.08 in the low-attention regime (with a s.e. of 0.034, *p*-

[48] The standard errors for $\hat{\gamma}_{\pi,H}$ are quite large because the sample here is restricted to after 1990, and hence, there are not that many observations in which inflation was above the threshold.

value of 0.02), and 0.06 in the high-attention regime (s.e. 0.06, p-value 0.337). Hence, the coefficient on long-run expectations in the low-attention regime is statistically significant. The magnitude, however, is small. A 1 percentage point higher long-run forecast means that short-run expectations are on average 0.08 percentage points higher (in the low-attention regime).

Table 8: Estimated attention levels and the attention threshold for long-run expectations

	Threshold $\hat{\bar{\pi}}$	Low Att. $\hat{\gamma}_{\pi,L}$	High Att. $\hat{\gamma}_{\pi,H}$
Mean expectations	3.91%	0.19	0.30
s.e.		(0.016)	(0.036)

Notes: This table shows the results from regression (4), while controlling for long-run expectations. Standard errors are robust with respect to heteroskedasticity.

Gasoline prices. Gasoline prices are a key predictor of inflation expectations (Coibion and Gorodnichenko, 2015b). To test whether accounting for gasoline price expectations affects my baseline results, I re-estimate my baseline threshold regression (4) while additionally controlling for the 1-year-ahead gasoline price expectations from the Michigan Survey of Consumers, lagged by one month. This series is available from 1982-1992, and then from 2005 until the end of my sample. The estimated attention parameters are $\hat{\gamma}_{\pi,L} = 0.19$ and $\hat{\gamma}_{\pi,H} = 0.29$ (compared to 0.18 and 0.35 in my baseline estimation). The gasoline-price coefficients are positive in both regimes, but only the one in the low-attention regime is statistically significantly different from 0.[49] The results are very similar when using actual gasoline prices instead of expectations. For this, I use the West Texas Intermediate (WTI) Crude oil price. The implied attention gains are 0.20 and 0.30, and the oil price coefficient is only significant in the low-attention regime. The magnitude of the oil price coefficient is somewhat higher than when considering gasoline-price expectations.

One-year-ahead expectations. As my baseline in Section 2, I focus on one-quarter-ahead expectations and compare them to quarter-on-quarter inflation. When using one-year-ahead expectations and comparing them to year-on-year inflation instead (and using last month's y-o-y inflation), I estimate an attention threshold at 5.03% and that attention increases from 0.54 to 1.22 from the low-attention to the high-attention regime. The difference between the two is highly statistically significant with an associated p-value of 0.000.

Professional forecasters. Professional forecasters behave quite differently from households and firms. To show this, I run my baseline regression using average 3-month-ahead CPI inflation expectations from the Survey of Professional Forecasters from the Philadelphia Fed. According to the Bayesian Information Criterion, the data prefer the specification with only one regime. The

[49] The gas-price-expectations coefficient in times of low attention is 0.015 (with s.e. of 0.007) and in times of high attention, it is 0.008 (with s.e. 0.006). The standard deviation of gas-price expectations over the whole sample is 9.4%. Hence, a coefficient of 0.01 means that a one standard deviation increase in gas price expectations increases inflation expectations by 9.4 basis points.

estimated attention parameter is 0.13. This very low updating gain is close to the one estimated in Coibion and Gorodnichenko (2012) for professional forecasters, and consistent with evidence of strongly anchored and stable inflation expectations of professional forecasters (see, e.g., Reis (2022a,b); Coibion and Gorodnichenko (2025)). Through the lens of the model, a likely explanation of these findings is that professional forecasters' prior uncertainty is very low. However, there is also substantive evidence that professional forecasters behave strategically (Gemmi and Valchev, 2023). Such strategic behavior is not captured in my model.

If I enforce a threshold, I estimate this threshold to be at an inflation rate of 3.92%, which is practically identical to the one of 3.91% that I estimate for households. The estimated attention parameters are $\hat{\gamma}_{\pi,L} = 0.07$ and $\hat{\gamma}_{\pi,H} = 0.17$. Thus, professional forecasters also update their short-run inflation expectations more strongly when inflation is above 4%. The differences across regimes and the degrees of updating overall, however, are smaller than for households.

C Additional Results and Robustness to Section 3

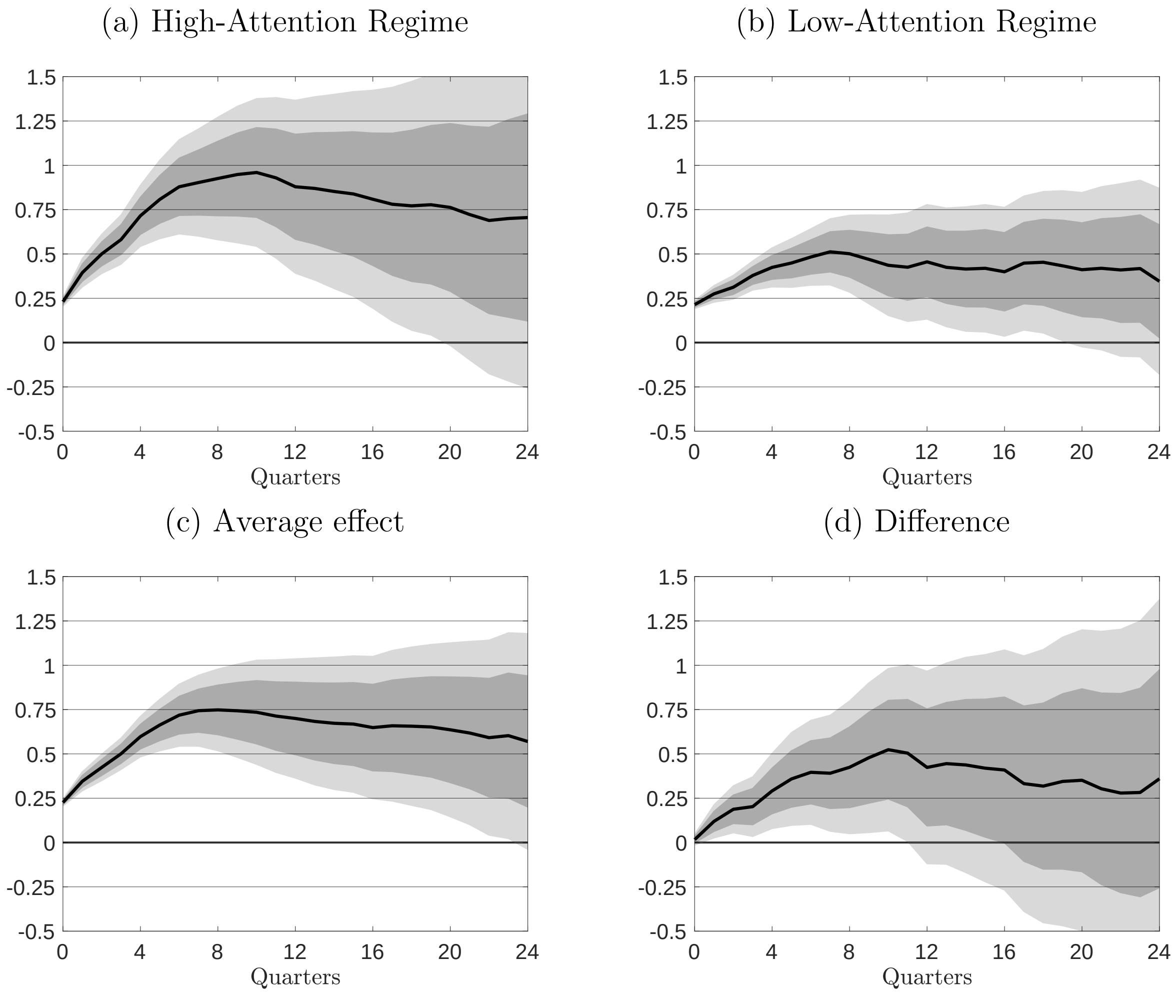
C.1 Other shocks

Inflation shock. Angeletos et al. (2020a) estimate a VAR including 10 key macroeconomic variables and then identify different shocks by maximizing their contribution to the volatility of a given variable over business-cycle frequency (6-32 quarters). I use their shock that contributes most to the volatility of inflation (using the GDP deflator, as in Angeletos et al. (2020a)). These shocks are available at quarterly frequency and span the period 1960 until the end of 2017. I use the previous quarter's CPI inflation as an indicator of whether the economy is in the high-attention regime or the low-attention regime (i.e., was the previous annualized CPI inflation rate above or below 4%.)

The dependent variable is the change in the log of the GDP deflator (times 100) from quarter $t-1$ to quarter $t+h$. Figure 12 shows the results. As in the specification with oil news shocks, inflation responds about twice as much to the *inflation shock* when hit in the high-attention regime compared to the low-attention regime. As panel (d) shows, these differences are highly statistically significant and quite persistent: the difference is largest 10 quarters after the shock and still highly statistically significant.

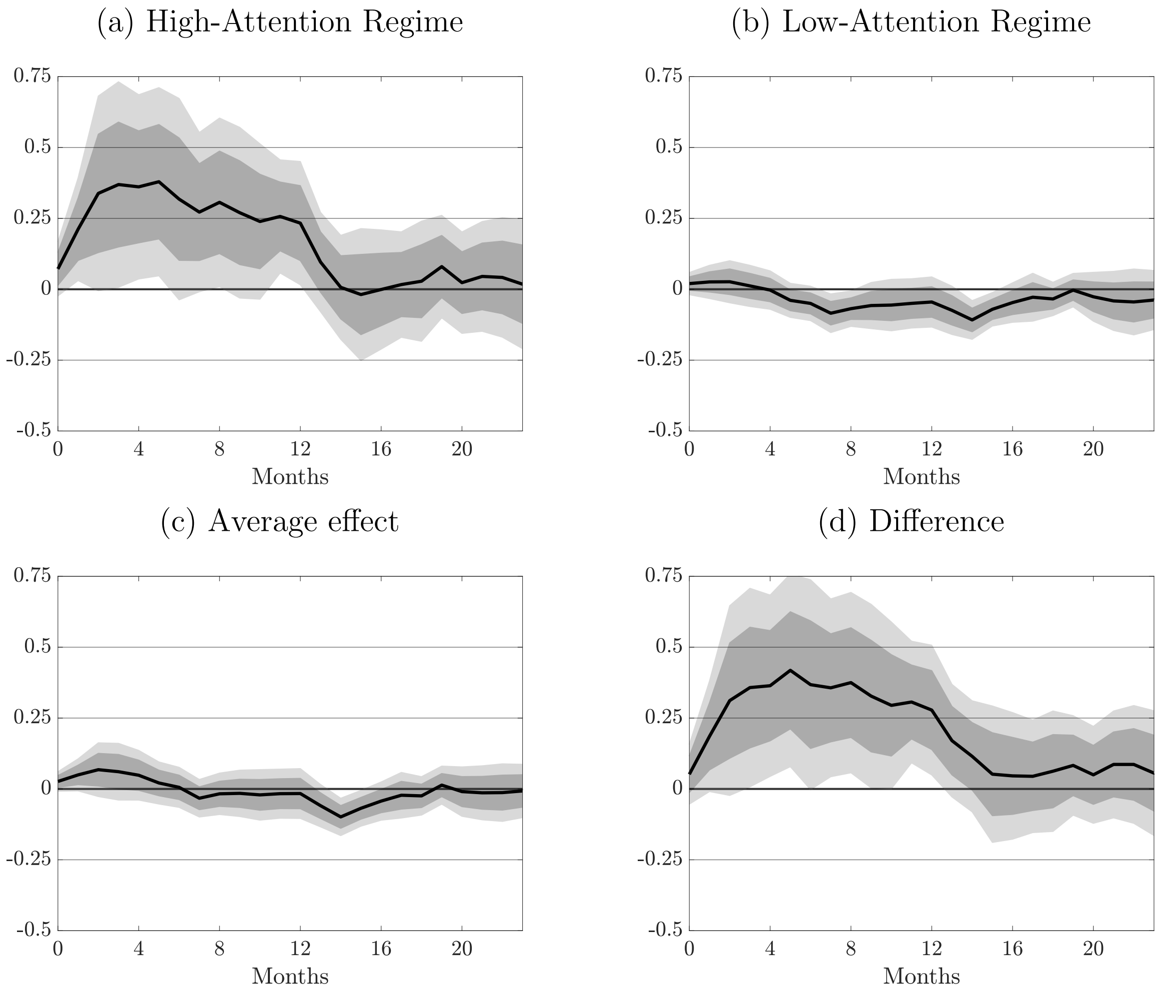
Monetary policy shocks. In Figure 13, I show that inflation also responds more strongly in the high-attention regime to monetary policy shocks identified using a high-frequency identification (the shocks are taken from Jarociński and Karadi (2020) and are purged from the information effects of monetary policy statements; I show here the shocks identified using sign restrictions, but the results are practically identical when using the shock series based on the "poor-man approach"). A drawback of using these monetary policy shocks, however, is that they are only available for the period 1990-2019. In particular, there are 49 months out of 353 in which quarter-on-quarter CPI inflation was above 4% during that period, so about 14% of the time, whereas in the full sample the economy spends about 30% of the time in the high-attention regime. Thus, the sample does not include most of the high-inflation periods. Therefore, while the differences across regimes are substantial in magnitude, the differences are less statistically significant.

Figure 12: Price level response to the shock targeting inflation



Notes: This figure shows the cumulative price level (using the GDP deflator) response to the shock of Angeletos et al. (2020a) that targets inflation in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 4 lags). The attention regimes are defined based on the previous quarter's CPI inflation rate.

Figure 13: Inflation response to a monetary policy shock

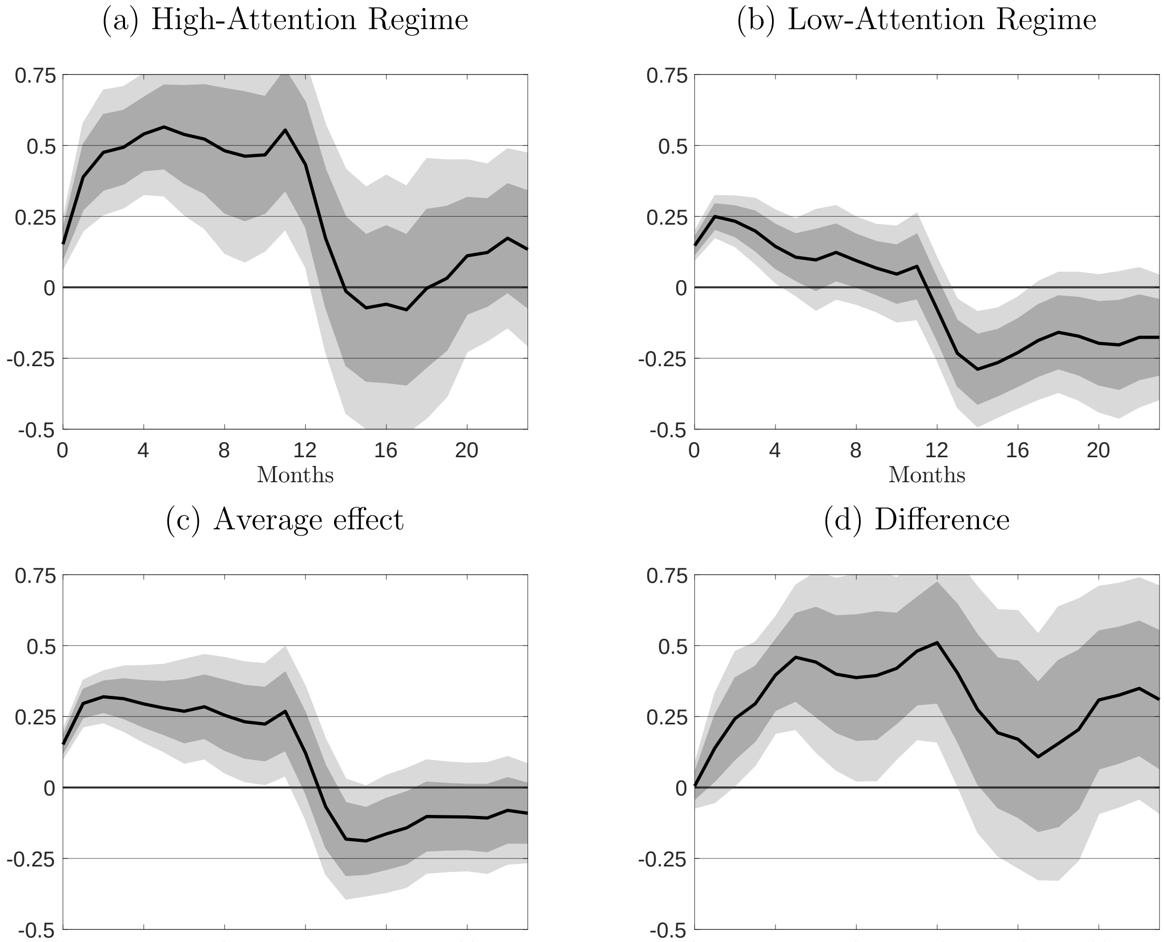


Notes: This figure shows the inflation response to a monetary policy shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags). The shocks are from Jarociński and Karadi (2020) and are based on a high-frequency identification and cleaned from information effects of monetary policy.

