Self Reference Diagnostic Expectations*

Koutaroh Minami
Tokyo Keizai University
E-mail: kminami@tku.ac.jp

ABSTRACT

Diagnostic expectations incorporate the psychological heuristic that people excessively inflate the probability of events which are typical or representative of the newly observed information. It accounts for overreaction, systematic forecast errors, and revisions. However, the overestimated states are determined by the comparison between past and current rational distributions. It makes current distorted beliefs independent of past and future distorted beliefs because past distorted beliefs are not used to form current beliefs. This is inconsistent with the empirical literature on experience effects such as the influence of lifetime experiences on belief formation. This paper aims to develop a self-reference diagnostic expectations model in which current distorted beliefs are formed from past distorted beliefs. This model suggests that the former model may overestimate the strength of the psychological heuristic. This paper also conducts simulations to evaluate the self-reference model and compares it with the previous model. I find that the impact of the psychological heuristic is weak and analysts refer to more recent beliefs in the self-reference model. Still, analysts significantly overreact to information. The self-reference model has lower Euclidean distances to realized data, indicating a better fit. Overall, the findings suggest that stock market analysts exhibit overreaction tendency but they update their beliefs based on their own distorted beliefs.

Key words: Diagnostic expectations, Self-reference, Overreaction, Representativeness heuristic

JEL Classification: D84, G41

Received: September 17, 2025 Revised: April 7, 2026 Accepted: May 12, 2026

*I am grateful for helpful comments from Takashi Misumi, Hisashi Nakamura, Yukihiro Yasuda, Takayuki Tsuruga, Syrine Gacem, Jean-Paul L'Huillier, Jinill Kim and participants of the seminars at Hitotsubashi University. I also thank the anonymous reviewers for their helpful comments. This paper was presented at Nippon Finance Association 33rd Annual Conference and European Financial Management Association 2025 Annual Conference. The former title was "Dynamically-Consistent Diagnostic Expectations." All errors are mine.

1. Introduction

It is widely known that many types of agents in the stock market overreact to newly observed information. La Porta (1996) shows that stock market analysts excessively expect firms' future growth. Greenwood and Shleifer (2014) show that investors' expectations tend to be extrapolative. Importantly, such overreactions are followed by predictable forecast errors. For example, Gennaioli and Shleifer (2018) show overreaction and subsequent predictable forecast errors in analysts' forecasts. Bordalo et al. (2018) show that credit spreads overreact to news and entail predictable reversals.

Rational expectations models do not account for overreaction and the resulting predictable forecast errors so that psychology-based models are proposed. A prominent example is the "representativeness heuristic." It is the tendency that people overweight the probability of an event when it is representative of characteristics to its parent population (Kahneman and Tversky, 1972).

Bordalo et al. (2018) and Bordalo et al. (2019) propose diagnostic expectations (hereafter DE) which are based on the representativeness heuristic. DE is a psychologically founded model and there is an expanding literature.

The intuition behind DE is as follows. When new information arrives, agents compare the past rational belief with the current updated belief.[1] If some events are more likely to occur under the current updated belief, agents overinflate the probabilities of such events. For example, when agents observe positive signal about productivity, then the probability of a high-productivity state increases relative to that of a low-productivity state. As a result, agents overestimate the likelihood of high productivity and underestimate that of the low-productivity state. Agents' ultimate beliefs are distorted.

DE successfully explains the overreaction and subsequent forecast errors, but it makes current distorted beliefs independent of past distorted beliefs. Because past rational beliefs are used to form the current distorted beliefs, DE always starts from rational beliefs which no agents actually hold. In that sense, DE can be seen as a one-shot deviation from rational expectations. I refer to this type of DE as the one-shot model or one-shot DE.

The direction of overreaction to newly observed information is guaranteed to coincide with that implied by rational expectations by assuming that the current distorted beliefs are independent of past and future beliefs. In the one-shot DE, past distorted beliefs and subjective surprises defined relative to those beliefs do not affect belief updating. However, this implication is inconsistent with the empirical literature on experience effects, such as the influence of lifetime experiences on belief formation and risk-taking attitudes (Malmendier, 2021). To apply diagnostic expectations across several fields, models that account for these effects are needed.

The purpose of this paper is to develop a diagnostic expectations model in which current distorted beliefs are formed from past distorted beliefs. With this model, this paper extends the DE literature and derives new implications for agents' belief formation. Existing studies assume that the reference belief underlying DE is a past rational belief. In contrast, I replace this assumption by allowing agents to refer to their own past distorted beliefs. To distinguish it from the one-shot model, I refer to this model as the self-reference model. I also conduct the simulated method of moments (SMM) to estimate the model parameters and compare this model and the one-shot model.

The self-reference model delivers several important results. First, it suggests that the strength of the representativeness may be overestimated in the one-shot model. In the DE framework, overreaction arises from two distinct sources: agents' perceived surprise and the distorted updating rule induced by representativeness. In the one-shot DE model, the magnitude of perceived surprise is not large because it is always computed by past rational expectations. Under rational expectations, surprises are driven solely by the realization of the error term, which makes it a reasonable magnitude. As a consequence, overreaction is only driven by the representativeness heuristic, making the strength parameter large.

In contrast, in the self-reference DE model, agents' past beliefs are already distorted, and perceived surprise tends to be large as a natural consequence of prior overreaction. For example, if investors observed positive signal in the previous period, they overestimated the probability of high-productivity states, but current realized signal is typically lower than expected values, which makes perceived surprise large. When perceived surprise is large, overreaction can arise even with a relatively weak representativeness heuristic. I derive a sufficient condition under which the self-reference DE model is stable, which limits the extent of representativeness-driven distortions.

Second, simulations show that there is a significant belief distortion by representativeness even in the self-reference model. The simulations follow the methodology of Bordalo et al. (2019) and include the rational expectations case where there are no effects of representativeness. I find that the strength of representativeness is significantly positive, rejecting the rational expectations model. Although the self-reference model allows for larger perceived surprises and does not require a strong representativeness, the belief updating rule remains distorted.

Thirdly, I compare the self-reference and one-shot DEs and find that the self-reference model provides a better fit to the realized data. Consistent with the theoretical mechanism, the estimated strength of the representativeness in the self-reference model is lower than in the one-shot model. The other important parameter in DE is the reference period. It captures the lag of the reference distribution used in belief formation. The results show that the self-reference model refers to more recent beliefs than the one-shot model.

As a robustness check, I test the null hypothesis that the data are generated by the estimated one-shot model by using bootstrap simulations. Specifically, I simulate time-series data based on the estimated parameters of the one-shot model and then re-estimate the optimal parameters for each DE model. I compute the Euclidean distance between the moments of the optimal parameters and those of bootstrap-generated data for both models. Repeating this procedure across independent bootstrap samples, I obtain the frequency with which the self-reference model fits the data better than the one-shot model under the null hypothesis. I interpret this frequency as the bootstrap p-value. The results confirm that the self-reference model provides a better fit and that the difference is statistically significant.

The main contribution of this paper is to propose a DE model in which belief distortions are formed from own past distorted beliefs. In the model, agents update their beliefs based on their own distorted beliefs, and their update rule is distorted by the representativeness heuristic. In the one-shot DE, agents update their beliefs based on past rational expectations, which enables us to understand the mechanism of agents' overreaction concisely (Bordalo et al., 2018; Bordalo et al., 2019), but it offers limited insight into the dynamic evolution of agents' beliefs. By developing a self-reference DE model, this paper extends the applicability of DE to a broader range of settings. Moreover, because belief distortions propagate over time in this framework, stability is not guaranteed. This paper derives a sufficient condition under which the self-reference DE model is stable. I also provide evidence that the one-shot model tends to overestimate the strength of the representativeness.

This paper also offers new insights into analysts' belief formation. In the simulation, I used the analysts' forecasts of EPS. Both the self-reference and one-shot models reject the null hypothesis that analysts have rational expectations. Importantly, the self-reference model fits the data better than the one-shot model. These results suggest that analysts are influenced by the representativeness heuristic, yet their beliefs are formed with their own past beliefs. Analysts overreact to newly observed information, but they do not discard their prior beliefs. In this sense, the paper sheds light on belief formation: analysts rely on their experiences and beliefs, but their belief updating remains systematically distorted.

This paper is organized as follows. Section 2 reviews the existing literature on DE. Section 3 describes the one-shot and the self-reference DE. Section 4 describes the simulation methodology, and section 5 presents the result of baseline simulation and robustness check. Section 6 concludes.

2. Prior Literature

If agents have rational expectations and update their beliefs according to Bayes' rule, they do not over- or underreact to information. However, a large body of empirical literature shows systematic overreaction by agents. For example, in the stock market, La Porta (1996) and Gennaioli and Shleifer (2018) show the overreaction by analysts and Greenwood and Shleifer (2014) show that many types of U.S. investors exhibit extrapolative characteristics. d'Arienzo (2020) shows that analysts overreact in long-term bond returns. Augenblick et al. (2021) find that almost all participants update their beliefs in the right direction, but there is heterogeneity in how much they revise.

