American Economic Review 2022, 112(7): 2269–2315
https://doi.org/10.1257/aer.20201713

Belief Distortions and Macroeconomic Fluctuations[†]

By Francesco Bianchi, Sydney C. Ludvigson, and Sai Ma*

This paper combines a data-rich environment with a machine learning algorithm to provide new estimates of time-varying systematic expectational errors (“belief distortions”) embedded in survey responses. We find sizable distortions even for professional forecasters, with all respondent-types overweighting the implicit judgmental component of their forecasts relative to what can be learned from publicly available information. Forecasts of inflation and GDP growth oscillate between optimism and pessimism by large margins, with belief distortions evolving dynamically in response to cyclical shocks. The results suggest that artificial intelligence algorithms can be productively deployed to correct errors in human judgment and improve predictive accuracy. (*JEL* C45, D83, E23, E27, E31, E32, E37)

How important are belief distortions in economic decision-making and what is their relation to macroeconomic fluctuations? Large theoretical literatures have emerged to argue that systematic expectational errors embedded in beliefs can have important dynamic effects on the economy. Less is known about the empirical relation of any such distortions with macroeconomic activity.

To formalize our notion of “belief distortion,” let us define it in general terms as *an ex ante expectational error generated by the systematic misweighting of available information demonstrably pertinent to the accuracy of the belief*. This definition nests those that consider errors generated by merely omitting relevant information to include any instance where information is suboptimally given too much or too little

*Bianchi: Department of Economics, Johns Hopkins University, and Department of Economics, Duke University, CEPR, and NBER (email: fb36@duke.edu); Ludvigson: Department of Economics, CEPR, and NBER (email: sydney.ludvigson@nyu.edu); Ma: Federal Reserve Board of Governors (email: sai.ma@frb.gov). Emi Nakamura was the coeditor for this article. Ludvigson acknowledges financial support from the C. V. Starr Center for Applied Economics at NYU. We thank Marios Angeletos and Fabrice Collard for providing data on their estimated cyclical shocks, and Michael Boutros, Josue Cox, Justin Shugarman, and Yueteng Zhu for excellent research assistance. We are grateful to Marios Angeletos, Rudi Bachmann, Fabrice Collard, Andrew Foerster, Xavier Gabaix, David Hershleifer, Cosmin Ilut, Anil Kashyp, Yueran Ma, Laura Veldkamp, and to seminar participants at the Bank of Israel, Chicago Booth, Duke, the Federal Reserve Board, King’s Business School, MIT, the Richmond Federal Reserve Bank, UC Berkeley, the 2021 annual meeting of the American Economic Association, the 2022 annual meeting of the American Finance Association, the July 2020 NBER Behavioral Macro workshop, the 2019 New Approaches for Modeling Expectations in Economics Conference (London), the 2019 Conference on Applied Macro-Finance (Melbourne), the 2020 Federal Reserve System Econometrics Meeting, and the 2020 Stanford Institute for Theoretical Economics Workshop on Asset Pricing, Macro Finance, and Computation for many helpful comments. The views expressed are those of the authors and do not necessarily reflect those of the Federal Reserve Board or the Federal Reserve System.

[†]Go to https://doi.org/10.1257/aer.20201713 to visit the article page for additional materials and author disclosure statements.

weight. In the theoretical macroeconomic literatures where distorted beliefs play a role, economic agents make systematic expectational errors due to a wide variety of reasons. These include the presence of information frictions driven by rational or behavioral inattention, the use of simple extrapolative rules, the intentional adoption of conservatively pessimistic beliefs, the overreaction to incoming news, or the presence of skewed priors, among others.

A fundamental challenge in assessing the role of distorted beliefs in macroeconomic fluctuations is that no objective measure of such distortions exists. So far, empirical work has largely proceeded by investigating whether forecast errors made by survey respondents deviate from the standard of full information and rational expectations. Yet a review of the literature discussed below finds little agreement on how such a theoretical standard should be measured. Existing studies differ according to the specific surveys that are investigated, the segment of the population that is surveyed, the topic of the survey questions, the time period to which the survey questions pertain, and the empirical methodology used to identify systematic errors in expectations. Perhaps most important, given the wide-ranging theoretical literatures cited above and the vast amount of information that could be considered ex ante known and pertinent to economic decision making, it is not obvious what benchmark model of beliefs should be applied to measure any distortion in survey responses.

This paper proposes new measures of systematic expectational errors in survey responses and relates them to macroeconomic activity. Our objective is to construct and study a comprehensive, methodologically consistent, econometric measure of belief distortions in macroeconomic expectations by looking across a range of surveys, a range of agent types, and a range of questions about future economic outcomes. A general premise of our approach is that big data algorithms can be productively employed to reveal subjective biases in human judgments. Once we have a method for uncovering those biases, artificial intelligence algorithms can be deployed to "correct" those errors and improve predictive accuracy.

Returning to our definition of belief distortions above, it is clear that measuring the errors in human judgments requires four key ingredients. First, we require direct evidence on what economic decision-makers actually believe. For this we obtain data from several different surveys, different survey questions, and broad cross sections of survey respondents with different beliefs.