C.2 Robustness analysis

Using Google Trends as the threshold-defining variable. Instead of using the lagged inflation rate as the threshold-defining variable, I now use Google Trends data to define the regimes. As this data is only available since 2004, the high-attention regime largely coincides with the recent inflation surge. I use the same control variables as in my baseline specification and I further include four lags of the number of Google searches as control variables. Figure 14 shows the results. They confirm the baseline results that inflation responds substantially and significantly more to supply shocks when attention to inflation is high.

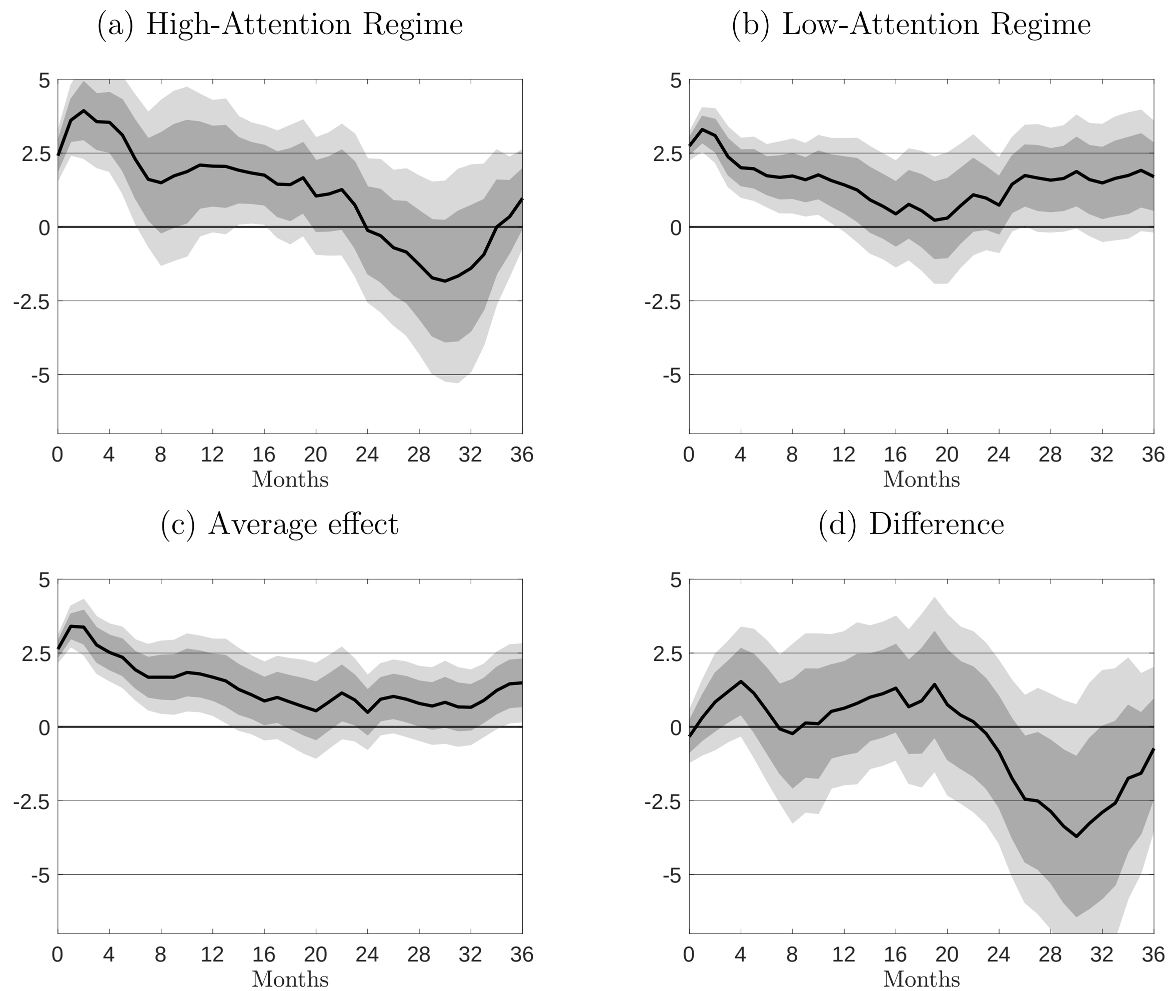
Figure 14: Inflation response to an oil supply shock with Google data as threshold-defining variable



Notes: This figure shows the inflation response to an oil supply news shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The attention regimes are defined based on the number of Google searches of the word inflation in the current month. The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

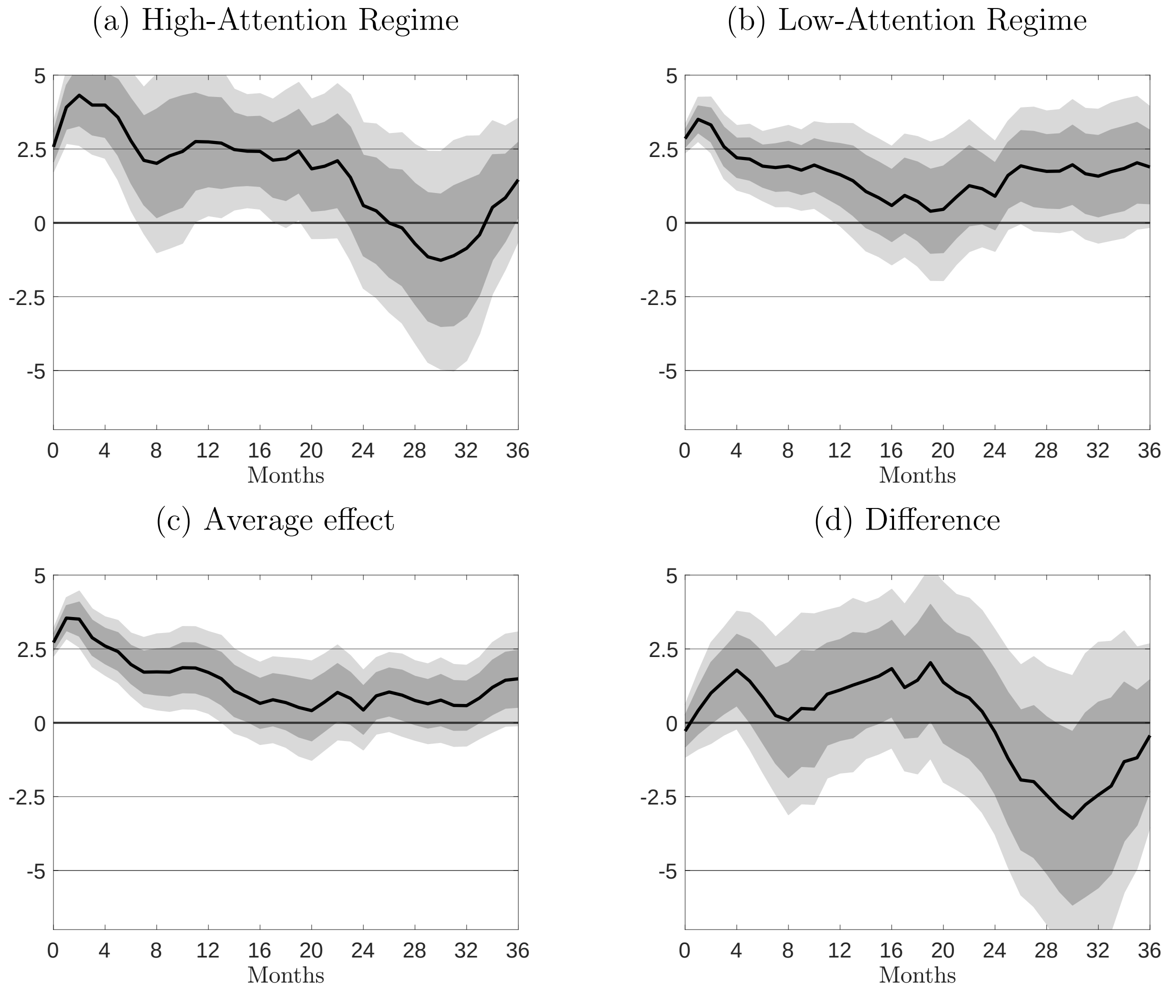
Oil price response. Figures 15 and 16 show the responses of the real and nominal oil price, respectively, to a negative oil supply news shock for the two attention regimes (panels (a) and (b)), the average response in panel (c), and the difference across regimes in (d). We see that the differences across regimes are not significant at the 10% significance level at any horizon, indicating that the differences in the inflation responses discussed in section 3 are unlikely to be driven by different oil price responses.

Figure 15: Real oil price response to an oil supply news shock



Notes: This figure shows the real oil price response to an oil supply news shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags). The attention regimes are defined based on the previous month's inflation rate.

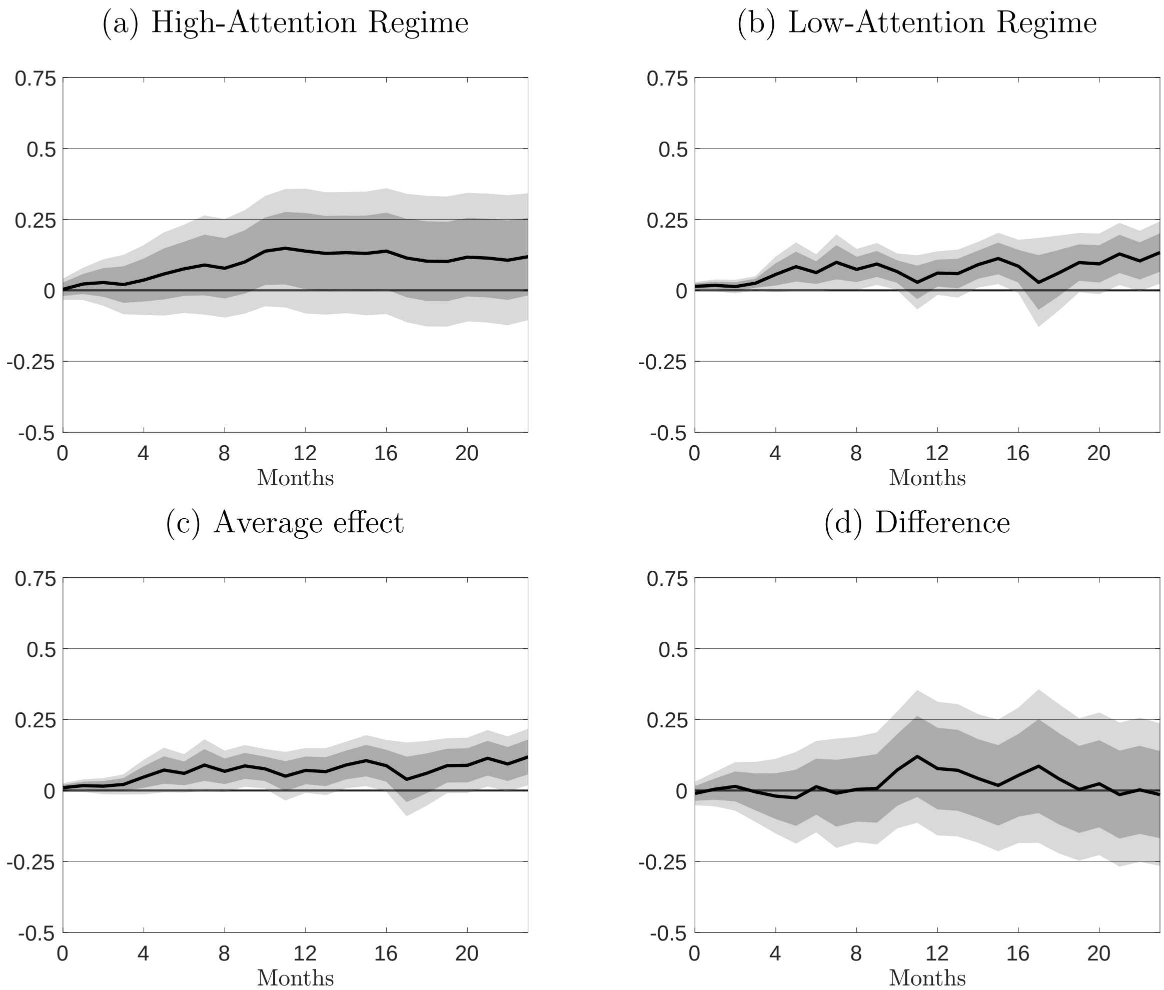
Figure 16: Nominal oil price response to an oil supply news shock



Notes: This figure shows the nominal oil price response to an oil supply news shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags). The attention regimes are defined based on the previous month's inflation rate.