To account for them, several studies have employed psychological foundations. Barberis et al. (1998) propose an investor sentiment-driven model. Barberis and Shleifer (2003) examine style investing where investors categorize assets into broad classes and allocate their funds. Gennaioli and Shleifer (2010) incorporate the representativeness heuristic, the tendency for individuals to overweight the probability of an event when it is representative of the characteristics of its parent population (Kahneman and Tversky, 1972).

Based on the psychological background, Bordalo et al. (2018) propose the DE. In this framework, expectations are formed and distorted by the representativeness heuristic. People overestimate the probability of events that are more likely to occur under the parent population. For example, when investors receive the positive signal, rational investors update their beliefs according to Bayes' rule, but diagnostic investors, who have DE, overestimate the high productivity state because such state is representative of positive signals. As a result, diagnostic investors overreact to positive signals and become more optimistic.

DE shares some characteristics with extrapolative expectations, but there is a clear difference. Under extrapolative expectations, agents assume that expected future changes such as returns or prices are determined by the weighted average of past changes (Barberis et al., 2018; Glaeser and Nathanson, 2017; Barberis et al., 2015). Extrapolative expectations are backward-looking, whereas DE are forward-looking because DE compares prior and posterior probabilities and overestimates the probability of representative states in the posterior probability distribution. Therefore, it is a context-dependent framework. The state to which agents overreact and the magnitude of overreaction depend on prior beliefs and newly observed information.

DE accounts for investors' tendency to overreact and for the presence of predictable forecast errors. Bordalo et al. (2018) show, under DE, that credit spreads are excessively volatile, overreact to news, and exhibit predictable reversals. Bordalo et al. (2021a) show that declines in productivity growth lead to excessively large increases in credit spreads through the DE mechanism. Bordalo et al. (2022) demonstrate that overreaction driven by DE generates predictable boom-bust cycles.

Bordalo et al. (2019) apply DE to stock return analysis and show that DE can explain the systematic forecast errors and revisions of analysts' forecasts. Bordalo et al. (2020b) show that systematic overreaction generates the price reversals and excess stock market volatility. Bordalo et al. (2021b) show that price overreaction leads to endogenous bubbles and crashes.

DE have also been used to explain consumption behavior. L'Huillier et al. (2024) show that consumption also overreacts to supply shocks. Bianchi et al. (2021a) explain persistent and hump-shaped boom-bust cycles in consumption within a DE framework. Bianchi et al. (2021b) show that consumption becomes time-inconsistent when the reference point is not recent.

Krishnamurthy and Li (2020) introduce DE into frictional financial intermediation models and show that banks with DE have higher risk tolerance, decrease risk premia and increase credit before crisis. Maxted (2024) shows that disappointment after a boom driven by excessive optimism under DE makes banks tighten their lending, leading to crisis.

In the current DE model, current beliefs are distorted by the representativeness with which past and current rational expectations are compared. Because current distorted beliefs are not updated from past distorted beliefs, they are independent of past and future beliefs. This mechanism ensures that overreaction to newly observed information coincides with the direction implied by rational expectations. In contrast, agents form their beliefs anew in each period, rendering current beliefs independent of past experiences and beliefs.

A large body of empirical evidence shows that individuals' lifetime experiences have a substantial impact on belief formation and risk taking in finance (Malmendier and Nagel, 2011; Malmendier, 2021). Bianchi et al. (2022) show that individuals overweight private information relative to public information and that even professionals exhibit similar behavior. Bouwman and Malmendier (2015) argue that experience effects are present even at the institutional level. Malmendier et al. (2020) analyze an asset pricing model in which agents form expectations by overweighting the dividends observed during their lifetime. Malmendier and Wellsjo (2024) show that past inflation experiences affect the homeownership decision. Because the current DE model does not use past distorted beliefs to form current distorted beliefs, it cannot account for these experience effects.

Bianchi et al. (2024a) show that, under the current DE model, the law of iterated expectations holds if and only if the past beliefs—the comparison distribution—are one period ahead. They study a multi-period consumption-savings model and demonstrate that when past beliefs are distant, the model generates boom-bust cycles.

This paper fills an important gap in the DE literature. I develop a self-reference DE and compare it with a one-shot deviation model from rational expectations. I consider how agents' beliefs are updated based on their own past distorted beliefs. This setting avoids situations in which agents effectively forget their distorted beliefs each period.

The most closely related paper to this analysis is Bianchi et al. (2024b). They introduce uncertainty effects into the DE framework. In the baseline DE models of Bordalo et al. (2018) and Bordalo et al. (2019), subjective uncertainty is assumed to be stationary and is not revised, allowing the analysis to focus on the impact of psychological heuristics on expected values. In contrast, Bianchi et al. (2024b) propose a smooth DE model in which subjective volatility is also updated to examine the uncertainty effects. This paper takes a different approach by analyzing belief updating when reference beliefs are past distorted beliefs. This framework allows us to examine directly how self-referencing shapes the dynamics of agents' expected values.

DE are also closely related to the concept of selective memory. When decision makers observe new information, states or events whose probability increases the most come to mind first and they are oversampled in mind because of the representativeness heuristic. In the selective memory literature, individuals recall the memories from their internal database, and the order of recalling or the weights typically decline with temporal distance. While the standard setting assumes that observing events or information are sources of current beliefs (Bordalo et al. 2020a; Nagel and Xu 2022), my results suggest that agents' own past beliefs can also serve as salient reference points in belief formation.

3. Model: Self-Reference Diagnostic Expectations

Assume that the fundamental of firm i follows an AR(1) process

$$f_{i,t} = af_{i,t-1} + \eta_{i,t} \tag{1}$$

where $a \in [0,1]$ and $\eta_{i,t} \sim N\left(0, \sigma_\eta^2\right)$ is an i.i.d. normally distributed shock. It is assumed that investors cannot observe $f_{i,t}$ directly. Instead, they observe the signal $x_{i,t}$ which is given by

$$x_{i,t} = bx_{i,t-1} + f_{i,t} + \epsilon_{i,t} \tag{2}$$

where $b \in [0,1]$ and $\epsilon_{i,t} \sim N\left(0, \sigma_\epsilon^2\right)$ is an i.i.d. normally distributed shock. Imposing $b \leq a$ ensures stationarity. Bordalo et al. (2019) interpret this signal as the natural logarithm of a firm's earnings per share (EPS).

Rational investors update their beliefs using the Kalman filter to infer the current fundamental based on the information set after observing signal $x_{i,t}$. The expected value of $f_{i,t}$ is updated as follows:

$$\mathbb{E}\left[f_{i,t}\middle|x_{i,t}\right] = \hat{f}^{RE}_{i,t} = a\hat{f}^{RE}_{i,t-1} + K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{RE}_{i,t-1}\right) \tag{3}$$

where $\mathbb{E}$ is an expectation operator for rational expectations, $\hat{f}^{RE}_{i,t}$ is the expected value of $f_{i,t}$ under rational expectations, and $K \equiv \left(a^2\sigma_f^2 + \sigma_\eta^2\right)/\left(a^2\sigma_f^2 + \sigma_\eta^2 + \sigma_\epsilon^2\right)$ is the signal-to-noise ratio, also called Kalman gain.[2] This indicates the extent to which signal fluctuations are attributed to the fundamental.

The Kalman filter is known to be the optimum recursive estimator for linear Gaussian state-space models, in which both the state and observation equations are linear, and the error terms follow a Gaussian distribution. This optimality comes from the fact that the Kalman filter is the minimum mean square error estimator.

DE are a framework in which agents' beliefs are distorted by the representativeness heuristic. Representativeness heuristic is a human's tendency to overestimate the probability of events that are representative or typical of a parent class (Kahneman and Tversky, 1972).

Gennaioli and Shleifer (2010) propose the measurement of the representativeness. Building on this approach, Bordalo et al. (2016) assume that probability judgements are formed using the representativeness-distorted density

$$h^\theta(T = \tau \mid G) = h(T = \tau \mid G)\left(\frac{h(T = \tau \mid G)}{h(T = \tau \mid -G)}\right)^\theta Z \tag{4}$$

where $h(T = \tau|G)$ is the distribution of a variable T in a group G, $-G$ is a comparison (reference) group, $\theta \geq 0$, and Z is a constant ensuring that the distorted density $h^\theta(T = \tau \mid G)$ integrates to 1. θ controls the impact of the representativeness. As θ increases, the distortion of the probability increases; when $\theta = 0$, the distribution is not distorted from the true distribution $h(T = \tau \mid G)$ and it corresponds to the rational benchmark.