Second, we must cope with the theoretically vast quantity of available information that is possibly pertinent to belief accuracy. For this, we use tools for data-rich environments along with machine learning to process hundreds of pieces of information that would have been available to survey respondents in real time at daily, quarterly, and monthly sampling intervals. Our approach also explicitly recognizes that forecasters may combine public information with private information or judgment in their predictions, and this too must be taken into account in the quantification of forecaster distortion. We argue below that artificial intelligence algorithms can effectively control for that intangible information using a specification that conditions on the current survey forecast.

Third, we must account for other bona fide features of real time decision-making, such as the out-of-sample nature of forward-looking judgments. Failure to properly account for either the data-rich environment in which survey respondents operate or the out-of-sample nature of their forecasts can lead to erroneous conclusions

about belief distortions and their relation to the macroeconomy. Conversely, using information that may have been *un*available to survey respondents to compute a standard of nondistorted beliefs could be equally erroneous. To address these issues, we develop a dynamic machine learning algorithm explicitly designed to combat overfitting in order to detect demonstrable, ex ante expectational errors in real time.

The fourth and final ingredient is the availability of observations on both survey responses and objective economic information over a sufficiently long time span. This is required to reduce sampling noise, as is necessary to distinguish bad luck in a random environment from a systematic misweighting of information, as well as to statistically infer the relation of any belief distortions to cyclical fluctuations.

With these ingredients in hand, we ask whether cross sections of survey respondents with different beliefs systematically misweight pertinent economic information. If the machine detects a sustained pattern of demonstrable, ex ante errors in survey respondents' forecasts, the magnitude of these distortions should be evident from the relative (machine versus respondent-type) out-of-sample forecast errors once averaged over a sample sufficiently long so as to eliminate differences in ex post predictive outcomes attributable to random error.

Machine learning is itself a model of belief formation. We argue that it provides an appropriate benchmark for quantifying biases in survey responses, for two reasons. One is that optimized approaches to real world decision and prediction problems almost always require the efficient processing of large amounts of information. This clearly applies to professional forecasters who are presumably among the most informed agents in the economy, but also to other agent types, including investors, firms, governments, and even households. Machine algorithms are advantageous in this regard because they are explicitly designed to cope with large amounts of information. This is important because a benchmark based on a small amount of arbitrarily chosen information could fail to reveal systematic expectational errors or, conversely, lead to spurious evidence of systematic error. A second reason is that a machine algorithm can easily be coded to systematically adapt to new information as it becomes available and to make out-of-sample forecasts on this basis. Thus the approach does not run the risk of spuriously indicating that respondent performance is suboptimal merely because of the existence of structural breaks and/or the arrival of new information that even an efficient information processing algorithm could have learned about only slowly over time. More generally, we argue that the machine-based methods offer hope for improving prediction and estimation in a range of settings that rely on human surveys as a major empirical input.

Inherent in our machine-based approach is the idea that minding key features of real world expectation formation is essential when establishing a benchmark against which belief distortions are measured. Whether doing so matters in practice, however, is an empirical question. On this question we can report at least three ways in which our results differ from some in the extant literature. First, in contrast to well known results from in-sample regressions, we find little evidence that lagged ex ante revisions in survey forecasts have predictive power for average survey forecast errors. Second, information found elsewhere to be consequential for out-of-sample prediction in a low-dimensional setting is often found to be unimportant in our high-dimensional, data-rich setting. Third, measures of belief distortions created

by comparing ex ante survey expectations with theoretical benchmarks that rely on ex post historical outcome data overstate the magnitude of distortion.

Our main economic findings may be summarized as follows. First, across a range of surveys, variables, and respondent-types with heterogeneous beliefs, the machine model produces lower mean squared forecast error over long external evaluation samples, sometimes by large margins. The magnitude of improvement is especially large in the last five years of the sample, from 2013:II to 2018:II. A key finding is that survey respondents of all types place too much weight on the private or judgmental component of their forecasts and too little weight on objective, publicly available economic information. We present below a simple model of public and private signals that facilitates this interpretation.

Second, survey expectations of inflation for the median respondent of all surveys are biased upward on average, a direction we shall refer to as "pessimistic." By contrast, survey expectations of economic growth by professional forecasters and corporate executives are "optimistic" on average—i.e., biased upward, while they are very slightly pessimistic for households. These biases are found to be largest at the end of our sample, from 2013:II–2018:II, when the median forecast of economic growth from the Survey of Professional Forecasters (SPF) was biased upward by an amount equal to 38 percent of actual GDP growth over this period, resulting in forecasts that were 18 percent less accurate on average than the machine specification. The median SPF forecaster also persistently overestimated inflation during this time period, resulting in forecasts that were 37 percent less accurate than the machine specification.

Third, although our machine learning algorithm indicates that sparse specifications are often optimal, this is not the case in every period. Moreover, even with sparse specifications, the precise information utilized changes from period to period. These results underscore the importance of using a dynamic, large-scale information processing algorithm to reduce errors in human judgment, even if much of the information the algorithm considers is associated with a coefficient that is shrunk all the way to zero most of the time.