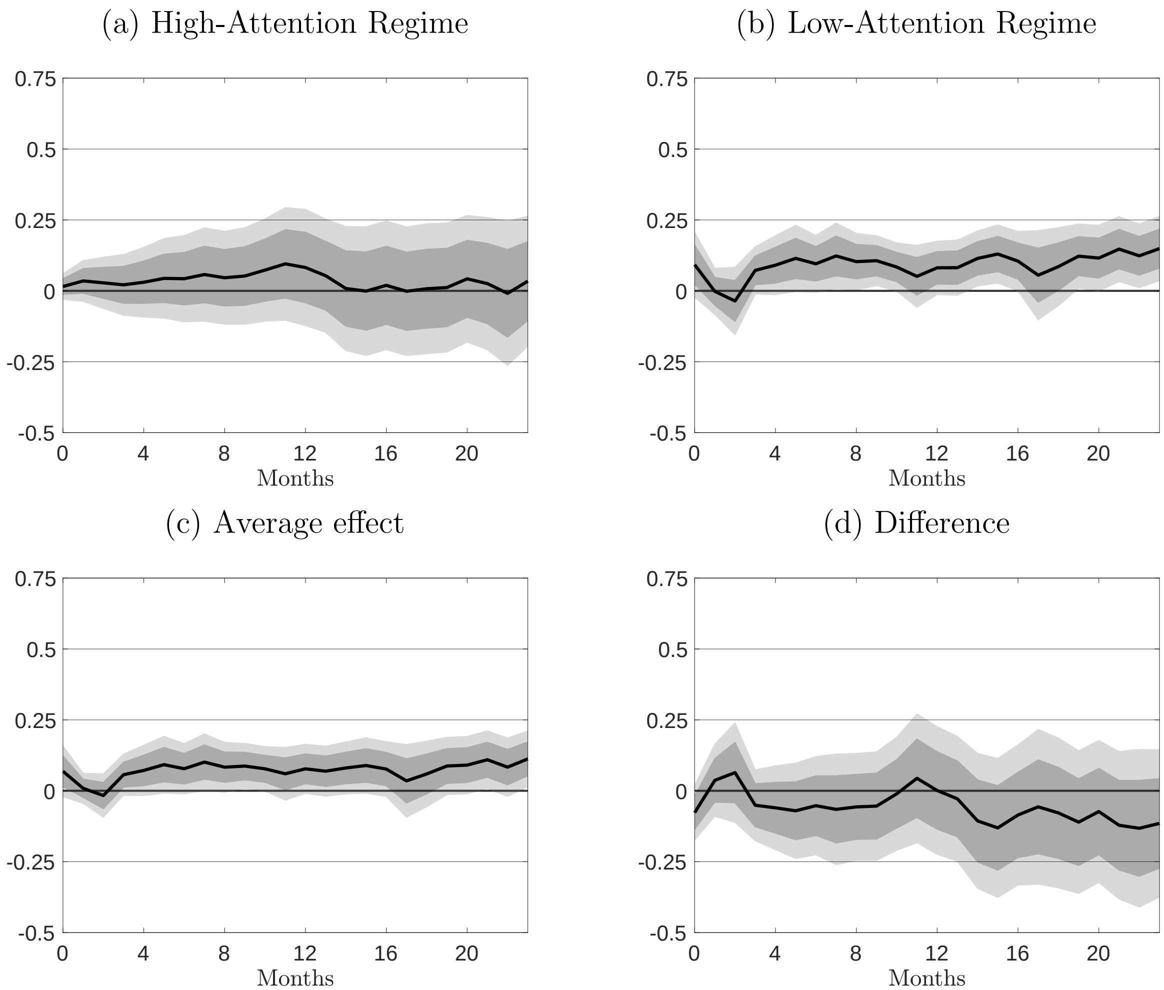
Unemployment response. Figures 17 and 18 show the responses of the unemployment rate (excluding the Covid period 2020-2023 and including it, respectively) to a negative oil supply news shock for the two attention regimes (panels (a) and (b)), the average response in panel (c), and the difference across regimes in (d). We see that the differences across regimes are not significant at any conventional significance level at any horizon, indicating that the differences in the inflation responses discussed in section 3 are unlikely to be driven by generally more responsive macroeconomic variables.

Figure 17: Unemployment response to an oil supply news shock



Notes: This figure shows the unemployment rate response to an oil supply news shock when excluding the Covid-19 period (2020-2023) in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

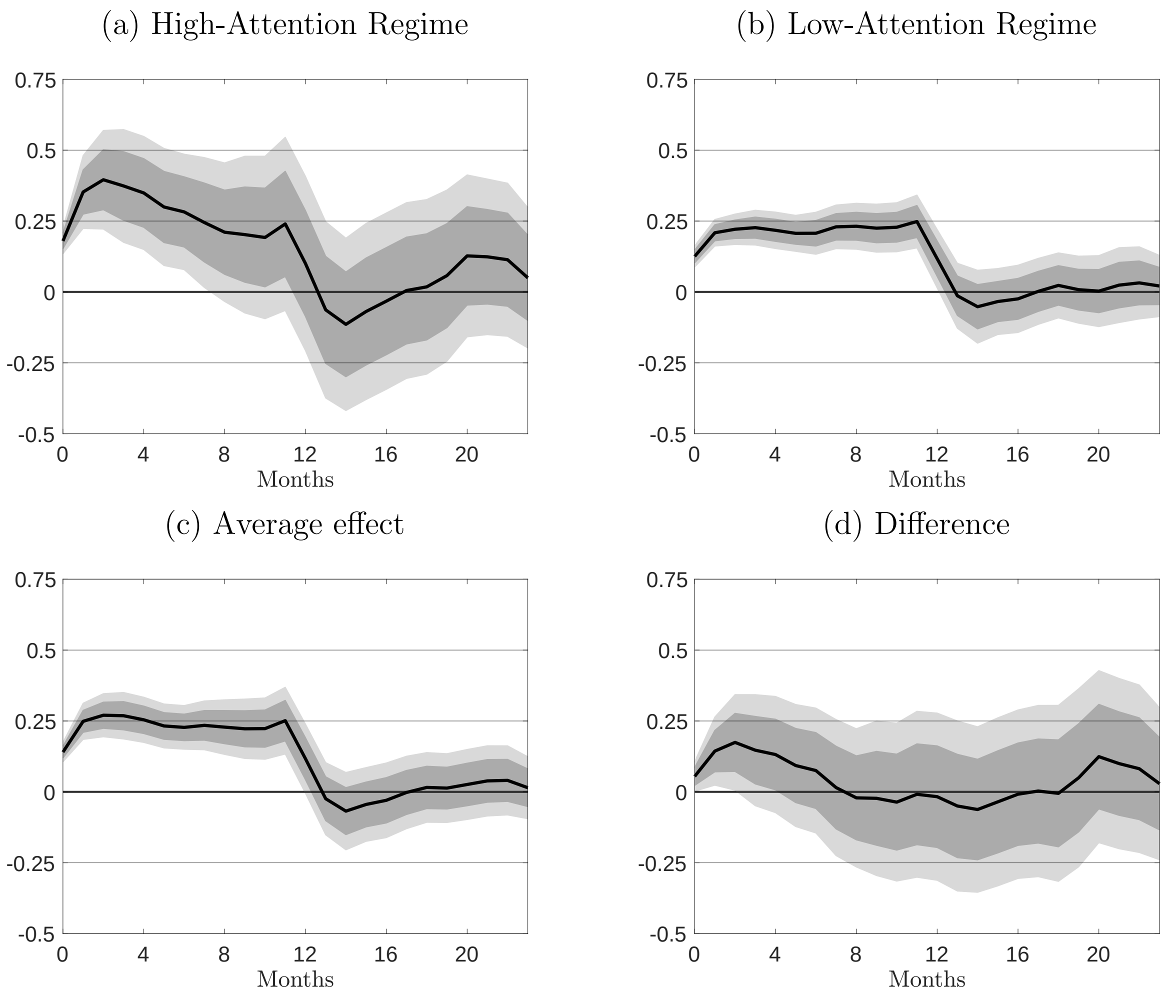
Figure 18: Unemployment response to an oil supply news shock including Covid



Notes: This figure shows the unemployment rate response to an oil supply news shock when including the Covid-19 period (2020-2023) in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Covid. Figure 19 shows the inflation response to a negative oil supply news shock for the two attention regimes (panels (a) and (b)), the average response in panel (c), and the difference across regimes in (d) when excluding the Covid period 2020-2023. We see that the differences across regimes are slightly weaker than for the baseline case in Figure 3 but still substantial.

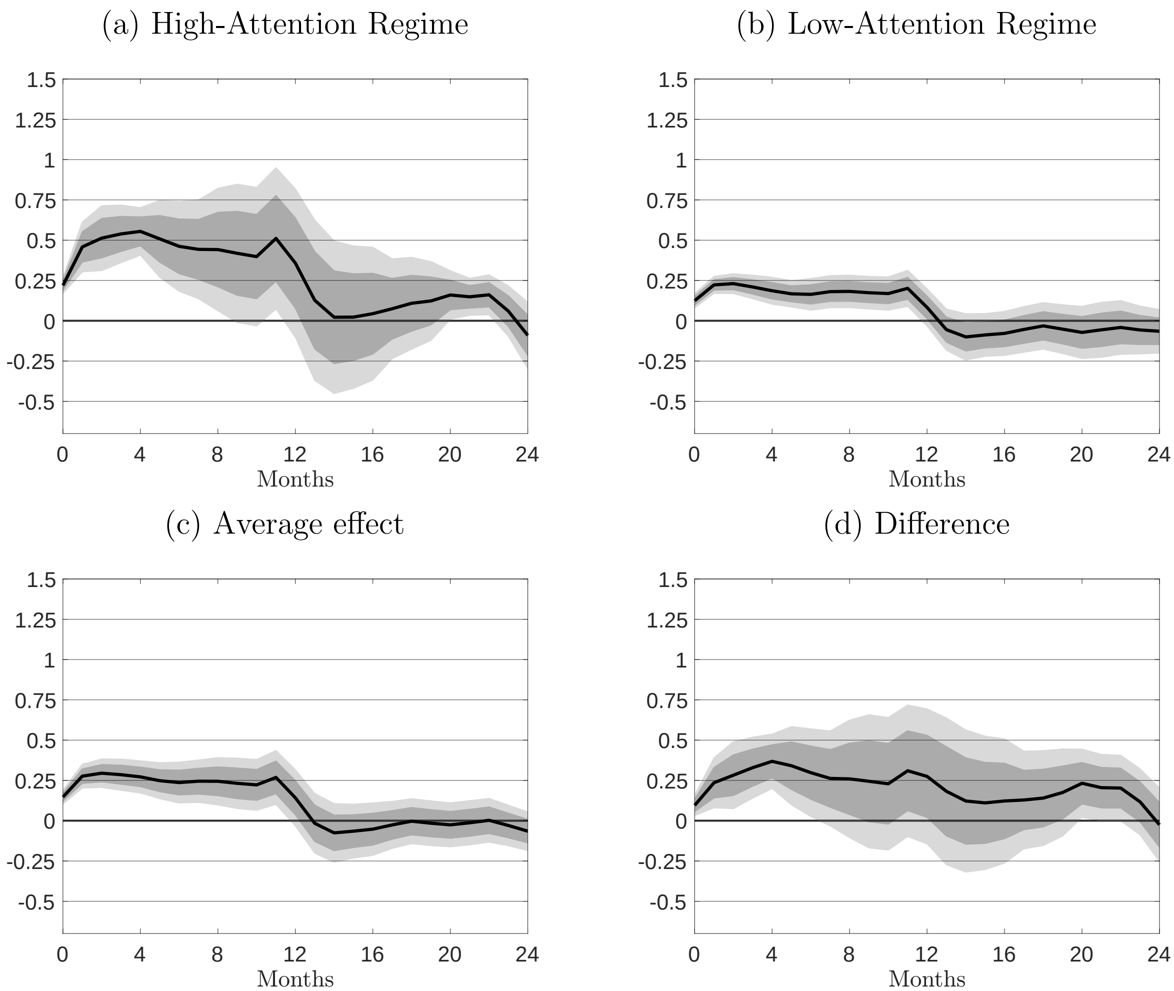
Figure 19: Inflation response to an oil supply news shock excluding Covid



Notes: This figure shows the inflation response to an oil supply news shock when excluding the years 2020, 2021, 2022 and 2023 in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Excluding the Great Inflation period. Figure 20 shows the inflation response to a negative oil supply news shock for the two attention regimes (panels (a) and (b)), the average response in panel (c), and the difference across regimes in (d) when excluding the Great inflation period, i.e., when excluding observations before 1990. We see that the differences across regimes are even slightly stronger than in the baseline specification.

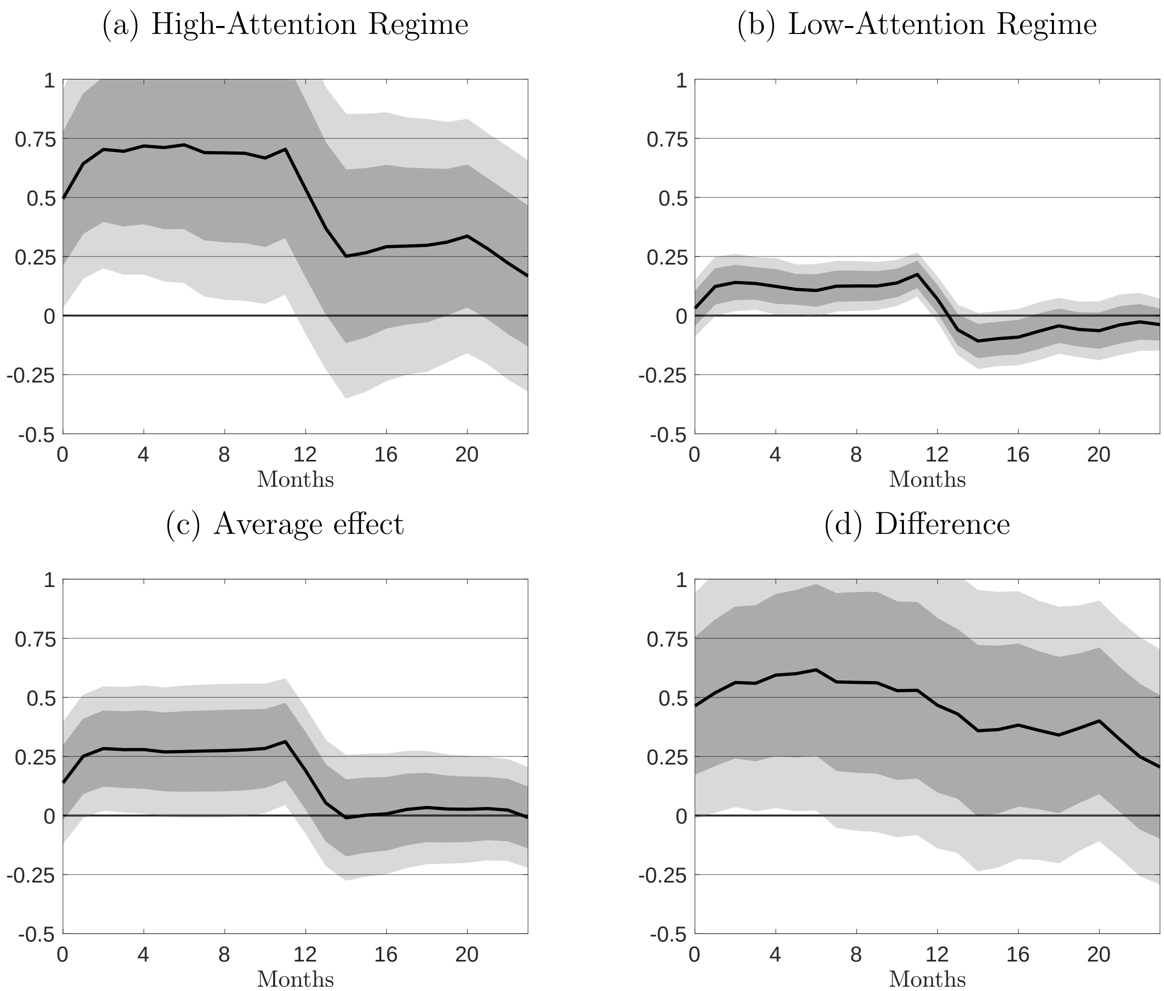
Figure 20: Inflation response to an oil supply news shock after 1990



Notes: This figure shows the inflation response to an oil supply news shock when excluding the years prior to 1990, in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Role of control variables. Figure 21 shows the inflation response to a negative oil supply news shock for the two attention regimes (panels (a) and (b)), the average response in panel (c), and the difference across regimes in (d) when not including any control variables. We see that the differences are still substantial, but the estimates, especially in the high attention regime, are noisier.