Bordalo et al. (2019) apply this distorted probability judgment and propose the DE. They use the rational expectations $N\left(\hat{f}^{RE}_{i,t}, \sigma_f^2\right)$ after observing the signal $x_{i,t}$ as a target distribution $h(T = \tau \mid G)$ and the rational expectations $N\left(a\hat{f}^{RE}_{i,t-1}, \sigma_f^2\right)$ in the counterfactual case in which the signal surprise is zero $x_{i,t} = bx_{i,t-1} + a\hat{f}^{RE}_{i,t-1}$ as a reference distribution $h(T = \tau \mid -G)$. Substituting these distributions into equation (4), the DE belief is obtained through a distorted Kalman filter.

$$\mathbb{E}^{OS}\left[f_{i,t} \mid x_{i,t}\right] = \hat{f}^{OS}_{i,t} = a\hat{f}^{RE}_{i,t-1} + K(1+\theta)\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{RE}_{i,t-1}\right) \tag{5}$$

where $\mathbb{E}^{OS}$ is an expectation operator for DE proposed by Bordalo et al. (2019) and $\hat{f}^{OS}_{i,t}$ is the expected value of $f_{i,t}$ under this DE.

Note that investors are assumed to know the state-space model characterized by equation (1) and (2), do not observe the fundamental directly, and observe signal instead. Investors with DE systematically overreact to signals and do not improve their learning process over time.

Under DE, investors compare the current conditional rational distribution with the past one and overestimate(underestimate) the probability of events that become more(less) likely to occur after observing new information. Because the reference distribution is a past conditional rational distribution, past distorted beliefs (past DE) do not affect current beliefs. Instead, belief distortions arise from the comparison between past and current rational beliefs. From this perspective, we can interpret DE as a one-shot deviation strategy from rational expectations.

This property exposes DE to several important critiques. First, the interpretation of rational prior beliefs is problematic. There is only an agent with diagnostic expectations in this economy. It means that agents' beliefs at time $t-1$ are distorted and given by $\hat{f}^{OS}_{i,t-1}$, not by the rational expectations $\hat{f}^{RE}_{i,t-1}$. What is $\hat{f}^{RE}_{t-1}$, which no agents actually hold at time $t-1$? How do diagnostic agents get that? This ambiguity poses a challenge to the one-shot DE.

Second, the assumption that current distorted beliefs are independent of past distorted beliefs is not aligned with empirical evidence on experience effects. A large body of empirical research shows that personal experiences play a crucial role in shaping belief formation and risk-taking behavior. In the one-shot DE framework, current distorted beliefs are constructed from past and current rational beliefs rather than from agents' own beliefs. Consequently, agents' personal experiences—embodied in their past distorted beliefs—do not influence current belief formation.

To address these issues and apply DE in various analyses, I use past distorted beliefs as the reference distribution and consider the distorted belief updating based on distorted beliefs.

Suppose that agents' past beliefs are not rational.

$$\mathbb{E}^{RD}\left[f_{i,t-1} \mid x_{i,t-1}\right] = \hat{f}^{SR}_{t-1} \tag{6}$$

where $\mathbb{E}^{RD}$ is an expectation operator which differs from both rational expectations and the one-shot DE, and $\hat{f}^{SR}_{t-1}$ is the expected value of $f_{i,t-1}$ under this expectation.

The forecast error of the signal and its expected value based on this distorted past belief are defined as follows.

$$v^{RD}_{i,t} = x_{i,t} - \mathbb{E}^{RD}\left[x_{i,t} \mid x_{i,t-1}\right]$$
$$\mathbb{E}^{RD}\left[v^{RD}_{i,t} \mid x_{i,t-1}\right] = \mathbb{E}^{RD}\left[x_{i,t} \mid x_{i,t-1}\right] - \mathbb{E}^{RD}\left[\mathbb{E}^{RD}\left[x_{i,t} \mid x_{i,t-1}\right] \mid x_{i,t-1}\right] = 0$$

where $\mathbb{E}^{RD}$ is an expectation operator based on equation (6). If investors' prior beliefs are given by equation (6) and they update their beliefs using the Kalman filter, their posterior beliefs are given by

$$\mathbb{E}^{RD}\left[f_{i,t} \mid x_{i,t}\right] = a\hat{f}^{SR}_{i,t-1} + K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{SR}_{i,t-1}\right) \equiv \hat{f}^{RD}_{i,t} \tag{7}$$

Notice that the Kalman filter with distorted beliefs is identical to that under rational beliefs. Because only the prior mean is altered, the subjective variance of fundamental is still assumed to be stationary. Since the Kalman gain depends on variance terms, it remains unchanged.

Next, I consider the DE when prior beliefs are given by equation (6). The distorted density shown in equation (4) requires a target distribution and a comparison distribution. As the target distribution, I use the current conditional distribution, $N\left(\hat{f}^{RD}_{i,t}, \sigma^2_f\right)$. As the reference distribution, I use $N\left(a\hat{f}^{SR}_{i,t-1}, \sigma^2_f\right)$, corresponding to the conditional distribution under distorted prior beliefs in the case of zero signal surprise. Throughout, $\hat{f}^{SR}_{i,t}$ denotes the expected value of the fundamental conditional on the current signal and distorted priors.

Proposition 1 (Self-Reference Diagnostic Expectations). Under self-reference diagnostic expectations, the expected value of the fundamental is updated as follows.

$$\hat{f}^{SR}_{i,t} = a\hat{f}^{SR}_{i,t-1} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{SR}_{i,t-1}\right) \tag{8}$$

Proof is shown in Appendix A.

Table 1 summarizes agents' beliefs about the fundamental as a function of prior beliefs and the updating method. Top-left corresponds to rational expectations, in which prior beliefs are rational and investors update their beliefs using the Kalman filter. The posterior belief in this case is

Table 1. Beliefs by prior beliefs and updating rule

prior\update	Rational	Diagnostic
Rational ($\hat{f}^{RE}_{i,t-1}$)	[Rational expectations] $\hat{f}^{RE}_{i,t} = a\hat{f}^{RE}_{i,t-1} + K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{RE}_{i,t-1}\right)$	[One-shot diagnostic expectations] $\hat{f}^{OS}_{i,t} = a\hat{f}^{RE}_{i,t-1} + K(1+\theta)\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{RE}_{i,t-1}\right)$
Distorted ($\hat{f}^{SR}_{i,t-1}$)	$\hat{f}^{RD}_{i,t} = a\hat{f}^{SR}_{i,t-1} + K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{SR}_{i,t-1}\right)$	[Self-reference diagnostic expectations] $\hat{f}^{SR}_{i,t} = a\hat{f}^{SR}_{i,t-1} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{SR}_{i,t-1}\right)$

Note: This table summarizes beliefs as a function of prior beliefs and the updating rule. Rows indicate prior beliefs, and columns indicate the updating rule.

given by equation (3). Top-right corresponds to the DE proposed by Bordalo et al. (2019), as shown in equation (5). Because prior beliefs are assumed to be rational in this framework, the analysis focuses on comparing update rules and isolating the effect of distorted updating.

In contrast, in the bottom row, prior beliefs are not rational. In the bottom-left, investors update their beliefs using the Kalman filter without correcting prior beliefs, shown in equation (7). Bottom-right corresponds to the DE with distorted prior beliefs, in which investors update their beliefs using the distorted Kalman filter characterized by equation (8).

Compared with the top-right one-shot DE, the bottom-right self-reference DE is derived from past distorted beliefs. The signal surprise for self-reference diagnostic investors is given by $x_{i,t} - \left(bx_{i,t-1} + a\hat{f}^{SR}_{i,t-1}\right)$. If the surprise in the previous period was positive and $\hat{f}^{SR}_{i,t-1}$ increased, the current surprise is more likely to be negative. Moreover, this surprise is incorporated into beliefs through the amplified gain $(1+\theta)K$, rather than through the standard Kalman gain K. As a result, the self-reference DE qualitatively generates larger signal surprise and overreacts to signals than rational expectations and one-shot DE.

In the one-shot DE, belief overreaction is solely driven by θ, which captures the strength of the representativeness. In contrast, in the self-reference DE, overreaction can arise from θ and endogenously inflated surprises. This implies that estimated θ under the one-shot DE may be upward biased.

Bordalo et al. (2019) consider the comparison period to be not only one period prior but also $s \geq 1$ periods prior.