Fourth, although the machine is able to detect patterns in the data that notably improve predictive accuracy over human forecasts, these improvements produce smaller estimates of belief distortion than suggested by some previous empirical studies in the behavioral macro literature, as discussed below. In contrast to these studies, however, our benchmark measure of nondistorted beliefs is formed by (i) relying exclusively on information that we can verify could have been known to survey respondents on or before the survey response deadline, (ii) requiring the machine to choose every aspect of the forecasting specification (including the predictor variables, the lag orders, the window lengths, etc.) ex ante rather than with hindsight, and (iii) employing genuine out-of-sample prediction in the external validation step. The strict adherence to these principles means that we find a smaller magnitude of belief distortion in the predictions of professional forecasters than some previous studies. Moreover, there are times in our sample when the machine makes large ex post mistakes. A notable example is the Great Recession, which the machine failed to recognize in real time, resulting in large forecast errors similar in magnitude to those made by professional forecasters during this episode. We argue that such episodes underscore the role of largely unforeseen events in generating

occasionally large prediction error, not all of which can be attributed to a systematic bias in expectations.

The rest of this paper is organized as follows. Section I reviews the related literature. Section II describes our econometric and machine learning framework. Section III describes results pertaining to our estimates of belief distortions and relates our estimates with those from approaches used in some well-known prior empirical studies. Section IV contains results on how belief distortions change over the business cycle. Section V concludes. A large amount of additional material on our data construction, estimation procedures, and additional robustness checks have been placed in an online Appendix.

I. Related Literature

Our estimates provide a benchmark to evaluate theories for which information capacity constraints, extrapolation, sentiments, ambiguity aversion, and other departures from full information, rational expectations play a role in business cycles.

In these theoretical literatures, economic agents make systematic expectational errors for a variety of reasons. These reasons include the presence of information frictions that lead agents to act in a "boundedly rational" manner because they are incapable of attending to or processing all the available information at a given moment (e.g., Mankiw and Reis 2002; Woodford 2002; Sims 2003; Reis 2006a, 2006b; Eusepi and Preston 2011; Gabaix 2014). Alternatively agents may be inattentive for broader behavioral reasons (e.g., Gabaix 2020). A key implication of these theories, explored in well-known work by Coibion and Gorodnichenko (2015), is that individuals *under*react to objective economic information. Our finding that belief distortions for professional forecasters are larger at the end rather than the beginning of our sample suggests that forms of bounded rationality attributable solely to limitations in information processing capacity are unlikely to fully explain our results. Similarly, our estimates are little changed if we allow the machine to observe every respondent-type's current forecast, suggesting that information frictions based on noisy "dispersed information" are also unlikely to be the most relevant source of systematic error we uncover.

Other theories postulate that individuals use simple extrapolative rules or overweight "representative" events in reacting to incoming news (e.g., De Long et al. 1990; Barberis, Shleifer, and Vishny 1998; Barberis et al. 2015; Bordalo, Gennaioli, and Shleifer 2018; Gennaioli and Shleifer 2018; Bordalo et al. 2018). Related theories propose that individuals overweight their personal experiences (e.g., Malmendier and Nagel 2011, 2015). A key implication of many of these theories is that individuals *over*react to objective information.

A literature on "sentiments" postulates that communication frictions can cause aggregate expectations to exhibit statistical biases (e.g., Angeletos and La'O 2013; Angeletos, Collard, and Dellas 2018b; Milani 2011, 2017). Other models feature "confidence shocks," or ambiguity averse agents who are deliberately pessimistic on average (e.g., Hansen and Sargent 2008; Epstein and Schneider 2010; Ilut and Schneider 2014; Bianchi, Ilut, and Schneider 2018; Ilut and Saijo 2021; Bhandari, Borovicka, and Ho 2019), or agents with skewed priors (Afrouzi and Veldkamp 2019). There remains a question of whether ambiguity aversion or skewed priors

would be revealed in survey responses. If not, such models need some other mechanism to explain the systematic expectational errors documented here and elsewhere.

Finally a theoretical literature in economic psychology studies how basic properties of cognition can give rise to human biases in expectation formation (e.g., Woodford 2013; Khaw, Stevens, and Woodford 2017).

Any of the theories above provide a mechanism through which a relatively unbiased and potentially more information-efficient machine operating in a data-rich environment would provide forecasts that deviate from those made by humans and possibly be more accurate. The objective of this study is to provide new measures of such deviations and to investigate their relation to macroeconomic fluctuations.

On the empirical side, our work follows a growing body of literature that reports evidence of belief distortions and relates them to economic activity. These papers include those that find evidence of departures from rational expectations in predicting inflation and other macro variables (Coibion and Gorodnichenko 2012, 2015; Fuhrer 2018), the aggregate stock market (Bacchetta, Mertens, and van Wincoop 2009; Amromin and Sharpe 2014; Greenwood and Shleifer 2014; Adam, Marcet, and Beutel 2017), the cross section of stock returns (Bordalo et al. 2019), credit spreads (Greenwood and Hanson 2015; Bordalo, Gennaioli, and Shleifer 2018), and corporate earnings (DeBondt and Thaler 1990; Ben-David, Graham, and Harvey 2013; Gennaioli, Ma, and Shleifer 2016; Bouchaud et al. 2019). Although these studies differ widely according to their empirical design, none take into account the data-rich context in which survey respondents operate or the dynamic, out-of-sample nature of their forecasts, gaps our study is designed to fill.