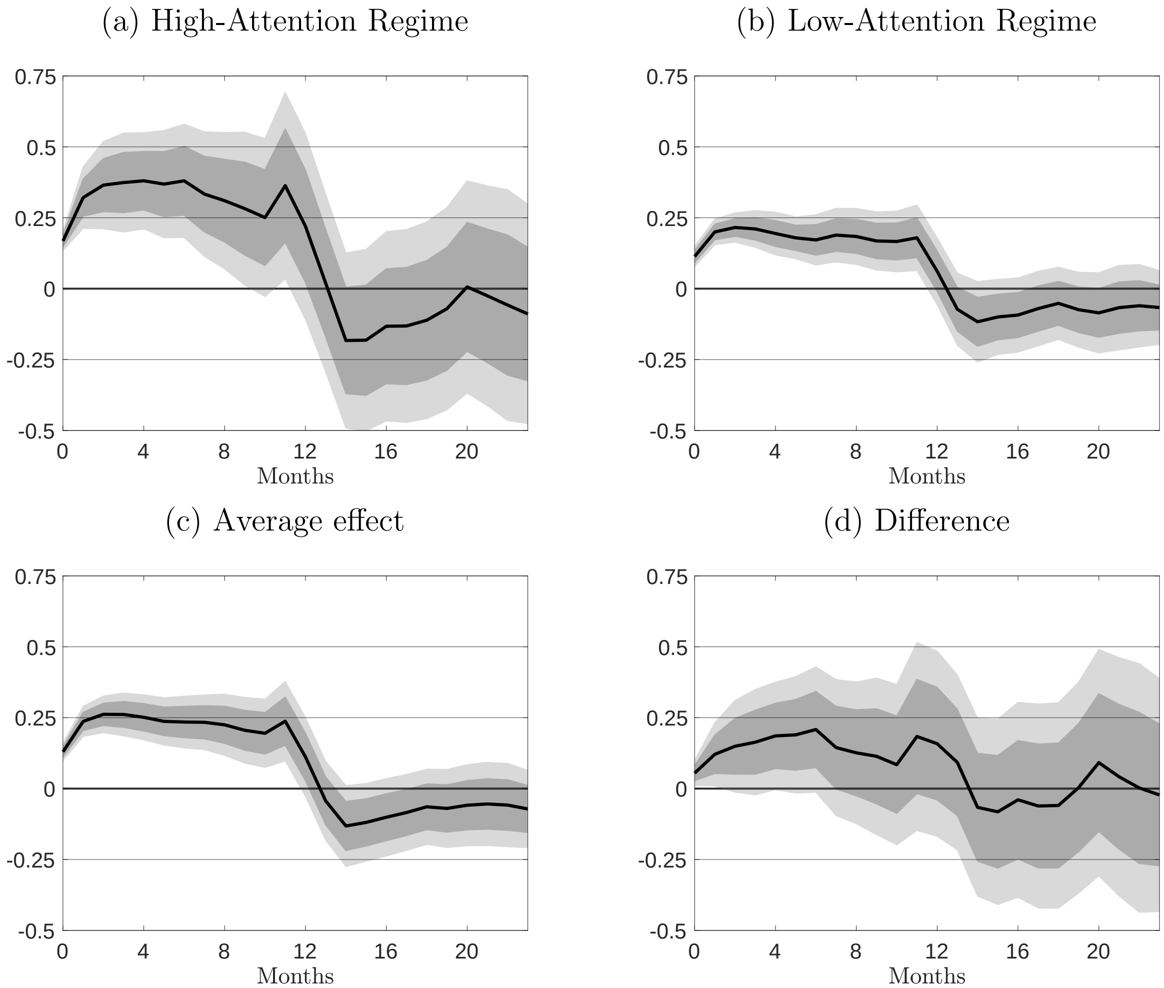
Figure 21: Inflation response to an oil supply news shock (no controls)



Notes: This figure shows the inflation response to an oil supply news shock when using no control variables in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Figure 22 shows the effects when controlling for four lags of the shock, inflation, and the unemployment as well as the interaction of those variables with a regime-dummy (for all four lags). The standard errors for the high-attention regime estimates increase. But the difference across regimes still remains significantly different from 0 at the 10% in the first two months, and the overall results align with the results in Figure 3.

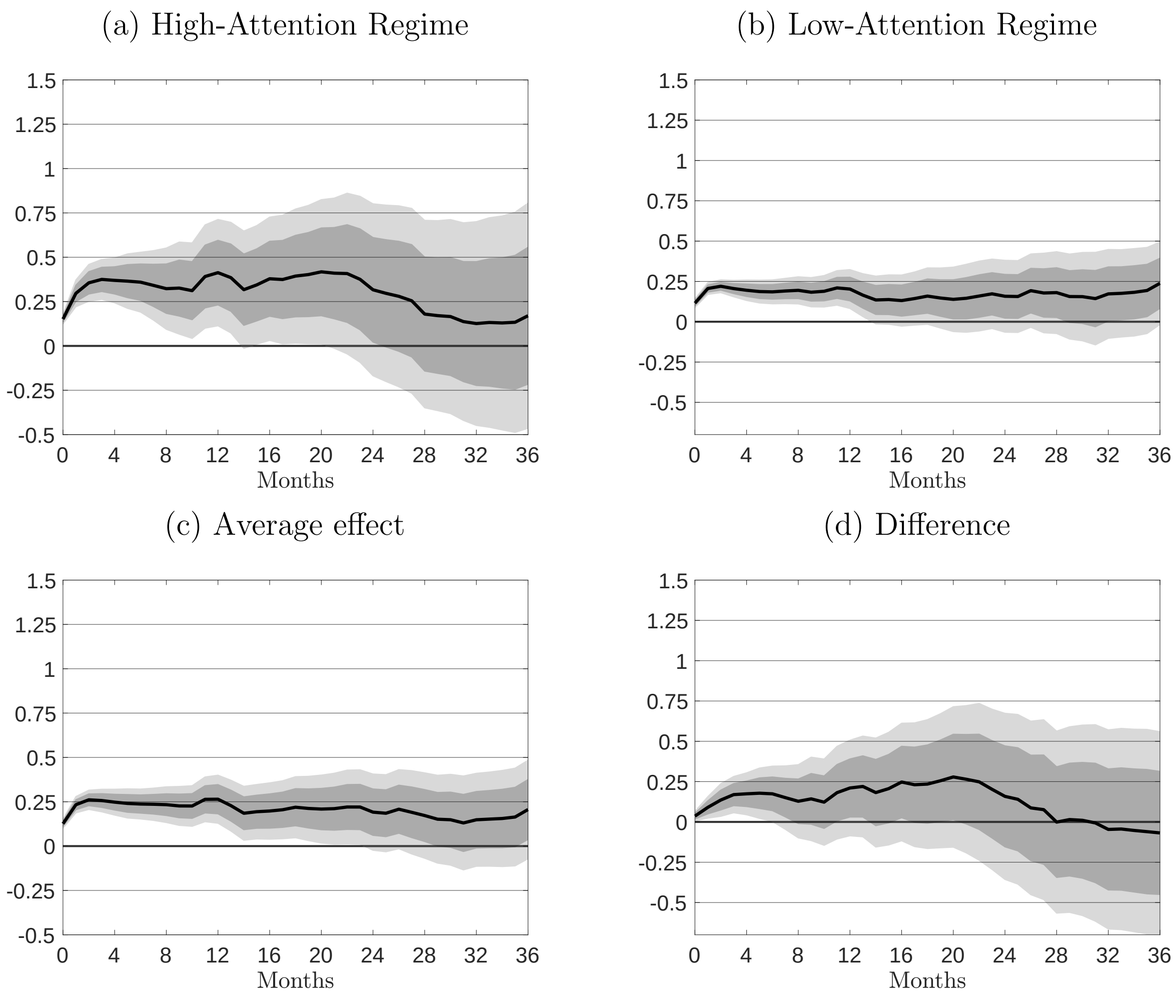
Figure 22: Controlling for regime-dependent macroeconomic variables



Notes: This figure shows the inflation response to an oil supply news shock (equation (5)) in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The controls include four lags of the shock, inflation, and the unemployment as well as the interaction of those variables with a regime-dummy (for all four lags). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

Price-level response. Figure 23 shows the case where the dependent variable is the cumulative change in the price level: $P_{t+h} - P_{t-1}$, where P_t is the natural logarithm (times 100) of the price level using the CPI. We see from Figure 23 that the results are similar to the baseline case in Figure 3: inflation responds almost twice as strongly in the high-attention regime and the differences are highly statistically significant in the first few months.

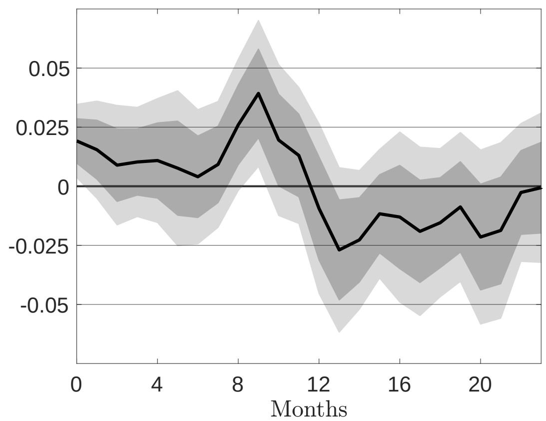
Figure 23: Price level response to an oil supply news shock



Notes: This figure shows the cumulative price level (using the CPI) response to an oil supply news shock in the high-attention regime (panel (a)), the low-attention regime (panel (b)), on average across regimes (panel (c)), and the difference between the two regimes (panel (d)). The dark shaded areas depict the 68% confidence bands and the light-shaded area the 90% confidence bands. Standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags). The attention regimes are defined based on the previous month's inflation rate.

Alternative attention measures. Figure 24 shows the interaction effect of the inattention measure based on Bracha and Tang (2025) and the oil supply news shock. The response is normalized to show the effect of a one standard-deviation shock with the inattention measure being one standard deviation above its mean. The coefficients have been normalized such that a higher value indicates a more inflationary effect in times of higher attention. I use regional data to construct the measures and I control for time- and regional fixed effects in the local projections, as well as four lags of the inattention measure and four lags of regional inflation. If inattention is higher, oil supply news shocks are less inflationary and these effects are often significant in the first 10 months. These findings thus support the findings in Section 3, using a different measure of (in)attention to inflation.

Figure 24: Alternative measure of inattention and inflation response

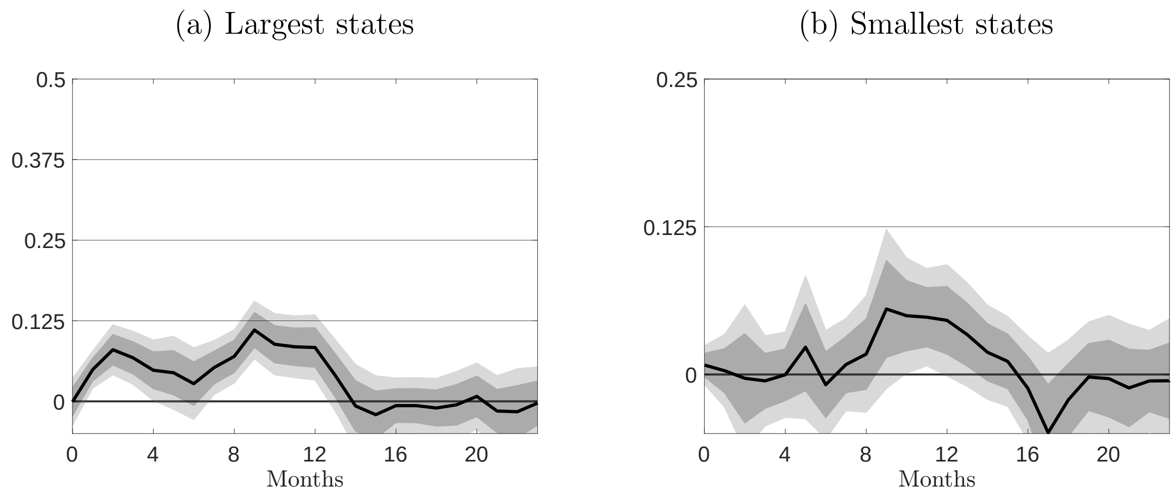


Notes: This figure shows the inflation response to an oil supply news shock. In particular, it shows the interaction term of the Bracha-Tang measure of inattention with the shock (for a one standard deviation shock and the Bracha-Tang measure being one standard deviation above its mean) when controlling for time and region fixed effects, as well as four lags of the inattention measure and four lags of regional inflation. The coefficients have been normalized such that a higher value indicates a more inflationary effect in times of higher attention. Standard errors are robust with respect to heteroskedasticity as well as serial and cross-sectional correlation.

Alternative regional measures of Google Trends. For my baseline measure of Google Trends at the regional level, I average across all states within the different regions. Within every state, Google Trends are normalized such that the highest value is set to 100. A value of 100 in Texas, however, might not be the same as 100 in Florida. I therefore consider two alternatives. In the first one, for every census region, I use the region's largest state's Google searches as the measure of attention of that region. Second, I use the smallest state's Google searches. Both according to their 2024 population.[50] Thus, the Google Trends data used in these exercises does not feature any within region cross-sectional differences. Differences across regions are absorbed via the region fixed effects. Figure 25 shows the results. The results are similar to my baseline estimates, even though the ones using the smallest states (panel (b)) are somewhat noisier.

[50] The largest states are New York, Texas, California and Illinois. They account for about 30% of the US population, and about 35% of U.S. GDP according to the BEA in 2023. The smallest states are Vermont, Delaware, North Dakota, and Wyoming.

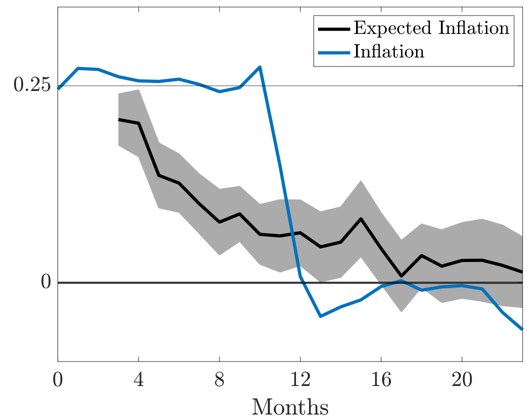
Figure 25: Interaction with Google Trends using the Largest and Smallest States Only



Notes: This figure shows the inflation response to an oil supply news shock. In particular, it shows the interaction term of Google Trends with the shock (for a one standard deviation shock and Google Trends being one standard deviation above its mean) when controlling for time and region fixed effects, as well as four lags of Google Trends and four lags of regional inflation. Instead of using an aggregate measure of Google Trends for every region, I here use Google Trends of the largest (panel (a)) and smallest (panel (b)) state (in terms of population in 2024) within every census region. Standard errors are robust with respect to heteroskedasticity as well as serial and cross-sectional correlation.

Expectations and inflation. Figure 26 shows the response of inflation forecasts to an adverse oil supply news shock using average expectations from the Survey of Consumers (black-solid line) together with the response of actual inflation (blue-solid line). The response of expectations is shifted such that the vertical distance between the two lines captures forecast errors. Consistent with the model, we see that expectations initially undershoot, followed by a delayed overreaction.

Figure 26: Inflation and expected inflation



Notes: This figure shows the response of inflation forecasts using average expectations from the Survey of Consumers (black-solid line) together with the response of actual inflation (blue-solid line). The response of expectations is shifted such that the vertical distance between the two lines captures forecast errors. The shaded areas depict the 68% confidence bands, standard errors are robust with respect to serial correlation and heteroskedasticity (Newey and West (1987) with 12 lags).

C.3 The Role of Attention in the Post-Covid Inflation Surge

In this section, I show that the increase in people's attention to inflation played a quantitatively important role in the post-Covid inflation surge in the United States.[51] The black-solid line in Figure 27 shows the evolution of CPI inflation for the period 2020-2023.