Since the expected value under one-shot DE can be written as $\mathbb{E}^{OS}\left[f_{i,t}\right] = \mathbb{E}\left[f_{i,t} \mid x_{i,t}\right] + \theta\left(\mathbb{E}\left[f_{i,t} \mid x_{i,t}\right] - \mathbb{E}\left[f_{i,t} \mid x_{i,t-s}\right]\right)$, the one-shot DE belief is given by

$$
\begin{aligned}
\hat{f}^{OS}_{i,t} &= \hat{f}^{RE}_{i,t} + \theta\left(\hat{f}^{RE}_{i,t} - a^s\hat{f}^{RE}_{i,t-s}\right) \\
&= a\hat{f}^{RE}_{i,t-1} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{RE}_{i,t-1}\right) + \theta\left(\hat{f}^{RE}_{i,t} - a^s\hat{f}^{RE}_{i,t-s}\right)
\end{aligned}
\tag{9}
$$

Next, consider the self-reference DE in which the reference distribution is also s periods prior. In this case, the target distribution is the current conditional distribution updated from the prior distorted distribution, denoted by $N\left(\hat{f}^{RD}_{i,t}, \sigma^2_f\right)$. The reference distribution is the distorted belief based on $t-s$ periods given by $N\left(a^s\hat{f}^{SR}_{i,t-s}, \sigma^2_f\right)$. The self-reference DE is given by

$$
\begin{aligned}
\hat{f}^{SR}_{i,t} &= \hat{f}^{RD}_{i,t} + \theta\left(\hat{f}^{RD}_{i,t} - a^s\hat{f}^{SR}_{i,t-s}\right) \\
&= (1+\theta)(1-K)a\hat{f}^{SR}_{i,t-1} - \theta a^s\hat{f}^{SR}_{i,t-s} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1}\right)
\end{aligned}
\tag{10}
$$

The derivation of equation (10) is provided in Appendix B.

The one-shot DE is a model which always deviates from rational expectations. Because the rational expectations updating rule in equation (3) is stable, the one-shot DE updating rule in equation (9) is also stable and does not diverge. In contrast, it is not clear whether equation (10) is stable.

By equation (2) the third term is rewritten as $x_{i,t} - bx_{i,t-1} = f_{i,t} + \epsilon_{i,t}$. Given that $a \in [0, 1]$ by assumption, the third term is stable and non-explosive. Therefore, the stability of equation (10) depends on the first and second terms.

Proposition 2 (Sufficient condition for stability). The self-reference diagnostic expectations model in which the reference distribution is based on $t - s$ periods evolves according to equation (10). A sufficient condition for this process to be stable in the sense that the process does not diverge is given by

$$\left|(1+\theta)(1-K)a\right| + \left|-\theta a^s\right| < 1 \tag{11}$$

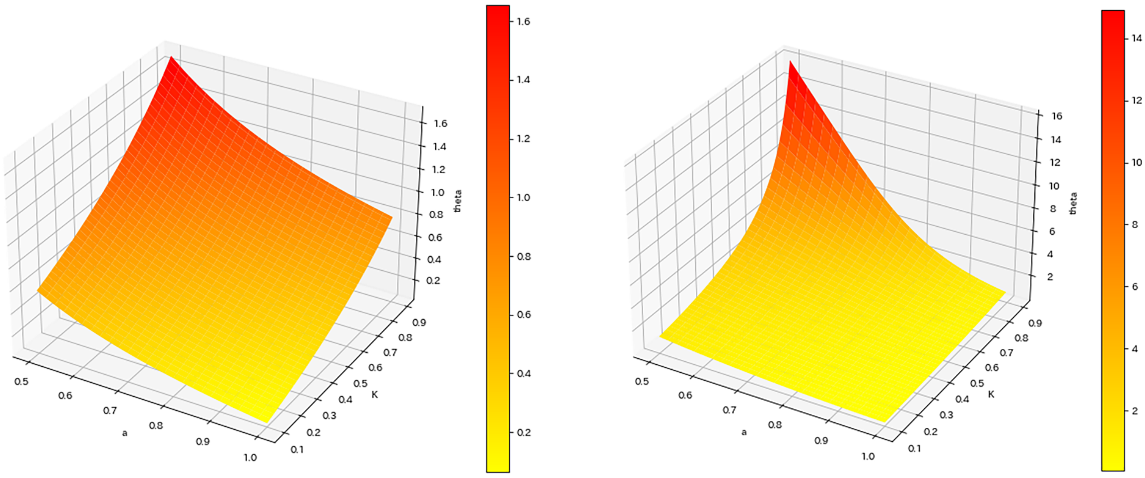
Proof. According to the Schur stability condition, the linear difference equation $y_t = \sum_{j=1}^{s} c_j y_{s-j}$ is stable if $\sum_{j=1}^{s} \left|c_j\right| < 1$. Setting $y_t = \hat{f}_{i,t}^{SR}$, $c_1 = (1+\theta)(1-K)a$, $c_s = -\theta a^s$, $c_k = 0$ for all $k \neq 1, s$, equation (11) is obtained.

Given $a \in [0, 1]$, $0 \leq K \leq 1$ and $\theta \geq 0$, the coefficient on the first term in equation (10) is always positive and the coefficient on the second term is always negative. Rearranging equation (11),

$$(1+\theta)(1-K)a + \theta a^s < 1$$

$$\theta < \frac{1-(1-K)a}{(1-K)a + a^s} \equiv \bar{\theta}(a, K, s) \tag{12}$$

The threshold $\bar{\theta}(K, a, s)$ can be interpreted as the upper bound for θ under which the self-reference DE remains stable. Taking partial derivatives, I obtain $\partial\bar{\theta}/\partial a \leq 0$, $\partial\bar{\theta}/\partial K \geq 0$, $\partial\bar{\theta}/\partial s \geq 0$. This implies that there is more room for belief distortion when the fundamental is less persistent, signal is more informative, or the reference period of comparison distribution is more distant. A lower value of a indicates that the fundamental is more influenced by the shock η_t, which is only indirectly inferred via signal surprises. This suggests that agents try to extract more information from surprises. Similarly, when signal is more informative, surprises contain more



(a) $s = 1$. (b) $s = 8$.

Figure 1. $\overline{\theta}$ in (a, K, θ) space

Note: These figures plot $\overline{\theta}$ as defined in equation (12). The left panel corresponds to $s = 1$, and the right panel corresponds to $s = 8$.

information about fundamental, leading to stronger overreaction. Finally, a more distant reference period makes current information more valuable, thereby amplifying overreaction. Figure 1 plots $\overline{\theta}$ in (a, K, θ) space.

It is not obvious whether self-reference DE satisfies a form of time consistency in the sense of a recursive expectation relation. Bianchi et al. (2024a) show that, for one-shot DE, a tower-like relation holds at the level of conditional means only when the reference period s is 1. Here, I examine whether a similar mean-level identity holds under self-reference DE. Consider the reference periods $s \geq 1$ and any integer $n \geq 1$.

Lemma 1 (A mean-level identity under self-reference DE). Under the self-reference DE, a tower-like relation at the level of conditional means holds if and only if the reference period is the immediate past, $s = 1$.

Proof is shown in Appendix C.

Note that this result does not imply that the law of iterated expectations holds for the self-reference DE in the usual sense. Although a tower-like relation holds at the level of conditional means when $s = 1$, it generally fails for higher moments if $\theta > 0$. This difference reflects the fact that the self-reference DE is a distorted belief updating rule.

When the reference period is not the most recent period, the target distribution $N\left(\hat{f}^{RD}_{i,t}, \sigma^2_f\right)$ incorporates not only the surprise realized at period t but also the surprises accumulated up to period $t-1$. In contrast, the reference distribution $N\left(a^s \hat{f}^{SR}_{i,t-s}, \sigma^2_f\right)$ is formed at period $t-s$. As a result, surprises from period $t-s+1$ onward are incorporated into beliefs only at time t, rather than being updated sequentially. Because belief updating skips intermediate revision, the tower-like relation at the level of conditional means fails when $s > 1$.

This result is more intuitive when compared to rational expectations. Under rational expectations, the prior distribution is implicitly assumed to be one in period $t-1$ so that surprises are incorporated sequentially. The self-reference DE yields the same mean-level identity for the same reason when $s = 1$.

More broadly, the self-reference DE can be interpreted as a framework in which agents overestimate the precision of signals due to the representativeness heuristic. Agents overreact to surprises because the effective Kalman gain is inflated. Under rational expectations, the dispersion of beliefs is governed by the Kalman gain K, whereas under self-reference DE it is governed by $(1+\theta)K$, implying that beliefs fluctuate more than the underlying signals. Because beliefs are updated from the agents' own past distorted beliefs in each period, they never converge to rational expectations. Consequently, agents continue to overreact to subjective surprises over time.

4. Simulation

In this section, I estimate the parameters of the self-reference DE and compare them with those obtained under the one-shot DE.

The estimation strategy follows the simulation procedure of Bordalo et al. (2019). I estimate the diagnostic parameter θ together with other macroeconomic parameters. Bordalo et al. (2019) use the natural logarithm of earnings per share (EPS) as the signal ($x_{i,t}$). For expectation data, I use data on analysts' expectations from Refinitiv Eikon. I obtain mean analysts' forecasts for EPS from 1982 to 2023. The sample consists of firms listed on major U.S. stock exchanges (NYSE, AMEX, and NASDAQ).