These very differences lead our findings to diverge in notable ways from some in the extant literature. For example, following Coibion and Gorodnichenko (2015), we ask whether ex ante revisions in the average forecast reduce average ex post forecast errors, as would be indicative of models that imply underreaction to economic news. Using the methodology proposed in this paper, we find no evidence that they do. Instead, the coefficients on forecast revisions are shrunk to zero by the dynamic machine algorithm in favor of placing greater absolute weight on other pieces of information. Even if we use the same empirical specification used in Coibion and Gorodnichenko (2015), forecast revisions cease to be a useful predictors of forecast errors in a dynamic context when predictions are simply made out of sample rather than in sample. Similarly, we ask whether survey respondents initially underreact to cyclical shocks but later overreact, as documented in Angeletos, Huo, and Sastry (2020). We confirm this general pattern but find that the magnitudes of under- and overreaction are much smaller than those using the methodology of Angeletos et. al., in which the benchmark for measuring nondistorted beliefs is based on historical outcome data that would not have been known to survey respondents in real time.

The literature discussed so far has little to say about overconfidence, a term generally reserved in the behavioral economics literature to describe an agent who overestimates the precision of her private signal. Yet our finding that respondents of all types place too much weight on the marginal information embedded in their own forecasts is one of the most robust and quantitatively important contributors to bias that we uncover. We present below a simple model of public and private signals to help interpret this finding. In this model, the machine will downweight the information

contained in the survey response if the forecaster either overestimates the precision of her private signal, as in traditional notions of overconfidence, and/or if she inefficiently combines the public information, thereby effectively underestimating the precision of her public signal. Either way, the forecaster gives too much weight in relative terms to the private, judgmental component of her forecast. In this regard, our findings relate to extensive finance literature that provides theory and evidence of overconfidence and its role in explaining a range of stylized facts about stock return predictability and trading patterns. Groundbreaking contributions include Odean (1998); Daniel, Hirshleifer, and Subrahmanyam (1998); Barber and Odean (2000); and Daniel, Hirshleifer, and Subrahmanyam (2001). Daniel and Hirshleifer (2015) provide an overview of this literature. More generally, our findings echo a large body of evidence in psychology showing that people—perhaps especially experts and professionals—give too much weight to their private judgments when making predictions (e.g., Kahneman, Sibony, and Sunstein 2021, ch. 10). To the best of our knowledge, this paper is the first to find evidence suggesting that systematic expectational errors in professional macroeconomic predictions are partly attributable to a strong overreliance on the implicit judgmental component of their forecasts.

Our work also connects with a pre-existing econometric forecasting literature, which finds that survey forecasts of inflation are extremely difficult if not impossible to beat with statistical models in out-of-sample forecasting (e.g., Ang, Bekaert, and Wei 2007; Del Negro and Eusepi 2011; Aiolfi, Capistrán, and Timmermann 2011; Genre et al. 2013; and Faust and Wright 2013). Indeed, these studies conclude that the very best forecasts of inflation are the subjective ones provided by surveys. By contrast, our machine learning algorithm, with its focus on detecting demonstrable ex ante errors, performs better in out-of-sample forecasting than every percentile of all of the survey forecast distributions that we study.

Finally, we are aware of relatively little work that has used machine learning as a benchmark against which belief distortions are measured. An important exception is Martin and Nagel (2019) who use it to study models of expected stock returns in the cross section. Although their context is very different from ours, they find, as we do, that accounting for the interplay between a data-rich environment and dynamic, out-of-sample forecasting generates findings about belief distortions that differ considerably from prior frameworks that sidestep these aspects of real world decision-making.

II. Econometric and Machine Learning Framework

This section describes our econometric and machine learning framework. This framework is applied to three different surveys that ask about expectations for future inflation and aggregate economic activity: the Survey of Professional Forecasters (SPF), the University of Michigan Survey of Consumers (SOC), and the Blue Chip Survey (BC). The first covers professional forecasters in a variety of institutions, the second covers households and is designed to be representative of the US population, and the third covers executives and professional forecasters at financial firms. Data from the SPF and the SOC are publicly available; BC data were purchased and hand coded for the earlier part of the sample.

A. *Overview*

Before getting into the details of our approach, we discuss two aspects of the general methodology.

First, as we argue below, a good benchmark against which any belief distortions in survey forecasts are measured must account for the information in the current survey forecast. At the same time, the analysis we undertake requires a sufficiently long time series of observations. Unfortunately, and the panel elements of the surveys are simply too limited to conduct the analysis on a respondent-level basis, since panelist participation is often brief and/or intermittent over time.[1] We therefore conduct the analysis for respondents of a particular *type* at time t, defined to be those in specific *percentiles* of the time t survey forecast distribution. A maintained assumption of our approach, explained in detail below, is that survey respondents know their own "type," so that they have a sense of where in the time t forecast distribution their belief is located. We argue that this assumption is a reasonable approximation to reality, at least for professional forecasters, who routinely and continuously telegraph updates of their forecasts to clients and the press, while at the same time monitoring the evolving predictions of other forecasters at "rival" institutions. Such forecasters are therefore likely to have very good real-time information about their location in the professional forecast distribution.