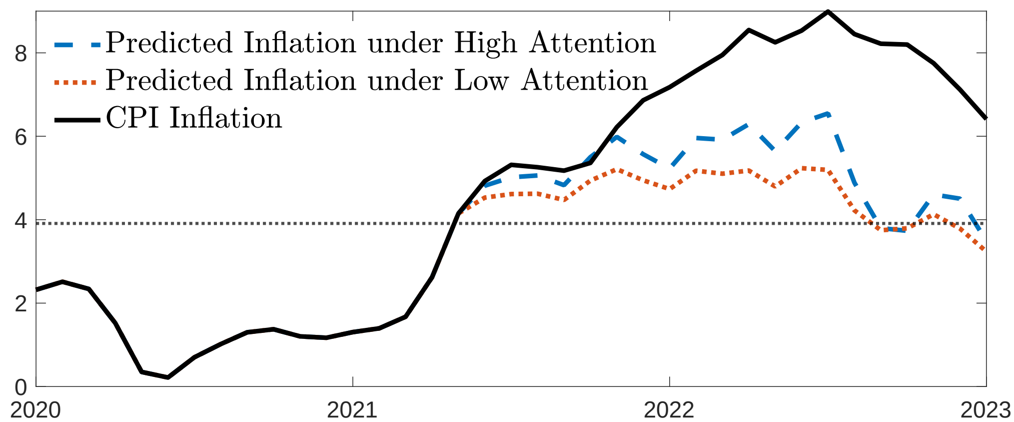
To quantify the importance of the change in attention and of oil supply shocks, I estimate the counterfactual evolution of CPI inflation when only considering oil supply news shocks starting in April 2021, which is when inflation exceeded 3.91% and thus, the economy entered the high-attention regime.[52] I do this twice: once, taking the increase in attention into account, and once, without the increase in attention. For the case with the increase in attention, I therefore feed in the identified oil supply shocks from April 2021 until the end of 2022 using the estimated impulse response functions

[51] A similar approach is, e.g., used in Coibion (2012) who analyzes the historical contribution of monetary policy shocks to industrial production, unemployment and inflation, or by Mitra (2023) who examines the role of noise shocks for the dynamics of labor markets.

[52] I ignore all previous shocks, i.e., I compute the counterfactuals as if the economy was in steady state in April 2021 and that there were no shocks before that period. Thus, the reported results in this section are likely to be at the lower end of the actual importance of attention for the inflation dynamics, given that inflation was already increasing before and that the supply shocks in the first three months of 2021 were all negative (i.e., inflationary). Consistent with early 2021 being the time attention increased, Hilscher et al. (2025) document a rise in the probability of persistent high inflation for the U.S. in mid-2021. Reis (2022a) also documents that household inflation expectations in mid-2021 were not very well anchored any more.

from panel (a) of Figure 3. The blue-dashed line shows the implied inflation rate. We see that the shape of the implied inflation rate is similar to the actual inflation rate, and they both peak in mid-2022. Furthermore, even though I only consider oil supply shocks and abstract from all other potential exogenous drivers of inflation, the implied inflation rate increases from the initial value of 4.1% up to almost 6.55%. Thus, the oil supply shocks account for about 50% of the additional increase from early 2021 until mid-2022. This is consistent with the findings in Shapiro (2024) who estimates the supply-driven part of (PCE) inflation in 2022 to be of similar magnitude.[53]

Figure 27: Supply shocks, attention and the post-Covid inflation surge



Notes: This figure shows the actual U.S. CPI inflation from 2020 until 2023, as well as the counterfactual inflation dynamics arising only from oil supply news shocks in the economy accounting for the increase in attention in April 2021 (blue-dashed line) and in the economy ignoring the increase in attention (red-dotted line).

How much of this is due to the increase in attention? The red-dotted line shows that if there had not been an increase in people's attention, that is when relying on the estimated impulse response functions from the low-attention regime (panel (b) in Figure 3), the implied inflation increase from April 2021 until mid-2022 would have only been about 1.1 percentage points (compared to the 2.55pp. when accounting for people's change in attention). These results suggest that the rise in the public's attention to inflation was quantitatively important in driving inflation further up in response to supply shocks during the recent inflation surge, and that without the increase in people's attention, supply-driven inflation would have been less than half as high as it was.

D Model Derivations and Calibration of γ_x

Derivation of equation (22). Linearizing the Euler equation (8) yields:

$$c_t = c^e_{t+1|t} - \left(i_t - \pi^e_{t+1|t} - (z_t - z^e_{t+1|t})\right) . \tag{43}$$

[53]Even though I focus on supply shocks, the increase in people's attention to inflation in 2021 likely also increased the inflationary effects of demand shocks, such as monetary stimulus, given that I also find demand shocks to be more inflationary in times of higher attention (Figure 13).

The representative household therefore needs to form expectations about future consumption $c^e_{t+1|t}$, inflation $\pi^e_{t+1|t}$ and the exogenous preference shock $z^e_{t+1|t}$. The household holds rational expectations about the shock, so $z^e_{t+1|t} = E_t z_{t+1}$, where $E_t(\cdot)$ denotes the rational expectations operator, and since z_t follows an AR(1) process with persistence ρ_z, it follows that $z^e_{t+1|t} = E_t z_{t+1} = \rho_z z_t$.

In order to express equation (43) in terms of the output gap, we need to derive the efficient allocation, i.e., the one that prevails in the economy with fully flexible prices. From the production function, we have $Y^*_t = H^*_t$. The real wage is constant $w^*_t = 1$. From the labor-leisure equation (9), we get that potential output is therefore also constant and equal to:

$$Y^*_t = \Xi^{-1}. \tag{44}$$

Thus, potential output in log-deviations is 0. The Euler equation in the flexible-price economy is therefore given by:

$$0 = -\left(r_t - (z_t - z^e_{t+1|t})\right).$$

Since the natural rate is defined as the real rate that prevails under flexible prices, r_t, it follows directly that:

$$r^*_t = z_t - z^e_{t+1|t}. \tag{45}$$

Substituting $z_t - z^e_{t+1|t}$ with r^*_t in (43) and using that $c_t = y_t = x_t$, since potential output is 0, yields the IS equation (22):

$$x_t = c^e_{t+1|t} - \left(i_t - \pi^e_{t+1|t} - r^*_t\right). \tag{46}$$

Note, however, that if we assume initial values $c^e_{0|-1} = x^e_{0|-1}$, it follows that $c^e_{t+1|t} = x^e_{t+1|t}$ because $c_t = x_t$ in equilibrium. This holds true, even if the household does not know that they are equal in equilibrium. We can thus write equation (46) as:

$$x_t = x^e_{t+1|t} - \left(i_t - \pi^e_{t+1|t} - r^*_t\right), \tag{47}$$

which is equation (22) in the main text.

Derivation of equation (21). To derive the New Keynesian Phillips Curve (21), we start by maximizing:

$$\Omega_0(j) = \tilde{E}_0 \sum_{t=0}^{\infty} \beta^t \left[(1-\tau_t) P_t(j) \left(\frac{P_t(j)}{P_t}\right)^{-\epsilon} \frac{Y_t}{P_t} - W_t H_t(j) - \frac{\psi}{2}\left(\frac{P_t(j)}{P_{t-1}(j)} - 1\right)^2 Y_t + T^F_t(j)\right],$$

which yields the first-order condition:

$$(1-\tau_t)(\epsilon-1)P_t(j)^{-\epsilon}\frac{Y_t}{P_t^{1-\epsilon}}=\epsilon MC_t\left(\frac{P_t(j)}{P_t}\right)^{-\epsilon-1}\frac{Y_t}{P_t}-\psi\left(\frac{P_t(j)}{P_{t-1}(j)}-1\right)\frac{Y_t}{P_{t-1}(j)}$$
$$+\beta\psi\tilde{E}_t\left[\left(\frac{P_{t+1}(j)}{P_t(j)}-1\right)\frac{P_{t+1}(j)}{P_t(j)}\frac{Y_{t+1}}{P_t(j)}\right].$$

I define $T_t\equiv 1-\tau_t$, and set it to eliminate markup distortions in steady state, i.e., $T=\frac{\epsilon}{\epsilon-1}$. Linearizing around the zero-inflation steady state in which $P(j)=P$ for all j, I obtain:

$$T(\epsilon-1)\frac{Y}{P}\left[t_t-\epsilon p_t(j)+y_t-(1-\epsilon)p_t\right] \tag{48}$$
$$=\epsilon MC\frac{Y}{P}\left[(-\epsilon-1)p_t(j)-(-\epsilon-1)p_t+mc_t+y_t-p_t\right] \tag{49}$$
$$-\psi\frac{Y}{P}\left(y_t-p_{t-1}(j)+\pi_t(j)-y_t+p_{t-1}(j)\right) \tag{50}$$
$$+\beta\psi\frac{Y}{P}\tilde{E}_t^j\pi_{t+1}^j. \tag{51}$$

Grouping terms, using $MC=1$, $T=\frac{\epsilon}{\epsilon-1}$ and dividing by $\frac{Y}{P}$ yields:

$$p_t(j)=\frac{1}{\psi+\epsilon}\left[\psi p_{t-1}(j)+\epsilon\left(mc_t-t_t+p_t\right)+\beta\psi\tilde{E}_t^j\pi_{t+1}^j\right]. \tag{52}$$

Given the assumptions discussed in Section 4, it follows that $\tilde{E}_t^j\pi_{t+1}^j=\pi_{t+1|t}^e$ and $p_{t-1}(j)=p_{t-1}$ such that:

$$p_t(j)=\frac{1}{\psi+\epsilon}\left[\psi p_{t-1}+\epsilon\left(mc_t-t_t+p_t\right)+\beta\psi\pi_{t+1|t}^e\right]. \tag{53}$$

Therefore, the optimal price $p_t(j)$ is the same for all firms j, $p_t(j)=p_t$, such that we get:

$$\psi\underbrace{\left(p_t-p_{t-1}\right)}_{=\pi_t}=\epsilon\left(mc_t-t_t\right)+\beta\psi\pi_{t+1|t}^e. \tag{54}$$

Dividing by ψ yields:

$$\pi_t=\frac{\epsilon}{\psi}\left(mc_t-t_t\right)+\beta\pi_{t+1|t}^e. \tag{55}$$

Using $mc_t=y_t$ and $y_t=x_t$ and defining cost-push shocks as $u_t\equiv-\frac{\epsilon}{\psi}t_t$ and the slope parameter $\kappa\equiv\frac{\epsilon}{\psi}$, I arrive at the linearized New Keynesian Phillips Curve under subjective expectations:

$$\pi_t=\kappa x_t+\beta\pi_{t+1|t}^e+u_t. \tag{56}$$

D.1 Unemployment expectations

In order to calibrate the attention parameter with respect to the output gap, γ_x, I use expectations about unemployment changes from the Survey of Consumers. The survey asks households whether they expect unemployment to increase, decrease or to remain about the same over the next twelve months. I follow Carlson and Parkin (1975), Bhandari et al. (2025) and Pfäuti and Seyrich (2023) to translate these categorical unemployment expectations into numerical expectations.

Let q_t^D, q_t^S and q_t^U denote the shares reported at time t that think unemployment will go down, stay roughly the same, or go up over the next year, respectively. I assume that these shares are drawn from a cross-sectional distribution of responses that are normally distributed according to $\mathcal{N}\left(\mu_t, (\sigma_t)^2\right)$ and a threshold a such that when a household expects unemployment to remain within the range $[-a, a]$ over the next year, she responds that unemployment will remain "about the same". We thus have:

$$q_t^D = \Phi\left(\frac{-a-\mu_t}{\sigma_t}\right) \qquad q_t^U = 1 - \Phi\left(\frac{a-\mu_t}{\sigma_t}\right),$$

which after some rearranging yields:

$$\begin{aligned}\sigma_t &= \frac{2a}{\Phi^{-1}\left(1-q_t^U\right) - \Phi^{-1}\left(q_t^D\right)} \\ \mu_t &= a - \sigma_t \Phi^{-1}\left(1-q_t^U\right).\end{aligned}$$

This leaves us with one degree of freedom, namely a. I follow Pfäuti and Seyrich (2023) and set $a = 0.5$ which means that if a household expects the change in unemployment to be less than half a percentage point (in absolute terms), she reports that she expects unemployment to be about the same as it is at the time of the survey. I use data from FRED for the actual unemployment changes and restrict the sample to end in 2019Q4, due to the extreme behavior of unemployment changes with the outbreak of the Covid-19 pandemic.

I then estimate γ_x separately for whether lagged inflation is above or below the estimated threshold of 3.91%. This results in estimates $\gamma_{x,L} = 0.254$ and $\gamma_{x,H} = 0.256$. Thus, there are practically no differences in attention to unemployment changes across regimes and hence, I impose γ_x to be the same across regimes in the model.

E Additional Figures and Model Results

E.1 AS-AD representation

To provide intuition for how the attention threshold can trigger self-reinforcing inflation surges, I use here a slightly stylized version of the model. In particular, I assume that agents are completely inattentive to the output gap, i.e., $\gamma_x = 0$, and that the Taylor rule is given by $i_t = \phi_\pi \pi_t$ with $\phi_\pi > 1$.

The economy starts in the steady state, with $u_0 = 0$, $r_0^* = 0$ and prior expectations at their long-run averages of 0: $\pi^e_{0|-1} = 0$ and $x^e_{0|-1} = 0$. The aggregate supply (AS) curve—derived by plugging in the inflation expectations in the Phillips Curve—in the initial period is given by:

$$AS_0: \quad \pi_0 = \frac{\kappa}{1-\beta\gamma_{\pi,L}} x_0, \tag{57}$$

and aggregate demand (AD), which follows from combining the Taylor rule with the aggregate Euler (or IS) equation, is given by:

$$AD_0: \quad \pi_0 = -\frac{1}{\phi_\pi - \gamma_{\pi,L}} x_0. \tag{58}$$

Panel (a) in Figure 28 depicts this initial situation graphically.[54] Inflation and the output gap are both at their steady state values of 0 and therefore, below the inflation attention threshold $\bar{\pi}$.

In period 1, a positive cost-push shock hits and I assume that it persists for two periods: $u_1 = u_2 > 0$ and returns to zero afterwards, $u_t = 0$ for $t \geq 3$. The AS and AD equations are now given by:

$$\begin{aligned} AS_1: \quad \pi_1 &= \frac{\kappa}{1-\beta\gamma_{\pi,L}} x_1 + \frac{1}{1-\beta\gamma_{\pi,L}} u_1 \\ AD_1: \quad \pi_1 &= -\frac{1}{\phi_\pi - \gamma_{\pi,L}} x_1. \end{aligned}$$

This situation is shown in panel (b) of Figure 28. The cost-push shock shifts the AS curve up along the AD curve. The resulting equilibrium is characterized by output below potential, i.e., a negative output gap, and positive inflation. The shock is assumed to be large enough, such that inflation exceeds the threshold.