Bordalo et al. (2019) obtain data on analysts' forecasts from IBES and their sample period is between 1981 and 2016. To check the differences in the data, this paper first conducts the test of Coibion and Gorodnichenko (2015) (CG test) to confirm that the analysts' expectations in this study also exhibit overreaction tendency.

$$x_{i,t+h} - x_{i,t} - LTG_{i,t,h} = \alpha + \gamma\left(LTG_{i,t,h} - LTG_{i,t-k,h}\right) + e_{i,t+h} + FE_t \tag{13}$$

where $LTG_{i,t,h}$ is the expected growth rate of the signal (EPS) over h periods, $LTG_{i,t,h} = E\left[x_{i,t+h} - x_{i,t} | x_{i,t}\right]$, and FE_t is a year fixed effect.

Compared to Bordalo et al. (2019)'s dataset, which offers annualized expected growth, Refinitiv offers mean analysts' forecasts $E\left[x_{i,t+h}\right]$. $LTG_{i,t,h}$ is computed by taking the logarithm of this value and subtracting the current realized value $x_{i,t}$. Annualized rate is obtained by dividing by h. Note that in equation (13), the expected growth rate and the realized growth rate are matched over the same h-period horizon.

If investors overreact to surprises, forecast revisions $LTG_{i,t,h} - LTG_{i,t-k,h}$ become excessive. As a result, forecast errors $x_{i,t+h} - x_{i,t} - LTG_{i,t,h}$ are, on average, negative. Therefore, if investors overreact to signals, γ would be negative.

Table 2 shows the number of samples for the CG test based on forecast horizon h and revision horizon k. Both h and k are measured in years. $(k, h) = (1, 1)$ has the most samples and the sample size decreases as h increases. In particular, the sample size decreases sharply for longer forecast length ($h > 3$).

Table 3 reports the results of the CG test. Standard errors are shown in parentheses and clustered at the firm level. All combinations of (k, h) have negative γ and they are statistically

Table 2. Sample sizes for the Coibion and Gorodnichenko (2015) test

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
$k = 1$	31480	29289	20470	7066	4624
$k = 2$	28797	26958	18573	5900	3740
$k = 3$	26603	25002	16971	5121	3171
$k = 4$	24644	23238	15569	4473	2719
$k = 5$	22831	21586	14272	3903	2317

Note: This table reports the number of observations used in the CG tests. Row k indicates the revision horizon in years, and column h indicates the forecast horizon in years.

Table 3. Results of the Coibion and Gorodnichenko (2015) test

	$h = 1$	$h = 2$	$h = 3$	$h = 4$	$h = 5$
$k = 1$	−0.754***	−0.601***	−0.597***	−0.685***	−0.677***
	(0.023)	(0.014)	(0.015)	(0.029)	(0.032)
$k = 2$	−0.781***	−0.594***	−0.62***	−0.706***	−0.678***
	(0.022)	(0.014)	(0.015)	(0.03)	(0.039)
$k = 3$	−0.745***	−0.599***	−0.625***	−0.706***	−0.664***
	(0.021)	(0.015)	(0.015)	(0.033)	(0.042)
$k = 4$	−0.779***	−0.597***	−0.625***	−0.671***	−0.682***
	(0.021)	(0.016)	(0.017)	(0.035)	(0.052)
$k = 5$	−0.795***	−0.612***	−0.646***	−0.68***	−0.628***
	(0.022)	(0.017)	(0.018)	(0.039)	(0.046)

Note: This table shows the regression coefficients of forecast errors on forecast revisions. Standard errors shown in parentheses are clustered at the firm level. Row k indicates the revision horizon in years, and column h indicates the forecast horizon in years. *** indicates statistical significance at the 1% level.

significant. In particular, estimated coefficients for $h = 1$ are larger negative values. Compared with Bordalo et al. (2019), results share the sign of coefficients and exhibit larger magnitudes. These findings indicate that analysts in our data also exhibit overreaction to signals.

Next, I run the SMM to estimate the parameters $(a, b, \sigma_\eta, \sigma_\epsilon, \theta, s)$. The parameters $(a, b, \sigma_\eta, \sigma_\epsilon)$ are macroeconomic parameters governing the fundamental and signal processes. The parameter θ captures the strength of the representativeness heuristic. s is a sluggishness parameter that indicates the number of quarters used for the reference distribution.

I construct parameter combinations on a grid defined by $a, b \in [0, 1]$, $\sigma_\eta, \sigma_\epsilon \in [0, 0.5], \theta \in [0, 2]$, and $s \in \{1, \cdots, 20\}$ quarters. The step size is 0.025 for a, 0.1 for b, 0.05 for σ_η and σ_ϵ, 0.1 for θ.

For each parameter combination, I first simulate time series for the fundamental and the signal. I then compute the associated one-shot DE, $\hat{f}^{OS}_{i,t}$, and self-reference DE, $\hat{f}^{SR}_{i,t}$. Using these beliefs, I calculate the expected long-term growth of the signal under each DE model.

$$LTG^{OS}_{i,t,h} = \mathbb{E}^{OS}\left[x_{i,t+h} - x_{i,t} | x_{i,t}\right] = -\left(1 - b^h\right)x_{i,t} + a^h \frac{1 - (b/a)^h}{1 - (b/a)} \hat{f}^{OS}_{i,t} \tag{14}$$

$$LTG^{SR}_{i,t,h} = \mathbb{E}^{SR}\left[x_{i,t+h} - x_{i,t} | x_{i,t}\right] = -\left(1 - b^h\right)x_{i,t} + a^h \frac{1 - (b/a)^h}{1 - (b/a)} \hat{f}^{SR}_{i,t} \tag{15}$$

where $\mathbb{E}^{SR}$ is an expectation operator for self-reference DE.

Using $LTG^{d}_{i,t,h}$ where $d \in \{OS, SR\}$, I compute forecast errors and forecast revisions to run the CG test. Following Bordalo et al. (2019), I focus on $(k, h) = (1, 4)$ and $(3, 4)$ years. I denote the corresponding regression coefficients by $\hat{\gamma}_1$ and $\hat{\gamma}_3$, respectively. In addition, I compute the autocorrelation $\hat{\rho}_l = cov\left(x_{i,t}, x_{i,t-l}\right)/var\left(x_{i,t}\right)$ for lags $l = 1$ through 4 years. These moments form the coefficient vector.

$$v\left(a, b, \sigma_\eta, \sigma_\epsilon, \theta, s\right) = (\hat{\rho}_1, \hat{\rho}_2, \hat{\rho}_3, \hat{\rho}_4, \hat{\gamma}_1, \hat{\gamma}_3)$$

I repeat this procedure for all parameter combinations. The parameter estimates are obtained by selecting the combination that minimizes the Euclidean distance loss function

$$l(v) = \|v - \bar{v}\|$$

where $\bar{v}$ is the coefficient vector estimated from the realized data. It is referred to as the target

Table 4. Target coefficient values

$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	$\hat{\rho}_4$	$\hat{\gamma}_1$	$\hat{\gamma}_3$
0.8947	0.8225	0.802	0.7708	−0.685	−0.706

Note: $\hat{\rho}_l$ denotes the autocorrelation of pooled EPS at lag l=1,2,3,4 years. $\hat{\gamma}_h$ denotes the regression coefficient from the CG test for h=1,3 years. These values constitute the target vector used in the simulations.

vector in this paper and is reported in Table 4.

I conduct this simulation for 30 independent runs and obtain 30 sets of estimated parameters, following the previous literature. Using these combinations, I analyze the differences between the one-shot and self-reference DEs.

One might be concerned that self-reference DE ($\hat{f}^{SR}_{i,t}$) and rational expectations ($\hat{f}^{RE}_{i,t}$) are not distinguishable. However, there is an important distinction within the SMM framework. A comparison of the update rule for rational expectations in equation (3) and that for self-reference DE in equation (8) suggests that the two become observationally equivalent when $(1 + \theta)K \leq 1$. Yet the Kalman gain is computed from the variances of the error terms of the fundamental and signal. Increasing σ^2_η raises the Kalman gain, but it simultaneously makes the signal process more volatile. Since the target vector includes the autocorrelations of the signal, matching only the Kalman gain component worsens the overall fit to the data.

In contrast, within the self-reference DE framework, one can allow expectations to fluctuate while keeping the variances small. For this reason, the SMM procedure generates a clear distinction between the two models. They are qualitatively the same only when $\theta = 0$.