To provide support for this claim, Figure 1 reports the forecasts of four-quarter-ahead real GDP growth over time for each percentile of a given professional forecast distribution along with the same-percentile forecast of the same variable and for the same future time period, but from a closely related professional forecaster survey that was publicly available before the survey deadline faced by type i. The first panel considers the BC survey, where survey results are released every month, giving a frequent snapshot of the professional forecaster distribution. Figure 1 shows that BC forecast distribution is quite persistent from month-to-month: for every percentile of the distribution, we see that the ith percentile's time t forecast is highly correlated with last month's ith percentile forecast. Thus BC forecasters can form an excellent idea of where their current forecast is located in the time t forecast distribution by observing last month's published BC distribution. For a panelist in the quarterly SPF survey, last period's SPF forecast distribution provides more stale information by definition since it was released a quarter rather than a month ago. However, the BC survey is a similar professional panel, and may even have overlapping panelists. In addition, the timing of the two surveys' deadlines is such that an SPF panelist can observe the BC forecast distribution from approximately two weeks prior for predictions of the same variable and over the same future time period. The second panel of Figure 1 shows that the ith percentile's time t forecast from the SPF panel is highly correlated with the ith percentile's forecast from the most recent BC panel, which was released roughly two weeks before. Thus SPF forecasters can form an excellent idea of where their current forecast is located in the time t professional forecast

[1] The learning algorithm described below employs rolling estimation and training sample windows that could be as long as 34 quarters once combined, a span of data that must be available before the first out-of-sample machine forecast can be recorded. By contrast, the length of time that individual panelists remain in the survey samples is comparatively short. For example, for the SPF survey on inflation expectations, the average forecaster remains in our sample just 18.5 quarters, with gaps in participation that would require filling in missing values.

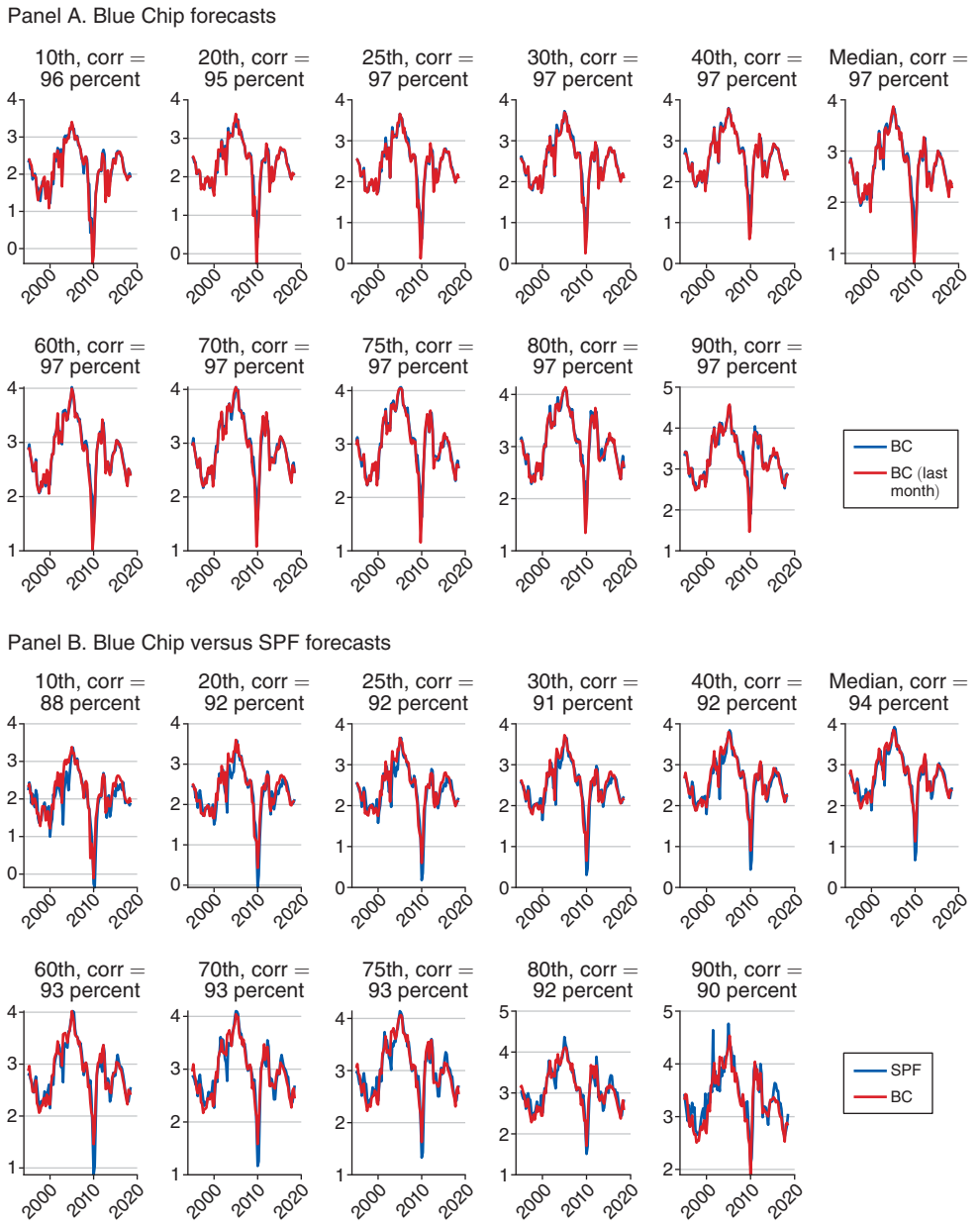

FIGURE 1. FORECASTER TYPES

Notes: Panel A of the figure reports the median Blue Chip (BC) four-quarter-ahead real GDP growth forecast (blue) along with last month's median BC forecast of the same variable for the same future subperiod (red). Panel B reports the SPF median forecast of four-quarter-ahead real GDP growth (blue) along with the most recent median BC forecast of the same variable for the same future subperiod (red). The most recent BC forecast is taken from the panel available on the closest month prior to the SPF survey deadline day, typically about two weeks in advance. The sample spans 1995:I–2018:II.

distribution from observing the most recent BC survey. These plots are consistent with the idea that professional forecasters have access to reliable observable proxies for their time t location in the professional forecast distribution.