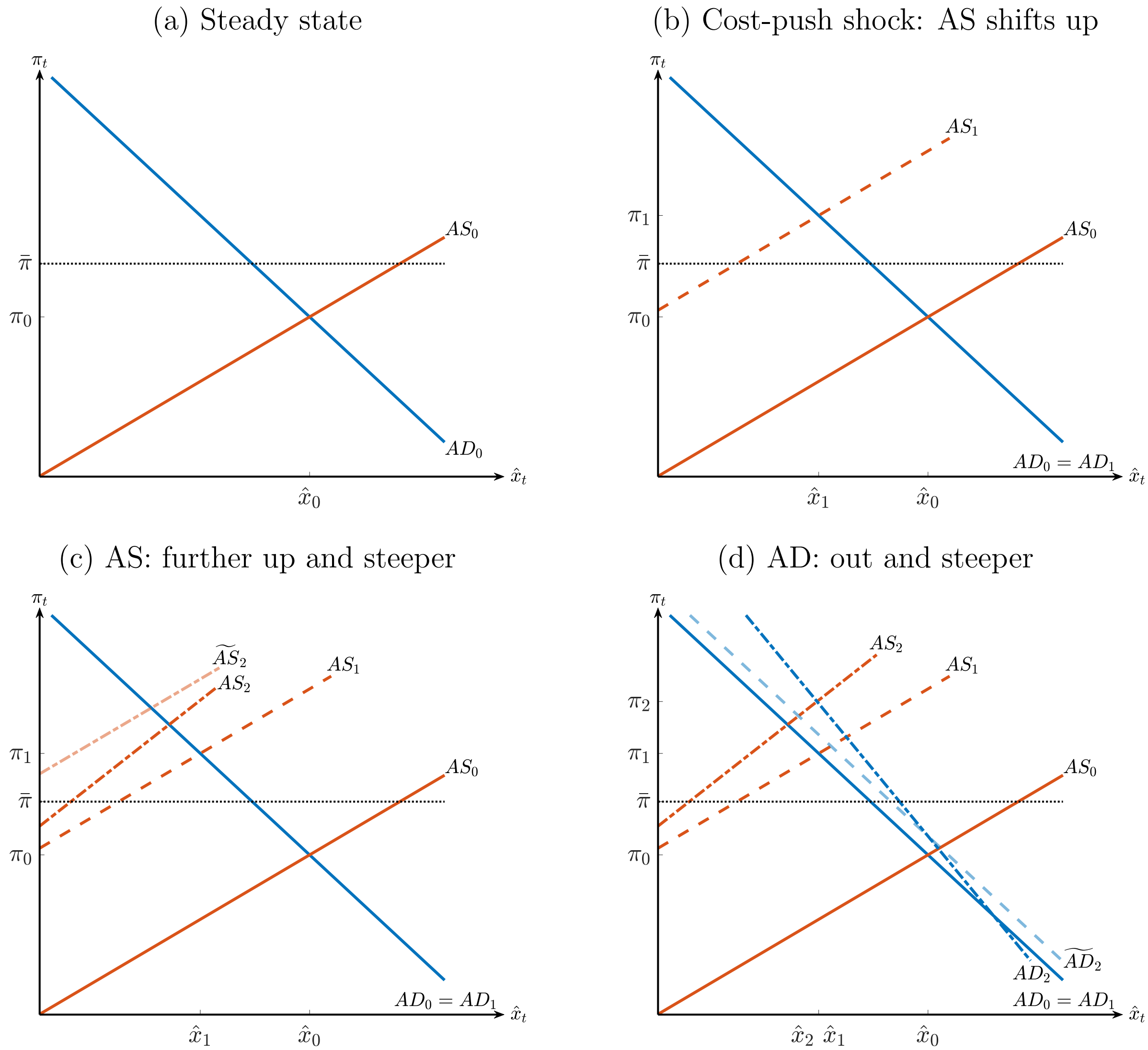
Due to the increase in inflation in period 1, firms enter period 2 with positive prior inflation expectations: $\pi^e_{2|1} = \gamma_{\pi,L}\pi_1 > 0$. These higher prior expectations together with the still ongoing cost-push shock shift the AS curve further up. This shift is illustrated in panel (c) of Figure 28 by the $\widetilde{AS}_2$ curve, which is given by:

$$\widetilde{AS}_2: \quad \pi_2 = \frac{\kappa}{1-\beta\gamma_{\pi,L}} x_2 + \underbrace{\frac{1}{1-\beta\gamma_{\pi,H}} u_2 + \frac{\beta(1-\gamma_{\pi,H})\gamma_{\pi,L}}{1-\beta\gamma_{\pi,H}} \pi_1}_{\text{Intercept}>0}.$$

The terms denoted "Intercept" capture this shift in the AS curve. Since inflation in the previous

[54] The values I use in this stylized example are: $\bar{\pi} = 4\%$, $\gamma_{\pi,L} = 0.2$, $\gamma_{\pi,H} = 0.4$, $\beta = 0.99$, $\kappa = 0.6$, $\phi_\pi = 1.05$ and $u_1 = u_2 = 10$.

Figure 28: An illustrative example



Notes: Panel (a) shows the initial situation in period 0 when the economy is in the steady state. In period 1, a positive cost-push shock hits the economy, leading to an upward shift of the AS curve (panel (b)). In period 2, the AS curve shifts further up (shift from AS_1 to $\widetilde{AS}_2$) and the curve becomes steeper (rotation from $\widetilde{AS}_2$ to AS_2), shown in panel (c). Simultaneously, the AD curve shifts out (shift from AD_1 to $\widetilde{AD}_2$) and becomes steeper (rotation from $\widetilde{AD}_2$ to AD_2), as shown in panel (d). The black-dotted line at $\bar{\pi}$ depicts the inflation attention threshold.

period exceeded the attention threshold, attention is now higher. This increase in attention leads to an unambiguously stronger effect of the cost-push shock compared to the case in which attention would have remained constant:

$$\frac{1}{1-\beta\gamma_{\pi,H}}u_2 > \frac{1}{1-\beta\gamma_{\pi,L}}u_2.$$

The cost-push shock leads to an increase in inflation, and this inflation increase now leads to a larger increase in firms' inflation expectations due to their higher attention. These higher expectations then feed back into higher prices and thus, higher inflation.

The effect of the increase in firm managers' prior expectations on inflation in the second period, captured by $\frac{\beta(1-\gamma_{\pi,H})\gamma_{\pi,L}}{1-\beta\gamma_{\pi,H}}\pi_1$, however, is smaller at the higher attention level. There are two counteracting forces. First, the increase in attention means that firm managers now update their expectations more strongly and put less weight on their prior expectations. This per se leads to a smaller effect. Second, the higher prior increases overall inflation expectations which, ceteris paribus, increases current inflation. But because inflation expectations are discounted by $\beta \leq 1$, the first effect dominates. Therefore, the shift in the AS curve due to the higher prior expectations is smaller at higher attention levels. Quantitatively, however, these differences are very small as $\beta \approx 1$. In fact, if $\beta = 1$, the change in attention has no effect on the shift due to higher prior expectations, and therefore, the total shift of the AS curve is unambiguously higher when attention is higher, due to stronger effect of the cost-push shock.

The shift of the AS curve to $\widetilde{AS}_2$, however, is only part of the story because the increase in attention also steepens the slope of the AS curve. That is, the aggregate supply curve becomes steeper in periods of high attention—a *dynamic* non-linearity. Taking this into account, the AS curve in the second period is given by:

$$AS_2 : \quad \pi_2 = \underbrace{\frac{\kappa}{1-\beta\gamma_{\pi,H}}}_{\text{Slope}} x_2 + \frac{1}{1-\beta\gamma_{\pi,H}} u_2 + \frac{\beta(1-\gamma_{\pi,H})\gamma_{\pi,L}}{1-\beta\gamma_{\pi,H}} \pi_1.$$

This steepening of the AS curve is illustrated in panel (c) of Figure 28 by the rotation of the AS curve from $\widetilde{AS}_2$ to AS_2. This steepening of the AS curve eases the inflationary pressures due to the negative output gap. Nevertheless, the steeper AS curve implies that if the AD curve would now shift out, the inflationary effects of this increase in demand would become larger.

It turns out that the higher prior expectations endogenously lead to such a demand increase. This is illustrated by $\widetilde{AD}_2$ in panel (d) of Figure 28, which is given by:

$$\widetilde{AD}_2 : \quad \pi_2 = -\frac{1}{\phi_\pi - \gamma_{\pi,L}} x_2 + \frac{(1-\gamma_{\pi,H})\gamma_{\pi,L}}{\phi_\pi - \gamma_{\pi,H}} \pi_1.$$

The higher prior expectations, ceteris paribus, decrease the real rate which leads to the outward shift of the AD curve. These effects, however, are smaller at higher levels of attention as long as the Taylor principle, $\phi_\pi > 1$, is satisfied, because in that case the higher inflation rates due to the higher prior expectations are counteracted by a more than one-for-one increase in the nominal rate. If monetary policy is relatively dovish, i.e., ϕ_π is close to 1, these differences are small. Furthermore, since the AD curve is now shifted along a steeper AS curve due to the heightened attention, the inflationary effects of a given shift are larger at the higher level of attention. This also implies that additional demand stimulus—for example, due to loose monetary policy or a fiscal stimulus—would have relatively large inflationary effects.

Additionally, the AD curve also becomes steeper, as illustrated by the rotation from $\widetilde{AD}_2$ to AD_2

in panel (d), where AD_2 is given by:

$$AD_2: \quad \pi_2 = -\frac{1}{\phi_\pi - \gamma_{\pi,H}} x_2 + \frac{(1-\gamma_{\pi,H})\gamma_{\pi,L}}{\phi_\pi - \gamma_{\pi,H}} \pi_1.$$

This leads to a further increase in inflation, especially now because the AS curve is steep. In this stylized example, inflation increases substantially from period 1 to period 2 through the change in attention, whereas the output gap remains practically constant. Thus, the attention threshold offers a potential mechanism for how inflation surges may occur and exhibit self-reinforcing dynamics without large changes in output.

After the shock. In the third period, when the shock has died out, the AS curve shifts back down. The AS curve is given by:

$$AS_3: \quad \pi_3 = \frac{\kappa}{1-\beta\gamma_{\pi,H}} x_3 + \frac{\beta(1-\gamma_{\pi,H})}{1-\beta\gamma_{\pi,H}} \pi^e_{3|2}.$$

Due to the positive prior expectations, $\pi^e_{3|2} > 0$, the AS curve does not fully shift back to its initial position but remains above it. Due to the increased steepness of the AD curve, however, the shift in the AS curve leads to a stronger reduction in inflation compared to a flatter AD curve.

While the AS curve comes back down, the AD curve shifts further out due to the positive prior expectations that agents have when going into period 3. The AD curve is given by:

$$AD_3: \quad \pi_3 = -\frac{1}{\phi_\pi - \gamma_{\pi,H}} x_3 + \frac{1-\gamma_{\pi,H}}{\phi_\pi - \gamma_{\pi,H}} \pi^e_{3|2}.$$

Thus, output recovers more strongly during this disinflationary period than what would be the case absent this shift in the AD curve, and disinflation occurs more gradually. These results are shown graphically in Figure 29.

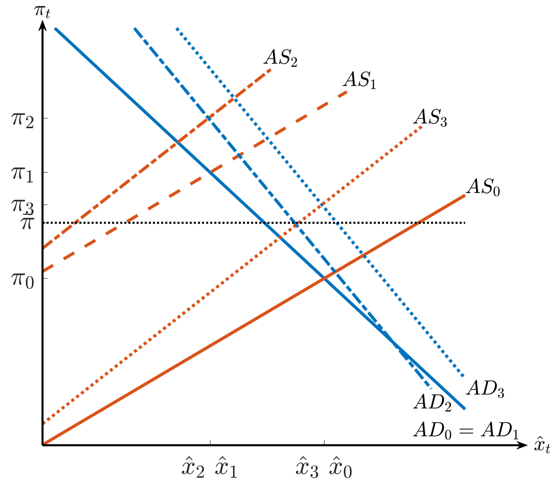
E.1.1 Testing for changes in the slopes of the AS and AD curve

I now test the predictions how changes in people's attention to inflation change the slope of the AS and the AD curve. As the output gap is unobserved, I use the unemployment rate and postulate a negative relationship between unemployment and the output gap: $x_t = -\Psi un_t$, where un_t is the unemployment rate and $-\Psi$ is some constant. Abstracting from shocks, in any given period t, the AD and AS curves can then be expressed as:

$$\text{AD:} \quad un_t = \frac{(\phi_\pi - \gamma_{\pi,r})}{\Psi} \pi_t - \frac{(1-\gamma_{\pi,r})}{\Psi} \pi^e_{t|t-1} \tag{59}$$

$$\text{AS:} \quad \pi_t = -\frac{\kappa}{1-\beta\gamma_{\pi,r}} \Psi un_t + \frac{\beta(1-\gamma_{\pi,r})}{1-\beta\gamma_{\pi,r}} \pi^e_{t|t-1}. \tag{60}$$

Figure 29: Illustrative example: after the shock has died out



Notes: This figure depicts the third period in which the cost-push shock has died out (after being positive for two consecutive periods). The black-dotted line at $\bar{\pi}$ depicts the inflation attention threshold.

Note that I represent the AD curve with unemployment on the left hand side of the equation, not inflation. This will help to illustrate the empirical approach I take to estimate the slope of the AD curve. The theory predicts that a given change in inflation should lead to a smaller increase in the unemployment rate when attention to inflation is higher, as can be seen from the AD curve. The reason is that higher inflation leads to a more than one-for-one response of nominal interest rates via the Taylor rule and the resulting higher real rates increase unemployment. When households are more attentive, however, their inflation expectations increase more strongly, and hence, the ex-ante real rate increases by less, so that unemployment increases by less.

For the slope of the AS curve, the theory predicts that a given decrease in the unemployment rate leads to more inflation when firm managers are more attentive. The reason is that their inflation expectations increase more strongly and hence, they increase their prices by more.

I test those predictions following the approach in Barnichon and Mesters (2021) and Barnichon and Mesters (2020). In all specifications, I use regional data on unemployment and CPI inflation, and I control for region specific fixed effects, as well as for past inflation expectations, as in equations (59) and (60).

To test the predictions regarding the AD curve, I regress the current unemployment rate on contemporaneous inflation and instrument inflation with contemporaneous supply shocks. In particular, I again use the oil supply news shocks from Känzig (2021).

Columns (1) and (2) in Table 9 show the results. We see that in both regimes the estimated slope is significantly positive in both regimes and consistent with the theory, it is flatter in times of high attention. The robust F-statistic is larger than 20 in both regimes.

To test the predictions regarding the slope of the AS curve, I regress inflation on the unemployment rate and instrument unemployment using monetary policy shocks, as in Barnichon and Mesters (2021). I use the shocks from Jarociński and Karadi (2020) as well as from Aruoba and

Table 9: Empirical AS and AD curve slopes

	AD slope		AS slope	
	$\pi_{t-1} < 3.91$	$\pi_{t-1} \geq 3.91$	$\pi_{t-1} < 3.91$	$\pi_{t-1} \geq 3.91$
	(1)	(2)	(3)	(4)
Slope estimates	0.25	0.17	−0.38	−0.74
s.e.	(0.147)	(0.103)	(0.113)	(0.091)
p-value	0.084	0.088	0.001	0.000
N	1,316	408	728	148
Control for $\pi^e_{t\|t-1}$	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes
Robust F-stat.	20.89	28.92	10.92	17.28

Notes: This table shows the estimated slopes of the AD (columns (1) and (2)) and the AS curves (columns (3) and (4)) for the two attention regimes. The last row provides the heteroskedasticity-robust first-stage F–statistic.

Drechsel (2024) to get more power. Consistent with the findings in Barnichon and Mesters (2021), I find very weak relations when considering contemporaneous effects as well as concerns regarding weak instruments. This is not surprising given that monetary policy shocks are often claimed to have delayed effects on unemployment (Aruoba and Drechsel, 2024). Barnichon and Mesters (2021) show that the relationship between unemployment and inflation strengthens when considering cumulative changes in those variables over longer horizons. I therefore follow their approach and use the 12-month-ahead cumulative change in the CPI as my dependent variable and the 12-month-ahead cumulative change in the unemployment rate. For the instruments, I include the contemporaneous shocks as well as six lags of the shocks.