5. Results

5.1. Baseline simulations

I begin by focusing on the simulations of the one-shot DE, which is a replication of Bordalo et al. (2019). Table 5 shows the mean and standard deviation of the 30 estimated parameter sets. Panel (a) presents the results of one-shot DE. All parameters are significantly different from zero according to Student's t-test. The estimates indicate that the fundamental process is highly persistent, while the estimate of b is 0.31, implying that the signal is not persistent. Examining the variances of the error terms, I find that σ_ϵ is larger than σ_η. As implied by the Kalman gain, this indicates that the signal contains large noise and it is less informative. Compared with Bordalo et al. (2019), the simulations in this paper yield a higher value of a and a lower value of b. They also observe that σ_η is larger than σ_ϵ, suggesting that signals are informative. These are major differences from previous research.

In contrast to the macroeconomic parameters, the most important parameter is the representativeness heuristic. The estimated value is close to that in previous literature. The mean estimate of θ is 1.079, which is slightly larger than one in previous research. Because signal surprises are incorporated into investors' beliefs by $(1+\theta)K$, a value of $\theta = 1$ implies that posterior beliefs incorporate signal surprises twice as strongly as under rational expectations. Although $\theta = 0$ is included in the parameter combinations, it is not selected as the optimal value. This result indicates that investors are likely to overreact to signals according to DE. Finally, the estimated sluggishness parameter, which indicates the lag of reference distribution, is 6.311 quarters in this simulation, which is lower than previous research. Note that θ and s are also statistically different from zero according to Student's t-test.

Next, I turn to the simulation results for the self-reference DE. Table 5 Panel (b) shows that the fundamental is still persistent, while the mean estimate of b is 0.461, indicating that the signal is not persistent as well. Regarding the variances of the error terms, σ_ϵ is larger than σ_η, but the difference between these variances is smaller than in the one-shot model. The implied Kalman gain is 0.508, suggesting that the self-reference diagnostic model features a relatively informative signal structure.

Turning to the most important parameters, the mean estimate of θ is 0.727. Since θ is greater than 0, this result indicates that agents' belief updating rule is distorted by the representativeness heuristic. It also implies that the self-reference DE is more likely to fit the data than the rational expectations model, which corresponds to $\theta = 0$.

Table 5. Estimated parameters

(a) One-shot diagnostic expectations.

	a	b	σ_η	σ_ϵ	θ	s	K
mean	0.991	0.31	0.16	0.35	1.079	6.311	0.36
std. dev	0.012	0.208	0.127	0.123	0.611	7.018	

(b) Self-reference diagnostic expectations.

	a	b	σ_η	σ_ϵ	θ	s	K
mean	0.984	0.461	0.21	0.286	0.727	5.744	0.508
std. dev	0.013	0.277	0.141	0.143	0.649	6.504	

(c) Differences between self-reference and one-shot models.

	a	b	σ_η	σ_ϵ	θ	s
mean	−0.006***	0.151***	0.05**	−0.064***	−0.352***	−0.567
std. error	0.002	0.037	0.02	0.02	0.094	1.009

Note: ***, ** and * indicate statistical significance at the 1, 5 and 10% levels, respectively. The Kalman gain (K) is calculated based on the mean estimates of a, σ_η and σ_ϵ.

The estimated sluggishness parameter is lower than in the one-shot model, suggesting that the self-reference model relies on a closer reference distribution. When agents refer to a more distant distribution, subjective surprises tend to be large and subsequent surprises also increase. To ensure stability of the self-reference DE process, the sluggishness parameter should not take large values. Note that all mean estimates of parameters are statistically different from zero according to Student's t-test.

Table 5 Panel (c) shows the difference between the two models, with standard errors from a t-test. I focus on θ and sluggishness, which are most closely related to DE. The self-reference model has a lower value of θ than the one-shot model. The difference in θ is statistically significant at the 1% level. The impact of representativeness in the self-reference model is weaker than that in the one-shot model. This result is consistent with the theoretical prediction.

Under DE, overreaction arises because agents overestimate the probability of the representative states when they observe a surprise. In the one-shot model, the size of the surprise is determined by rational expectations. Thus, the only source of overreaction is the strength of the psychological bias that distorts the update rule. In contrast, in the self-reference model, the magnitude of the surprise depends on past distorted beliefs. Because agents have already overreacted in previous periods, the resulting surprises become larger. As larger surprises make the states that come to mind more extreme, the update itself does not need to be as heavily distorted.

In the baseline results, agents in the one-shot model react about twice as strongly as under rational expectations, whereas the reaction in the self-reference model is about 1.7 times as large. It turns out that roughly 30% of the overreaction in the one-shot model is attributable to the fact that the surprise is derived from rational expectations, which agents do not actually hold.

5.2. Robustness check

The baseline simulations include parameter combinations with $\theta = 0$ and estimate θ by minimizing the Euclidean loss function. For both the one-shot and self-reference DE models, $\theta = 0$ corresponds to rational expectations. Nevertheless, the results show that both diagnostic models fit the data better than the rational expectations benchmark. As a robustness check, I compare the Euclidean distance across the two diagnostic models and show that the self-reference model yields a smaller distance than the one-shot model, suggesting that the former model fits better.

To formally assess this difference, I test the null hypothesis that the data are generated by the estimated one-shot model. In this test, I conduct a parametric bootstrap test. For each bootstrap sample, I compute the Euclidean distance between the coefficient vector implied by the estimated parameters and the realized target vector. Then, I evaluate how frequently the self-reference model outperforms the one-shot model under the null hypothesis.

First, I compute the Euclidean distance based on the estimated parameters for both models. The estimated parameters are reported in Table 5. Note that the sluggishness parameter, s, which captures the lag in quarters of the reference distribution, is rounded. The procedure for computing the Euclidean distance follows that of the baseline simulation.

I denote the Euclidean distance for model $d \in \{OS, SR\}$ as $l^d_{obs}(v) = \|v^*_d - \overline{v}\|$ where v^*_d is the coefficient vector computed by the estimated parameters for model d. Table 6 shows the resulting distances. The distance for the self-reference model l^{SR}_{obs} is 5.29, whereas that for the one-shot model l^{OS}_{obs} is 5.321.

The difference, $\Delta_{obs} = l^{SR}_{obs} - l^{OS}_{obs}$, is -0.032. The smaller Euclidean distance of the self-reference model indicates that the coefficient vector—comprising the autocorrelations of log EPS and the coefficients from the CG regression—computed under the self-reference model is closer to the realized data, implying that the self-reference model provides a better fit.

To confirm this, I generate B bootstrap samples. For each bootstrap sample $b = 1, \cdots, B$, I simulate time series of fundamentals and signals. Since the null hypothesis is that the data are generated by the estimated one-shot model, bootstrap samples are generated using the estimated parameter vector of the one-shot model reported in Table 5(a). Using the simulated signals, I compute their autocorrelations and estimate the CG regression coefficients to construct a bootstrap target vector, $\overline{v}^b$. Then, I estimate the optimal parameter combination for each DE model. The estimation procedure follows that of the baseline simulation, except that the target vector is $\overline{v}^b$ rather than the realized data.

For each model $d \in \{OS, SR\}$, the estimated parameter combination minimizes the Euclidean distance $l^d(v_b) = \|v - \overline{v}^b\|$. Based on this estimated parameter combination, I compute the corresponding Euclidean distance $l^d(v_b)$. Using these distances for both DE models, I construct the distance difference

$$\Delta^b = l^{SR}(v_b) - l^{OS}(v_b) \tag{16}$$

Table 6. Robustness test

$\mathbf{l}^{\mathbf{SR}}_{\mathbf{obs}}(\boldsymbol{v}) = \|\boldsymbol{v}^*_{\boldsymbol{SR}} - \overline{\boldsymbol{v}}\|$	$\mathbf{l}^{\boldsymbol{OS}}_{\mathbf{obs}}(\boldsymbol{v}) = \|\boldsymbol{v}^*_{\boldsymbol{OS}} - \overline{\boldsymbol{v}}\|$	Δ_{obs}	$p_{bootstrap}$
5.29	5.321	−0.032	0.011

Note: $l^d_{obs}(v) = \| v^*_d - \overline{v} \|$ where $d \in \{OS, SR\}$ denotes the Euclidean distance between the coefficient vector implied by the estimated parameters of model d and the realized target vector. $\Delta_{obs} = l^{SR}_{obs}(v) - l^{OS}_{obs}(v)$ is the difference in Euclidean distances between the self-reference and one-shot models. Δ_{obs} is computed from unrounded distances. $p_{bootstrap}$ is a bootstrap p-value. The bootstrap sample size is $B = 1000$.

In this robustness check, the empirical distribution of Δ^b is interpreted as the bootstrap null distribution. The bootstrap p-value is computed as

$$p_{bootstrap} = \Pr\left(\Delta^b \leq \Delta_{obs}\right) \tag{17}$$

I set the number of bootstrap samples to $B = 1000$, following the recommendation in Horowitz (2001) that B be chosen between 500 and 2000.