We also argue that percentile-level analyses are likely to be a more plausible description of the empirical specifications actually employed by professional forecasters. There are well-documented caveats with the assignment of individual identification numbers to panelists who change their place of employment but remain in the survey. When panelists change places of employment, they often join entirely new forecasting teams with bespoke modeling practices and forecasting perspectives.[2] Panelists also go in and out of the surveys, with sometimes extended gaps in their participation. Under these circumstances the use of types is likely to be more natural in an econometric modeling context than specifications based on a sporadic history of individual forecaster's own predictions.

We emphasize that our approach does not require, nor do we assume, that forecasters have a *time-invariant* type. As long as they know their contemporaneous type, the approach described below works even if respondents move around in the distribution. This follows because the one-period lagged values of every percentile's forecast are publicly available information, information that we find an efficient forecast would typically place nonzero weight on. Thus, a respondent can always ask how her time t type would have done historically by appending her current forecast onto the appropriate historical series.

With this in mind, we acknowledge that a core assumption of some economic theories is that individuals do not know their "rank" in the relevant cross-sectional distribution, in which case their "rank beliefs" can be important determinants of equilibria (e.g., Morris, Shin, and Yildiz 2016). We assess below the empirical relevance of such rank beliefs for our measured belief distortions, by comparing results using our baseline machine specifications with those from an alternative benchmark that uses the union of every type's forecast at time t.

A second aspect of the methodology pertains to our overall information processing approach. In order to identify possible distortions in beliefs, it is imperative that the benchmark model of belief formation use large and varied real time information sets, so that our measure of distortion does not miss pertinent information that could have been known to survey respondents, or pertain only to models with a small number of arbitrarily chosen information variables. To address this challenge, we take a two-pronged approach that combines diffusion index estimation with machine learning. Diffusion index forecasting, wherein a relatively small number of dynamic factors are estimated from hundreds of economic time series, has become common in data-rich environments, following on a long line of prior studies showing that the approach improves prediction accuracy in a manner similar to model averaging.[3] Diffusion indexes are also useful because some forms of nonlinearities are readily handled by including polynomial functions of estimated dynamic factors, or by forming additional factors from polynomials of the raw data. We use estimated

[2] For example, see the memo "Caveats on Using the Individual Identification Numbers in the Survey of Professional Forecasters," posted at philadelphiafed.org.

[3] An incomplete list of this literature includes Stock and Watson (1989, 1991, 2002a, 2002b, 2006); Ludvigson and Ng (2007, 2009, 2010).

factors as part of a dynamic machine learning algorithm of regularized estimation that chooses shrinkage and sparsity by optimally trading off the costs of down-weighting information against the benefits of reduced parameter estimation error.[4] The diffusion index aspect of our methodology is standard, so we cover this step in the online Appendix, focusing below on the dynamic machine learning framework.

B. *Machine Efficient Benchmark*

Let $y_{j,t+h}$ generically denote an economic time series indexed by j whose value in period $h \geq 1$ a survey forecaster is asked to predict at time t. Let $\mathbb{F}_t^{(i)}$ generically denote a survey forecast made at time t and let superscript (i) denote the ith *respondent-type*, where i denotes the respondent located at the ith percentile of the survey forecast distribution, i.e., "$i = 65$" refers to the belief of the respondent at the sixty-fifth percentile. Thus $\mathbb{F}_t^{(65)}\left[y_{j,t+h}\right]$ denotes the survey expectation of $y_{j,t+h}$ that is formed at time t by the respondent at the sixty-fifth percentile of the survey distribution.

Let $x_t^C = \left(x_{1t}^C, \ldots, x_{Nt}^C\right)'$ generically denote a dataset of economic information in some category C that is available for real-time analysis. We assume that x_{it}^C has an approximate factor structure as detailed in the online Appendix, where $\mathbf{G}_t^C$ is an $r_G \times 1$ vector of latent common factors ("diffusion indexes") with Λ_i^C a corresponding $r_C \times 1$ vector of latent factor loadings.