Columns (3) and (4) in Table 9 show the results. In both regimes, the slope is significantly negative, and as predicted by the theory, the slope steepens in times of high attention. The robust F-statistics are > 10 in both cases, even though they are weaker than when considering oil shocks. This evidence is consistent with the findings in Barnichon and Mesters (2021), who show that the slope of the AS curve has indeed been flatter after 1990 than before. The time before 1990 largely coincides with the high attention regime (their sample does not include the recent inflation surge).

A simple back of the envelope calculation shows that the very simple model in equations (59) and (60) does surprisingly well also quantitatively. The slope of the AD curve is $\frac{\phi_\pi - \gamma_\pi}{\Psi}$. If we fix $\phi_\pi = 2$, we can back out Ψ from the estimate of the slope in times of low attention. Doing so implies $0.25 = \frac{\phi_\pi - \gamma_\pi}{\Psi} = \frac{2-0.18}{\Psi}$, from which we obtain $\Psi = 7.3$. Using this Ψ implies a slope in times of high attention of $\frac{2-0.35}{7.3} = 0.22$ which is not too far from the empirical estimate of 0.17, even though the empirical findings suggest a somewhat more pronounced change in the slope of the AD curve across regimes. The AS curve slope is equal to $-\Psi\frac{\kappa}{1-\beta\gamma_\pi}$. Using my baseline calibration, I obtain a slope in times of low attention of −0.5 and in times of high attention of −0.64. Again, the change is somewhat stronger empirically, but given the simplicity of this model and that those predictions are untargeted, the model does surprisingly well.

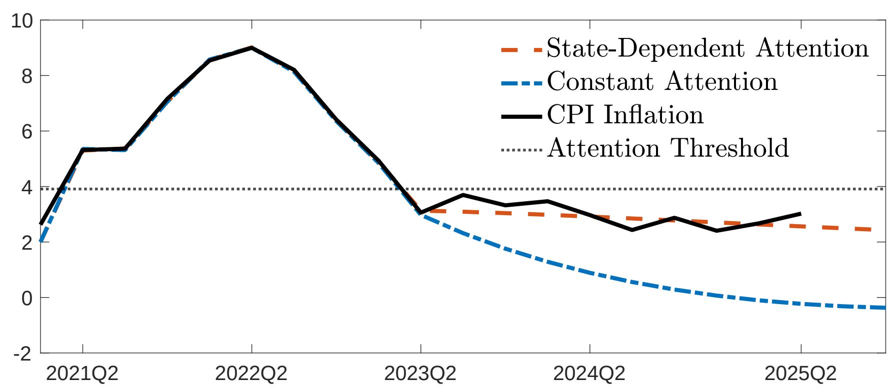
E.2 Additional figures to Section 5.1

In this section, I discuss alternative model specifications and additional sensitivity analyses related to the model results in Section 5.1.

The long last mile of disinflation. To provide an estimate of the quantitative role of the attention threshold for the recently observed long last mile of disinflation, I consider the following counterfactual. I feed in cost-push shocks to my model to perfectly replicate the recent inflation surge up until inflation fell back below the threshold. After that, I assume that there were no more shocks and examine what inflation dynamics the model produces endogenously. I do this exercise for my model featuring the attention threshold and the counterfactual model in which attention remains high after the inflation surge. Note that replicating the inflation surge before the return to low attention is important for this exercise as the inflation surge shapes people's prior expectations which are an important mechanism for the long last mile of disinflation in my model.

Figure 30 shows the results. The black-solid line shows actual U.S. CPI inflation, the red-dashed line shows my baseline model with the attention threshold, and the blue-dashed-dotted line shows the model where attention is constantly high. Up until 2023Q2, the shocks are set to match the actual inflation perfectly. The rest of the inflation dynamics are endogenously determined within the respective model. We can see that the model with the attention threshold matches the long last mile of disinflation almost perfectly. The model where attention remains high, on the other hand, performs poorly as it would predict a very sharp phase of disinflation and even deflation. Of course, there were exogenous shocks hitting the U.S. economy also after 2023Q2 that are not directly considered in this exercise that may have affected the last mile as well. The results in Figure 30, however, indicate that the net effects of these other forces may be small. If attention remains higher after the inflation surge, this would dampen the role of the attention threshold for the last mile of disinflation. But as long as attention does decrease, it contributes to the longer last mile. I discuss this possibility more extensively below in the discussion of Figure 36.

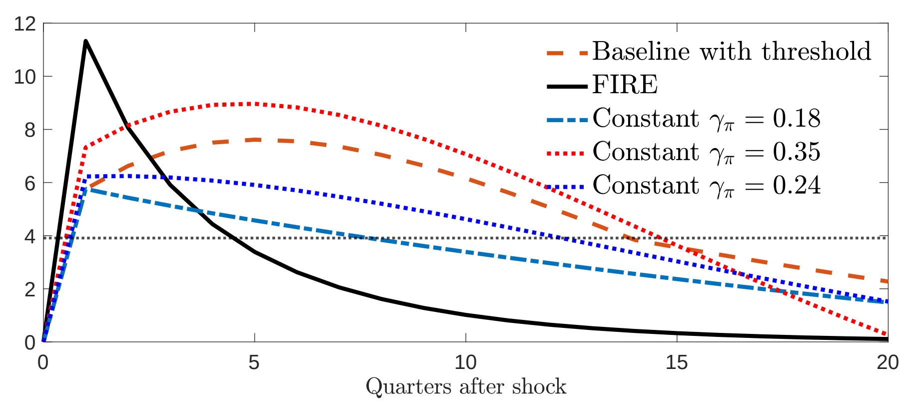
Figure 30: The last mile of disinflation



Notes: This figure shows the role of the attention threshold for the last mile of disinflation.

Alternative model specifications. Figure 31 shows the inflation response to a cost-push shock for five different models. My model with the attention threshold (labeled "Baseline with threshold" and depicted by the red-dashed line) shows the discussed hump-shaped pattern of the inflation surge and the decrease in the speed of disinflation once inflation falls back below the threshold. As the other four illustrate, these inflation dynamics are unique to the model with the attention threshold. In particular, the models under FIRE (black-solid line) or in which attention remains low (blue-dashed-dotted line) predict that inflation peaks on impact and then comes down relatively quickly. The same is true for the model with a constant attention parameter of 0.24, which corresponds to the estimated attention level in the data if I do not impose a threshold (the inflation response is captured by the blue-dotted line). Additionally, models with FIRE could not account for the observed dynamics of forecast errors. A model variant in which attention to inflation is always high (red-dotted), on the other hand, can produce the initial hump-shaped response but then predicts a very fast decline back to target, i.e., this model would not explain why we see such a long last mile of disinflation. Furthermore, a model in which attention is always high would have predicted a very pronounced deflationary period after the Great Financial Crisis. However, one thing that the model with high-attention illustrates, is that the inflation response is hump shaped when attention is high whereas it peaks on impact when the shock hits in times of low attention, consistent with the empirical findings (panels (a) and (b) in Figure 5).

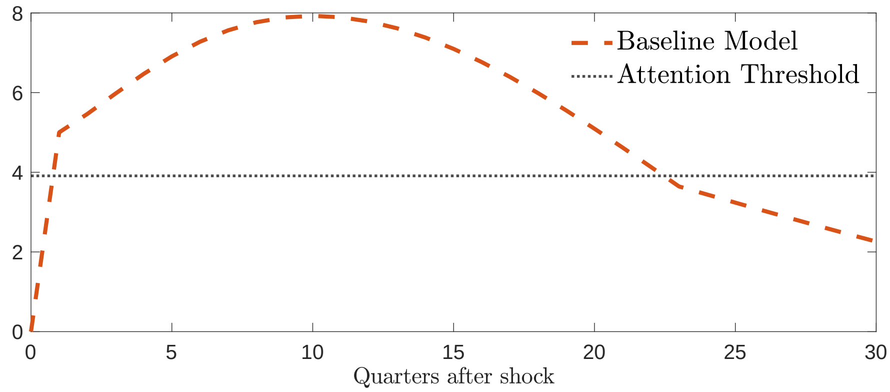
Figure 31: Inflation dynamics under FIRE or absent the attention change



Notes: This figure shows the inflation response to a cost-push shock for different model specifications.

Demand shocks. Figure 32 shows the inflation response to a positive demand shock. We see that the discussed inflation dynamics after a relatively large supply shock are similar to the ones after a demand shock: inflation keeps on increasing, fueled by the attention increase, and the speed of disinflation slows substantially once inflation falls back below the threshold and attention therefore decreases. Note, however, that the inflation surge is somewhat longer lived than in response to the supply shock.

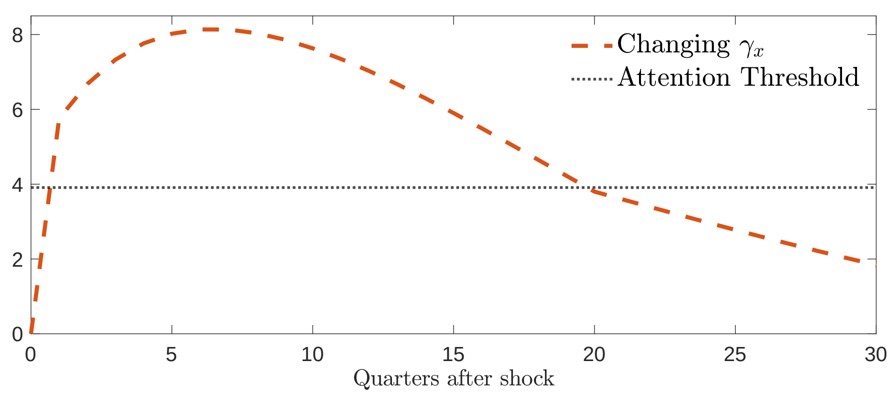
Figure 32: Inflation dynamics after a demand shock



Notes: This figure shows the inflation response in the model with the attention threshold in response to a demand shock r_t^* with persistence $\rho_r = 0.55$.

Time-varying attention to the output gap. Figure 33 shows the inflation response for the case in which attention to the output gap, γ_x, decreases when attention to inflation increases. This may reflect that agents shift their attention from the output gap to inflation when inflation is high. We see that the inflation surge becomes more persistent in that case (note that the x-axis now shows 30 periods rather than 20). The reason is that the negative output gap due to the adverse cost-push shock leads to a smaller downward revision of output gap expectations, thus, agents remain relatively more optimistic about the output gap, and hence, consume relatively more, leading to additional upward pressure on inflation.

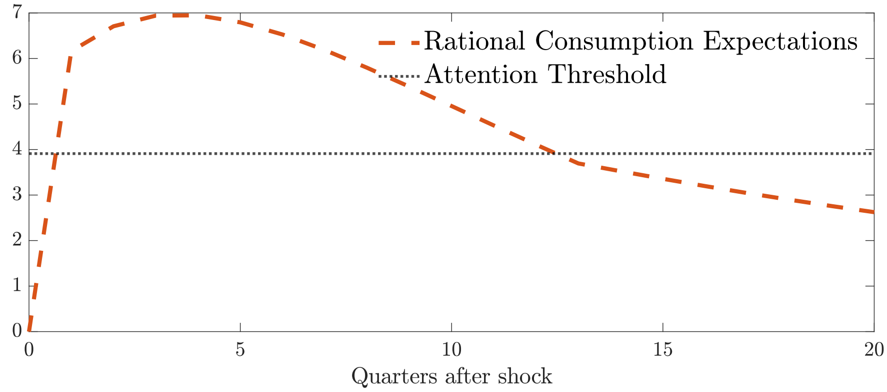
Figure 33: Inflation dynamics with $\gamma_{x,L} = 0.25$ and $\gamma_{x,H} = 0.1$



Notes: This figure shows the inflation response to a cost-push shock when attention to the output gap decreases in times of high attention to inflation.

Rational consumption expectations. Figure 34 shows the inflation response to a cost-push shock when consumption expectations are formed under full-information rational expectations. We see that the dynamics are qualitatively similar to the baseline model. The endogenous inflation increase is somewhat muted here because the forward-looking consumption expectations do not induce endogenous persistence in the way they do in the baseline model. When increasing the exogenous persistence of the shock slightly, however, the models are very similar (not shown).

Figure 34: Inflation dynamics with rational consumption expectations



Notes: This figure shows the inflation response to a cost-push shock when consumption expectations are formed under full-information rational expectations.

Smooth attention changes. What happens if attention changes smoothly with inflation? The upper panel in Figure 35 shows the inflation response to a cost-push shock when attention is a linear function of inflation (blue-dashed-dotted line) together with the inflation response in my baseline model with a threshold (red-dashed line). The specification I use for attention is $\gamma_{\pi,t} = \gamma_{\pi,L} + a \cdot \pi_t$, where $\gamma_{\pi,L} = 0.18$ is set equal to the attention estimate I obtain in the low-attention regime. I set $a = 7$ in which case $\gamma_{\pi,t}$ peaks at a similar level as my attention estimate in the high-attention regime.[55] The lower panel of Figure 35 shows the time series of $\gamma_{\pi,t}$ in response to the cost-push shock.

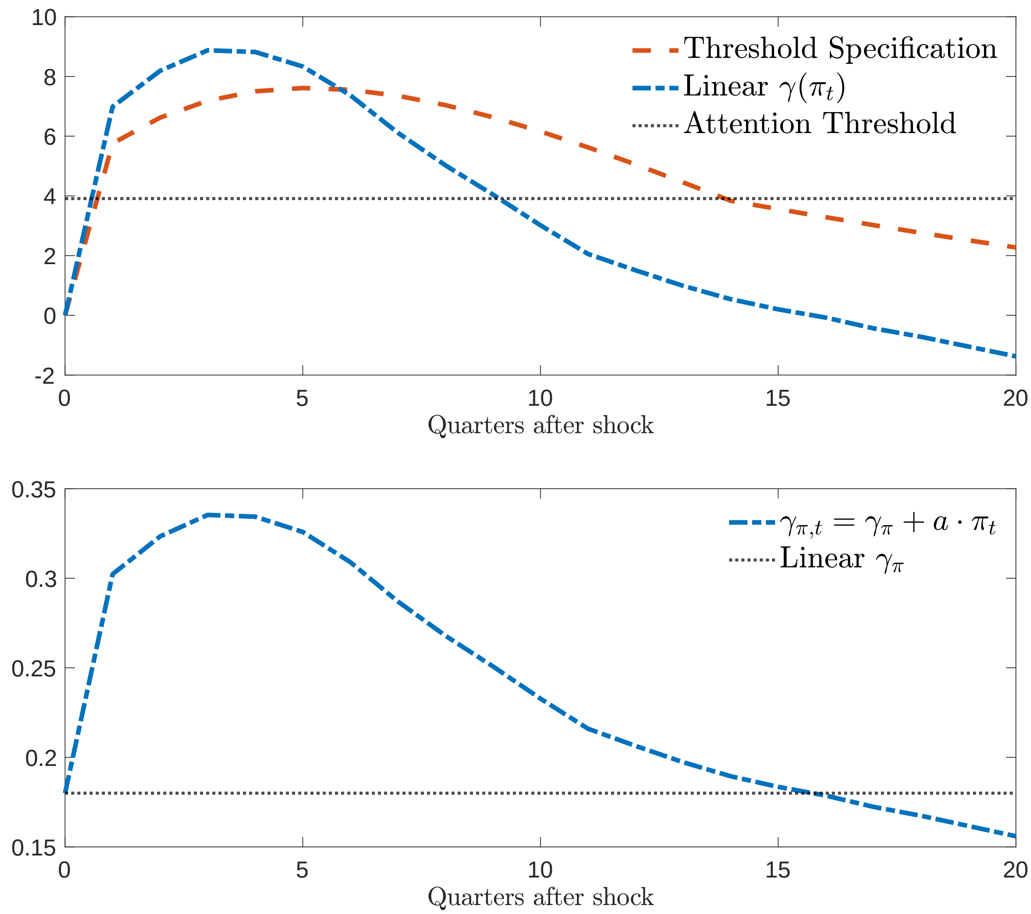
We see that the two models predict similar though slightly different inflation dynamics. Both predict a hump-shaped inflation response, a relatively fast first phase of disinflation which then slows down. The model with a continuous $\gamma(\pi_t)$, however, predicts a somewhat shorter-lived inflation surge. The reason is that $\gamma_{\pi,t}$ starts falling relatively quickly and hence, the inflationary pressure from inflation expectations becomes muted. Therefore, inflation itself starts falling which decreases $\gamma_{\pi,t}$ further.