As shown in Table 6, $p_{bootstrap}$ is 0.011. This small p-value indicates that, under the null hypothesis in which the data are generated by the estimated one-shot model, it is unlikely that the self-reference model exhibits a smaller Euclidean distance than the one-shot model. Nevertheless, the Euclidean distance difference observed in the realized data, Δ_{obs}, is negative. This implies that the superior fit of the self-reference model is unlikely to arise from random sampling variation. Rather, it provides statistical evidence that the self-reference model fits the data better than the one-shot model.

6. Conclusion

Under DE, investors overestimate representative states and overreact to observed signals. As a result, their beliefs are systematically distorted by the representativeness heuristic.

Most of the existing literature assumes that agents compare current rational beliefs with past ones and overestimate the representative states implied by current beliefs. However, this assumption renders current distorted beliefs independent of the past distorted beliefs and makes it inconsistent with experience effects.

This paper develops a self-reference DE in which agents update their beliefs based on their past distorted beliefs. Within this framework, I derive a sufficient condition on the strength of representativeness under which DE remains stable.

I estimate the model parameters using the SMM. The results show that the self-reference model is characterized by weak representativeness effects and shorter reference periods than the one-shot model, consistent with the mechanism of the self-reference model. Moreover, the self-reference model yields smaller Euclidean distances than the one-shot model, indicating a superior fit to the data.

The main contribution is the development of the self-reference DE model. I derive sufficient conditions under which DE remains stable and provide empirical evidence that the one-shot model tends to overestimate the strength of the representativeness heuristic.

This paper also has implications for analysts' belief formation. While many studies document analysts' overreaction to information (La Porta, 1996; Gennaioli and Shleifer, 2018), which is also observed in this paper, the results suggest that analysts update their beliefs in a self-referential manner, in the sense that their beliefs are formed with reference to their own past beliefs rather than past rational beliefs. These findings offer new insights into analysts' belief formation and behavior.

A caveat of this study is the difference in the data used. I obtained data from Refinitiv Eikon over the period from 1982 to 2023, whereas Bordalo et al. (2019) used IBES data between 1981 and 2016. These differences may partly account for the discrepancies in the estimated macroeconomic parameters shown in Table 5. Compared with prior literature, the results in this paper are characterized by higher persistence of the fundamental, lower persistence of the signal, and lower signal informativeness. Although the key parameters are in line with prior literature, these differences may affect the estimated values.

NOTES

1. More precisely, agents compare the current rational belief when the arriving signal is exactly same to the predicted values. In such atomic situation, no additional information is available and the updated belief does not change from the past belief.
2. In the steady state, the variance of fundamental is given as the solution to $a^2\sigma_f^4 + \sigma_f^2\left[\sigma_\eta^2 + \left(1 - a^2\right)\sigma_\epsilon^2\right] - \sigma_\eta^2\sigma_\epsilon^2 = 0$. In my model, subjective variance of fundamental is assumed to be stationary, which is the same as in the previous literature.

REFERENCES

Augenblick, N., Lazarus, E., & Thaler, M. (2021). Overinference from weak signals and underinference from strong signals. *arXiv preprint arXiv*, 2109, 09871.

Barberis, N., Greenwood, R., Jin, L., & Shleifer, A. (2015). X-CAPM: An extrapolative capital asset pricing model. *Journal of Financial Economics*, 115(1), 1–24.

Barberis, N., Greenwood, R., Jin, L., & Shleifer, A. (2018). Extrapolation and bubbles. *Journal of Financial Economics*, 129(2), 203–227.

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

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics*, 49(3), 307–343.

Bianchi, F., Ilut, C., & Saijo, H. (2024a). Diagnostic business cycles. *Review of Economic Studies*, 91(1), 129–162.

Bianchi, F., Ilut, C. L., & Saijo, H. (2021a). *Diagnostic business cycles* (Working paper).

National Bureau of Economic Research.

Bianchi, F., Ilut, C. L., & Saijo, H. (2021b). *Implications of diagnostic expectations: Theory and applications*.

Bianchi, F., Ilut, C. L., & Saijo, H. (2024b). *Smooth diagnostic expectations* (Technical report). National Bureau of Economic Research.

Bianchi, F., Ludvigson, S. C., & Ma, S. (2022). Belief distortions and macroeconomic fluctuations. *American Economic Review*, 112(7), 2269–2315.

Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., & Shleifer, A. (2020a). Memory and representativeness. *Psychological Review*, 128(1), 71.

Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. (2016). Stereotypes. *The Quarterly Journal of Economics*, 131(4), 1753–1794.

Bordalo, P., Gennaioli, N., Kwon, S. Y., & Shleifer, A. (2021b). Diagnostic bubbles. *Journal of Financial Economics*, 141(3), 1060–1077.

Bordalo, P., Gennaioli, N., La Porta, R., & Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6), 2839–2874.

Bordalo, P., Gennaioli, N., La Porta, R., & Shleifer, A. (2020b). *Belief overreaction and stock market puzzles*. National Bureau of Economic Research.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2018). Diagnostic expectations and credit cycles. *The Journal of Finance*, 73(1), 199–227.

Bordalo, P., Gennaioli, N., & Shleifer, A. (2022). Overreaction and diagnostic expectations in macroeconomics. *Journal of Economic Perspectives*, 36(3), 223–244.

Bordalo, P., Gennaioli, N., Shleifer, A., & Terry, S. J. (2021a). *Real credit cycles* (Working paper). National Bureau of Economic Research.

Bouwman, C. H., & Malmendier, U. (2015). Does a bank's history affect its risk-taking? *American Economic Review*, 105(5), 321–325.

Coibion, O., & Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8), 2644–2678.

d'Arienzo, D. (2020). *Maturity increasing overreaction and bond market puzzles*. Available at SSRN 3733056.

Gennaioli, N., & Shleifer, A. (2010). What comes to mind. *The Quarterly Journal of Economics*, 125(4), 1399–1433.

Gennaioli, N., & Shleifer, A. (2018). *A crisis of beliefs: Investor psychology and financial fragility*.

Glaeser, E. L., & Nathanson, C. G. (2017). An extrapolative model of house price dynamics. *Journal of Financial Economics*, 126(1), 147–170.

Greenwood, R., & Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3), 714–746.

Horowitz, J. L. (2001). The bootstrap. In *Handbook of econometrics* (Vol. 5, pp. 3159–3228). Elsevier.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430–454.

Krishnamurthy, A., & Li, W. (2020). *Dissecting mechanisms of financial crises: Intermediation and sentiment* (Working paper). National Bureau of Economic Research.

La Porta, R. (1996). Expectations and the cross-section of stock returns. *The Journal of Finance*, 51(5), 1715–1742.

L'Huillier, J.-P., Singh, S. R., & Yoo, D. (2024). Incorporating diagnostic expectations into the new Keynesian framework. *Review of Economic Studies*, 91(5), 3013–3046.

Malmendier, U. (2021). Experience effects in finance: Foundations, applications, and future directions. *Review of Finance*, 25(5), 1339–1363.

Malmendier, U., & Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1), 373–416.

Malmendier, U., Pouzo, D., & Vanasco, V. (2020). Investor experiences and financial market dynamics. *Journal of Financial Economics*, 136(3), 597–622.

Malmendier, U., & Wellsjo, A. S. (2024). Rent or buy? Inflation experiences and homeownership within and across countries. *The Journal of Finance*, 79(3), 1977–2023.

Maxted, P. (2024). A macro-finance model with sentiment. *Review of Economic Studies*, 91(1), 438–475.

Nagel, S., & Xu, Z. (2022). Asset pricing with fading memory. *The Review of Financial Studies*, 35(5), 2190–2245.