Collect all factors from different datasets of category C, as well as nonlinear components (polynomials of factors and factors formed from polynomials of raw data) into a single r_G dimensional vector $\mathbf{G}_t$. Let $\hat{\mathbf{G}}_t$ denote consistent estimates of a rotation of $\mathbf{G}_t$ and let the r_W dimensional vector $\mathbf{W}_t$ contain additional nonfactor information that will be specified below. Finally, let $\mathbf{Z}_{jt} \equiv \left(y_{j,t}, \hat{\mathbf{G}}_t', \mathbf{W}_{jt}'\right)'$ be a $r = 1 + r_G + r_W$ vector which collects the data at time t and let $\mathcal{Z}_{jt} \equiv \left(y_{j,t}, \ldots, y_{j,t-p_y}, \hat{\mathbf{G}}_t', \ldots, \hat{\mathbf{G}}_{t-p_G}', \mathbf{W}_{jt}', \ldots, \mathbf{W}_{jt-p_W}'\right)'$ be a vector of contemporaneous and lagged values of $\mathbf{Z}_{jt}$, where p_y, p_G, p_W denote the total number of lags of $y_{j,t}, \hat{\mathbf{G}}_t', \mathbf{W}_{jt}'$, respectively. Even with the use of factors, $\mathcal{Z}_{jt}$ can be of high dimension.

With these data in hand, consider the following machine learning empirical specification for forecasting $y_{j,t+h}$ given information at time t, to be benchmarked against the time t survey forecast of respondent-type i:

$$y_{j,t+h} = \alpha_{jh}^{(i)} + \beta_{jh\mathbb{F}}^{(i)} \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] + \mathbf{B}_{jh\mathcal{Z}}^{(i)'} \mathcal{Z}_{jt} + \epsilon_{jt+h}, \quad h \geq 1, \tag{1}$$

where $\alpha_{jh}^{(i)}$, $\beta_{jh\mathbb{F}}^{(i)}$, and $\mathbf{B}_{jh\mathcal{Z}}^{(i)}$ are parameters to be estimated, and where $\mathbf{B}_{jh\mathcal{Z}}^{(i)}$ is $K \times 1$, with $K = r + p_y + p_G \cdot r_G + p_W \cdot r_W$, the number of right-hand-side variables other than $\mathbb{F}_t^{(i)}$. Equation (1) is estimated using machine learning tools, as discussed below.

[4] It is straightforward to verify using hold-out samples and/or artificial data that combining diffusion index estimation with machine learning is often better than doing either one in isolation, for two reasons. First, the optimal number of factors can still be large enough that there are clear efficiency gains to using machine learning techniques for choosing shrinkage and sparsity even when all predictors are factors. Second, it is well known that the best approaches for choosing sparsity, such as those that use the L^1 "lasso" penalty, work poorly in the context of correlated regressors. Since our raw data are correlated, we have also verified that our elastic net estimator, which utilizes an L^1 penalty, works better when the data are first transformed into orthogonal diffusion indexes before estimation. This latter finding is consistent with results in Kozak, Nagel, and Santosh (2020).

Estimation of (1) delivers a time t machine "belief" about $y_{j,t+h}$, namely the machine forecast, denoted $\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right]$. We define the *machine efficient benchmark* as a set of parameter restrictions that would imply that the survey forecaster in the ith percentile processes all available information at time t as efficiently as the machine. This benchmark corresponds to the following parameter restrictions:

$$\beta_{jh\mathbb{F}}^{(i)} = 1; \quad \mathbf{B}_{jh\mathcal{Z}}^{(i)} = 0; \quad \alpha_{jh}^{(i)} = 0. \tag{2}$$

Systematic expectational errors in the survey forecast are revealed by deviations from the benchmark above, generated by a misweighting of information contained in $\mathcal{Z}_{jt}$ or "1" (i.e., $\mathbf{B}_{jh\mathcal{Z}}^{(i)} \neq 0$ or $\alpha_{jh}^{(i)} \neq 0$) and/or the survey respondent's own forecast, $\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right]$ (i.e., $\beta_{j\mathbb{F}}^{(i)} \neq 1$). Machine estimates $\widehat{\beta}_{j\mathbb{F}}^{(i)} \neq 1$ imply that the survey response $\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right]$ could have been improved by giving it more or less weight relative to other objective economic information than the implicit weight of one given this response by the survey respondent. The machine can correct systematic errors in the human forecast by optimally reweighting the marginal information contained in $\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right]$ against the publicly available information contained in $\mathcal{Z}_{jt}$ that all survey respondents also had access to.

Three points about the machine efficient benchmark bear emphasis. First, it is a *type-specific* benchmark that adopts the perspective of a forecaster who is in the ith percentile of the survey forecast distribution in period t. The machine is given any information that the survey forecaster in the ith percentile could have observed at time t, including her own forecast $\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right]$, as well as publicly available information contained in $\mathcal{Z}_{jt}$, where the latter includes lagged values of all other type's forecasts, since all surveys publish their results shortly after the response deadline. This approach explicitly recognizes that agents might have private information or use judgment, and this needs to be taken into account in the quantification of forecaster bias. Otherwise, the benchmark model of beliefs that we apply to measure any distortion in survey responses will have omitted a possibly pertinent piece of the time t information set of agents, leading to an erroneous measurement of systematic expectational errors. The next subsection argues that an artificial intelligence algorithm can effectively control for that intangible information using a specification, such as the one here, that conditions on the current survey forecast. For now we note that, even if we don't allow the machine to see the ith percentile's contemporaneous forecast and instead proxy for that observation using publicly available data prior to t, our findings on forecaster bias are very similar.[5] But we find that specification to be far less interesting, both because it necessarily provides an inaccurate account of any forecaster bias, and because it is silent on the possible role of private information or judgment in belief distortions.