Three attention regimes. In a recent paper, Korenok and Munro (2024) show that even though attention—measured using Google Trends—has decreased as inflation started to come down, it remained at a higher level than before the inflation surge. To understand how this possibility affects my model's predictions, I consider a third regime. In particular, I allow for a third regime in which attention takes on an intermediate value of 0.24.[56] The economy reaches that regime once inflation falls below the threshold. The economy is assumed to stay in that regime for six quarters before going back to the low-attention regime. The red-solid line in Figure 36 shows the inflation response to a cost-push shock in that case. For comparison, the blue-dashed-dotted line shows my baseline result with two regimes, and the yellow-dotted line shows the counterfactual scenario in which attention remains at its high level of $\gamma_{\pi,H}$ even after inflation has fallen back below the

[55] Note, here π_t is quarterly inflation in decimals; whereas the figures report annualized percent.

[56] 0.24 is the average attention parameter I estimate when I do not allow for multiple regimes.

Figure 35: Smooth attention changes

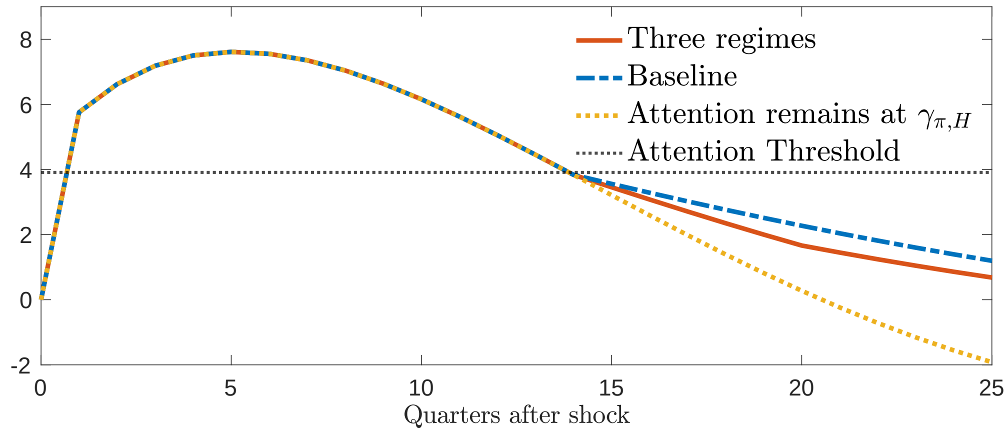


Notes: The upper panel shows the inflation response in my baseline model (red-dashed line) and in a model in which attention changes smoothly with inflation (blue-dashed-dotted line). The lower panel plots the dynamics of attention $\gamma_{\pi,t}$ (blue-dashed-dotted line). In contrast, the black-dotted line in the lower panel plots the constant low level of attention.

threshold. We can see that the slowdown in the speed of disinflation is somewhat muted, i.e., inflation comes down somewhat faster compared to my baseline model with two regimes. But nevertheless, there is still a pronounced slowdown in the inflation dynamics, indicating a longer last mile of disinflation compared to an economy absent a decrease in attention. With three regimes, there are two, not one, kinks in the disinflation phase. The first one occurs when inflation falls below the threshold and attention decreases from the high to the medium level. The second one happens in period 20, when attention decreases further to its low level. As long as there is some decrease in attention (which is consistent with the findings in Korenok and Munro (2024)), there is a slowdown in the speed of disinflation through the lens of my model.

No random walk. In the baseline calibration, I assume that the perceived persistence is set to $\tilde{\rho}_\pi = 1$. This random walk assumption ensures that inflation expectations and inflation coincide in the long run absent any shocks. Hence, I do not need to specify $\underline{\pi}$. I relax this assumption in the following.

Figure 36: Inflation dynamics with three attention regimes



Notes: The blue-dashed-dotted line shows the (annualized) inflation response to a cost-push shock in my baseline model with two attention regimes, the red-solid line shows the inflation response for the model with three regimes, and the yellow-dotted line shows the model response if attention remains at its high level of $\gamma_{\pi,H}$ even after inflation has fallen back below the threshold.

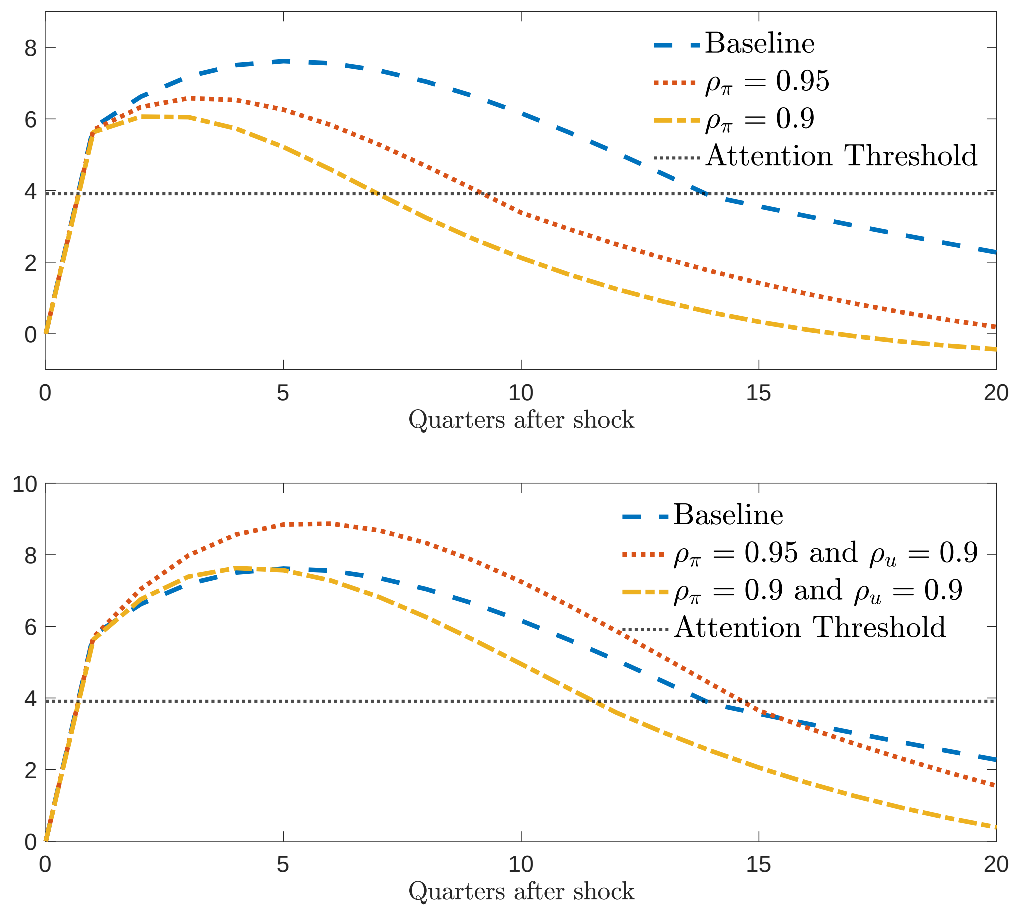
Figure 37 shows the inflation response to a cost-push shock for different calibrations. In all of them, I set $\underline{\pi} = 0$. This is a conservative assumption as imposing a positive $\underline{\pi}$ (as observed in the data) would lead to an upward bias in inflation expectations and hence, *ceteris paribus*, to more inflationary pressure. The blue-dashed line shows the baseline case with $\tilde{\rho}_\pi = 1$. The red-dotted line shows the case with $\tilde{\rho}_\pi = 0.95$ and the yellow-dashed-dotted line shows the case $\tilde{\rho}_\pi = 0.9$. Qualitatively, the three are all quite similar. Not surprisingly, however, the inflation surge is less persistent as the model produces less endogenous persistence when $\tilde{\rho}_\pi$ is lower. I therefore consider in the lower panel of Figure 37 the case in which the exogenous shock persistence ρ_u is increased from 0.8 to 0.9 and compare the results to my baseline calibration. We see that in this case, the three inflation responses are also quantitatively very similar.

Time-varying slope of the Phillips Curve. Figure 38 shows the model predictions if attention remains constant at its low level $\gamma_{\pi,L}$, but in which the price adjustment cost decreases, reflected in a temporarily higher κ. Decreasing the adjustment-cost parameter is a reduced-form way to capture the possibility of greater price flexibility. The blue-dashed-dotted line in the upper panel shows the inflation response to a cost-push shock assuming that κ doubles for 10 quarters after the shock. The lower panel shows the response to a demand shock. In both figures, the red-dashed line shows the responses in my baseline model with regime-dependent attention but constant κ.

The model with fixed attention and time-varying κ cannot explain the hump-shaped inflation response. In the case of the demand shock, disinflation happens very gradually. Thus, that model may explain the long last mile of disinflation, but it cannot simultaneously explain why disinflation happened relatively fast initially after inflation peaked. Nevertheless, the empirical evidence shows that the frequency of price adjustments does fluctuate over time (see, e.g., Blanco et al., 2024).

Figure 39 further provides the inflation dynamics after supply shocks for both small and large

Figure 37: Inflation dynamics with $\tilde{\rho}_\pi < 1$

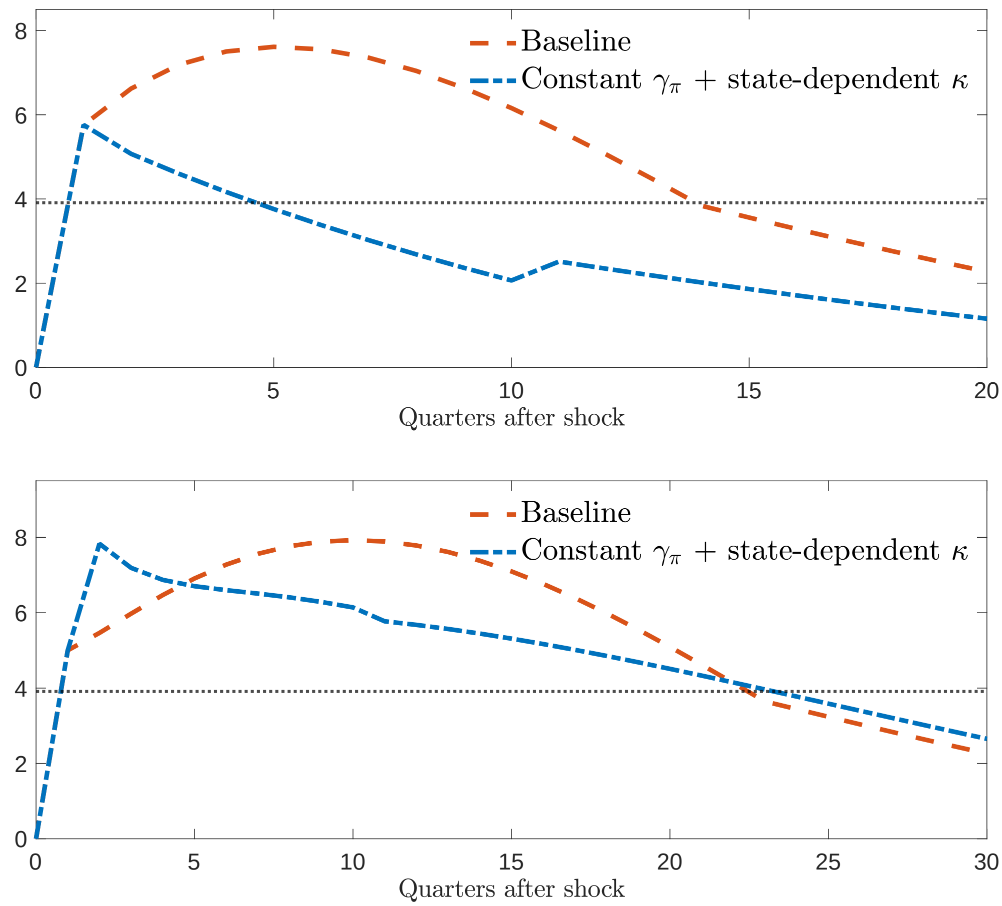


Notes: The blue-dashed line shows the (annualized) inflation response to a cost-push shock in the baseline model with $\tilde{\rho}_\pi = 1$ and $\rho_u = 0.8$. The red-dotted line shows the model with $\tilde{\rho}_\pi = 0.95$, and the yellow-dashed-dotted line shows the model with $\tilde{\rho}_\pi = 0.9$.

shocks for the two model specifications. The differences between the baseline model and the time-varying-κ alternative are smaller for the small shock, since inflation does not cross the attention threshold and hence attention remains fixed.

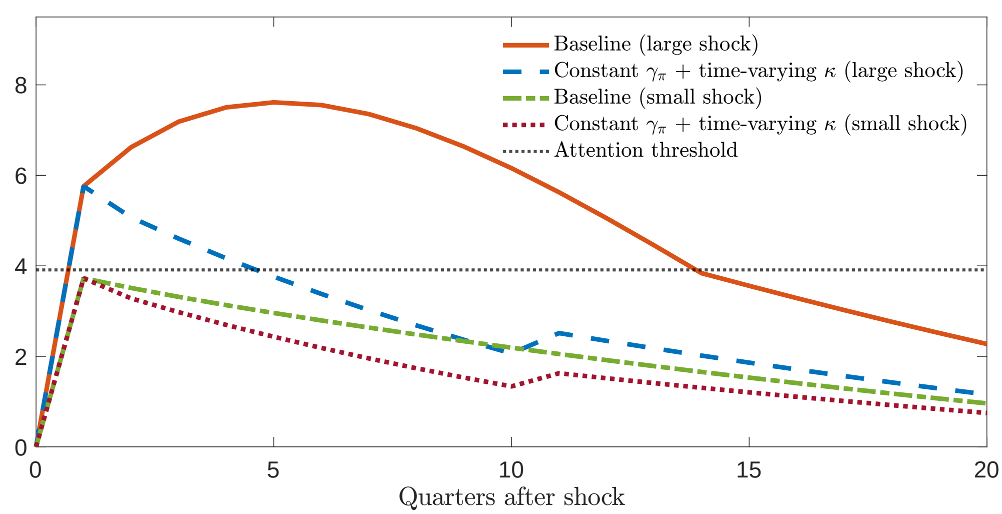
A full analysis of endogenous price adjustment frequency and the attention threshold is left for future work.

Figure 38: Inflation dynamics with state-dependent κ



Notes: The red-dashed line in the upper panel shows the (annualized) inflation response to a cost-push shock in the baseline model with constant slope of the Phillips Curve κ but two attention regimes. The blue-dashed-dotted line shows the inflation dynamics in response to a cost-push shock in the model with constant attention but in which κ doubles between periods 1 and 10. The lower panel shows the two model responses to a demand shock.

Figure 39: Inflation dynamics with state-dependent κ



Notes: The red-solid line shows the baseline response to a large shock in the model with the attention threshold, and the blue-dashed line shows the response to a large shock in the model with constant attention but in which κ doubles between periods 1 and 10. The green-dashed-dotted line shows the response of the attention-threshold model to a small shock, and the red-dotted line the model response to a small shock when κ doubles between periods 1 and 10.