Appendix A. Proof of Proposition 1

Proof. The probability density of DE is given by equation (4). For self-reference DE, the target distribution is a conditional distribution $N\left(\hat{f}^{RD}_{i,t}, \sigma_f^2\right)$ after observing the current signal, and the reference distribution is a conditional distribution $N\left(a\hat{f}^{SR}_{i,t-1}, \sigma_f^2\right)$ when the signal surprise is zero. Prior belief of both distributions is given by equation (6). I substitute these distributions into equation (4).

$$h^{SR}\left(f_{i,t} \mid x_{i,t}\right) = \frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-\hat{f}^{RD}_{i,t}\right)^2}{2\sigma_f^2}\right)\left[\frac{\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-\hat{f}^{RD}_{i,t}\right)^2}{2\sigma_f^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-a\hat{f}^{SR}_{i,t-1}\right)^2}{2\sigma_f^2}\right)}\right]^{\theta} Z$$

where Z is a constant ensuring that the distorted density $h^{SR}\left(f_{i,t}|x_{i,t}\right)$ integrates to 1. Collecting terms, I obtain

$$h^{SR}\left(f_{i,t} \mid x_{i,t}\right) = \frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-\hat{f}^{RD}_{i,t}\right)^2+\theta\left[\left(f_{i,t}-\hat{f}^{SR}_{i,t}\right)^2-\left(f_{i,t}-a\hat{f}^{SR}_{i,t-1}\right)^2\right]}{2\sigma_f^2}\right)Z$$

Since $\hat{f}^{RD}_{i,t}$ and $\hat{f}^{SR}_{i,t-1}$ are constant, the exponent term is

$$-\frac{\left(f_{i,t}-\left[\hat{f}^{RD}_{i,t}+\theta\left(\hat{f}^{RD}_{i,t}-\hat{f}^{SR}_{i,t-1}\right)\right]\right)^2+C}{2\sigma_f^2}$$

where $C=\left(\hat{f}^{RD}_{i,t}\right)^2+\theta\left[\left(\hat{f}^{RD}_{i,t}\right)^2-\left(a\hat{f}^{SR}_{i,t-1}\right)^2\right]-\left[\hat{f}^{RD}_{i,t}+\theta\left(\hat{f}^{RD}_{i,t}-a\hat{f}^{SR}_{i,t-1}\right)\right]^2$ is a constant and does not depend on $f_{i,t}$. Since $h^{SR}\left(f_{i,t} \mid x_{i,t}\right)$ is a probability density, it integrates to 1.

$$\int h^{SR}\left(f_{i,t} \mid x_{i,t}\right)df_{i,t} = \int\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-\left[\hat{f}^{RD}_{i,t}+\theta\left(\hat{f}^{RD}_{i,t}-\hat{f}^{SR}_{i,t-1}\right)\right]\right)^2}{2\sigma_f^2}\right)\exp\left(-\frac{C}{2\sigma_f^2}\right)Z\,df_{i,t}$$

$$=\exp\left(-\frac{C}{2\sigma_f^2}\right)Z\int\frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t}-\left[\hat{f}^{RD}_{i,t}+\theta\left(\hat{f}^{RD}_{i,t}-\hat{f}^{SR}_{i,t-1}\right)\right]\right)^2}{2\sigma_f^2}\right)df_{i,t}=1$$

The term inside the integral is interpreted as a normal density with mean $\hat{f}^{RD}_{i,t} + \theta\left(\hat{f}^{RD}_{i,t} - \hat{f}^{SR}_{i,t-1}\right)$ and variance σ_f^2 so that the latter term is 1.

$$\exp\left(-\frac{C}{2\sigma_f^2}\right)Z * 1 = 1$$

$$Z = \exp\left(\frac{C}{2\sigma_f^2}\right)$$

Substituting Z into the density function,

$$h^{SR}\left(f_{i,t} \mid x_{i,t}\right) = \frac{1}{\sqrt{2\pi\sigma_f^2}}\exp\left(-\frac{\left(f_{i,t} - \left[\hat{f}^{RD}_{i,t} + \theta\left(\hat{f}^{RD}_{i,t} - \hat{f}^{SR}_{i,t-1}\right)\right]\right)^2}{2\sigma_f^2}\right)$$

The self-reference DE is characterized by the mean $\hat{f}^{RD}_{i,t} + \theta\left(\hat{f}^{RD}_{i,t} - \hat{f}^{SR}_{i,t-1}\right)$ and variance σ_f^2. Because $\hat{f}^{RD}_{i,t}$ is given by equation (7), we have

$$\hat{f}^{SR}_{i,t} = \hat{f}^{RD}_{i,t} + \theta\left(\hat{f}^{RD}_{i,t} - a\hat{f}^{SR}_{i,t-1}\right) = a\hat{f}^{SR}_{i,t-1} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}^{SR}_{i,t-1}\right)$$

Appendix B. Derivation of Equation (10)

The probability density of self-reference DE is given by equation (4). Now consider the case in which the reference distribution is s periods prior. The target distribution is the current conditional distribution, $N\left(\hat{f}_{i,t}^{RD}, \sigma_f^2\right)$, obtained by updating the prior distorted distribution using the Kalman filter, and the reference distribution is the distorted distribution based on beliefs in $t-s$ periods, given as $N\left(a^s \hat{f}_{i,t-s}^{SR}, \sigma_f^2\right)$. Substituting these distributions into equation (4),

$$h^{SR}\left(f_{i,t} \mid x_{i,t}\right) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{\left(f_{i,t} - \hat{f}_{i,t}^{RD}\right)^2}{2\sigma_f^2}\right) \left[\frac{\frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{\left(f_{i,t} - \hat{f}_{i,t}^{RD}\right)^2}{2\sigma_f^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{\left(f_{i,t} - a^s \hat{f}_{i,t-s}^{SR}\right)^2}{2\sigma_f^2}\right)}\right]^{\theta} Z$$

Following the same steps as in the proof of Proposition 1, I obtain the following equation.

$$h^{SR}\left(f_{i,t} \mid x_{i,t}\right) = \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{\left(f_{i,t} - \left[\hat{f}_{i,t}^{RD} + \theta\left(\hat{f}_{i,t}^{RD} - a^s \hat{f}_{i,t-s}^{SR}\right)\right]\right)^2}{2\sigma_f^2}\right)$$

Under self-reference DE, when the reference distribution is s periods prior, the expected value of the current fundamental is

$$\hat{f}_{i,t}^{SR} = \hat{f}_{i,t}^{RD} + \theta\left(\hat{f}_{i,t}^{RD} - a^s \hat{f}_{i,t-s}^{SR}\right)$$

Substituting equation (7), I obtain the following equation.

$$\begin{aligned} \hat{f}_{i,t}^{SR} &= (1+\theta)\left[a\hat{f}_{i,t-1}^{SR} + K\left(x_{i,t} - bx_{i,t-1} - a\hat{f}_{i,t-1}^{SR}\right)\right] - \theta a^s \hat{f}_{i,t-s}^{SR} \\ &= (1+\theta)(1-K)a\hat{f}_{i,t-1}^{SR} - \theta a^s \hat{f}_{i,t-s}^{SR} + (1+\theta)K\left(x_{i,t} - bx_{i,t-1}\right) \end{aligned}$$

Appendix C. Proof of Lemma 1

Proof. Consider the reference periods $s \geq 1$ and any integer $n \geq 1$. In this case, the self-reference DE is given by equation (10). Given that the fundamental process follows equation (1) and that the expected value of η_t is 0,

$$\begin{aligned}\mathbb{E}_t^{SR}[f_{t+1+n}] &= \mathbb{E}_t^{SR}\left[a^{n+1} f_t + \sum_{j=0}^{n} a^j \eta_{t+n+1-j}\right] \\ &= a^{n+1} \hat{f}_t^{SR}\end{aligned}$$

Now consider the iterated expected values.

$$\begin{aligned}\mathbb{E}_t^{SR}\left[\mathbb{E}_{t+1}^{SR}[f_{t+1+n}]\right] &= \mathbb{E}_t^{SR}\left[a^n \hat{f}_{t+1}^{SR}\right] \\ &= \mathbb{E}_t^{SR}\left[a^n\left\{(1+\theta)(1-K)a\hat{f}_t^{SR} + (1+\theta)K(x_{t+1} - bx_t) - \theta a^s \hat{f}_{t-s+1}^{SR}\right\}\right] \\ &= a^n\left\{(1+\theta)(1-K)a\hat{f}_t^{SR} + (1+\theta)K\mathbb{E}_t^{SR}[(af_t + bx_t + \eta_{t+1} + \epsilon_{t+1}) - bx_t]\right. \\ &\quad \left. -\theta a^s \hat{f}_{t-s+1}^{SR}\right\} \\ &= a^n\left\{(1+\theta)(1-K)a\hat{f}_t^{SR} + (1+\theta)Ka\hat{f}_t^{SR} - \theta a^s \hat{f}_{t-s+1}^{SR}\right\} \\ &= a^n\left\{(1+\theta)a\hat{f}_t^{SR} - \theta a^s \hat{f}_{t-s+1}^{SR}\right\} \\ &= a^n\left\{a\hat{f}_t^{SR} + \theta\left(a\hat{f}_t^{SR} - a^s \hat{f}_{t-s+1}^{SR}\right)\right\} \\ &= \mathbb{E}_t^{SR}[f_{t+1+n}] + \theta\left(a\hat{f}_t^{SR} - a^s \hat{f}_{t-s+1}^{SR}\right)\end{aligned}$$

The second equality holds by substituting equation (10). The third equality holds by substituting equations (1) and (2) and taking all terms known at time t outside of the expectation operator.

The tower-like relation at the level of conditional means holds if $\theta\left(a\hat{f}_t^{SR} - a^s \hat{f}_{t-s+1}^{SR}\right) = 0$. For $\theta = 0$, self-reference DE reduces to rational expectations and the tower-like relation holds. In this case, the law of iterated expectations holds in the usual sense, not merely at the level of conditional means. For $\theta > 0$, the tower-like relation at the level of conditional means generally holds if and only if $s = 1$.