Second, the machine is given only that information at time t that the survey respondent-type in the ith percentile could have observed at time t, and nothing more. This is important because superior machine forecasts formed with ex post information that we cannot be certain the survey respondent could have observed

[5] See Table A13 of the online Appendix.

in real time might simply reflect the benefit of hindsight, rather than genuine systematic expectational error. For this reason, some popular techniques for forming benchmarks to measure forecaster bias, such as meta forecasts that pool multiple survey forecasts at time t to form a meta forecast, are ruled out by our procedure since we adhere to the principle of providing the machine only that information that we can verify was publicly available at the time of the survey forecast. This follows because the time t forecast distribution is publicly released only after all the analysts turn in their forecasts.

Third, since in principle all survey respondent-types could have accessed the same information given the machine, the time t machine forecast serves as a real-time check on whether the survey response may be making a demonstrable systematic expectational error. In practice, this artificial intelligence approach could be employed within institutions to check for, and possibly correct, biases in professional forecasts.

We compare below the forecast accuracy of the machine benchmark with the survey responses. If the machine systematically improves forecasts on average over an extended evaluation sample, we take that as evidence of belief distortion, or "bias" for short. In this event, we compute a *dynamic* measure of a survey respondent-type's belief distortion by taking the difference between the survey forecast and the machine forecast, $\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right]$, where we denote the bias of forecaster i at time t as

$$bias_{j,t}^{(i)} \equiv \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] - \mathbb{E}_t^{(i)}\left[y_{j,t+h}\right]. \tag{3}$$

Observe that $bias_{j,t}^{(i)}$ captures ex ante expectational errors, not ex post forecast errors, or "mistakes." In particular, bias in expectations is measured relative to the machine forecast, not relative to an ex post outcome. One implication of this is that it is possible that every respondent-type is biased vis-à-vis the machine ex ante, even though there will always be some respondent-type that is "right" ex post.

C. *A Model of Private and Public Signals*

Our approach explicitly recognizes that agents (most obviously professional forecasters) might have private information or use judgment, and this needs to be taken into account in the quantification of forecaster bias. Otherwise, the benchmark model of beliefs that we apply to measure any distortion in survey responses will have omitted a possibly pertinent piece of the time t information set of agents, thereby distorting our measurement of systematic expectational errors. In this subsection we argue that an artificial intelligence algorithm can effectively control for that intangible information using a specification that conditions on the current survey forecast. To do so, we present a model of forecasters as forming an overall prediction by combining a statistical forecast based on public information (a public signal) with a judgmental or private component based on information intangible to the machine (a private signal). This also facilitates an interpretation of the machine estimates of the parameters $\alpha_{jh}^{(i)}$, $\beta_{jh\mathbb{F}}^{(i)}$, and $\mathbf{B}_{jh\mathcal{Z}}^{(i)}$.

We suppose that forecasters form an overall prediction by combining a statistical forecast based on public information (a public signal) with a judgmental component based on information intangible to the machine (a private signal). Let x be publicly available information and let z be a private signal about an unknown variable y. Suppose these variables are related to one another according to the system

$$
\begin{aligned}
x &\sim iid\big(0, \sigma_x^2\big), \\
y &= \alpha x + u_2,\ u_2 \sim N\big(0, \sigma_2^2\big), \\
z &= y + u_1,\ u_1 \sim N\big(0, \sigma_1^2\big),
\end{aligned} \tag{4}
$$

where α is a parameter describing the mapping from x to y, and where x, u_1, and u_2 are i.i.d. and mutually uncorrelated with one another. In what follows, we will interpret y as the future value of a variable being forecast, z as a private signal representing judgment or intangible information that a survey forecaster can observe but the machine cannot, and x as public information that serves both as a predictor and, via the mapping αx, as a public signal. Note that x could also be a vector, while αx is still a scalar. For example, if y is future inflation, x could be a measure of the output gap, the central bank inflation target, and/or macro and financial factors formed from large datasets. The random variable u_2 is the unforecastable component of y and has the interpretation of a structural shock.

Conditional on observing both the private and public signals, the optimal forecast of y is

$$
\mathbb{E}_o\big[y \,|\, \alpha x, z\big] = \gamma z + \big(1 - \gamma\big)\alpha x, \quad \gamma \equiv \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} = \frac{\sigma_1^{-2}/\sigma_2^{-2}}{1 + \sigma_1^{-2}/\sigma_2^{-2}},
$$

where $\sigma_1^{-2}/\sigma_2^{-2}$ is the precision of the private signal relative to that of the public signal. More weight is given to the private signal when it is relatively more precise.

Suppose that a forecaster assigns weights to her private and public signals as follows:

$$
\mathbb{F} = \gamma^F z + \big(1 - \gamma^F\big)\alpha^F x.
$$

The survey forecast $\mathbb{F}$ is here interpreted as a prediction based partly on the respondent's statistical model using public information ($\alpha^F x$) and partly on a private signal z. The case $\alpha^F \neq \alpha$ arises when the forecaster inefficiently combines the public information when forecasting y, which could occur if she is inattentive to some variables and/or overly attentive to others.

How might γ^F be determined? One possibility is that the forecaster chooses $\mathbb{F}$ based on what would be optimal given the system (4), but employs incorrect estimates $\sigma_{1F}^2 \neq \sigma_1^2$, $\alpha^F \neq \alpha$, in which case we have

$$
\gamma^F = \frac{\sigma_{1F}^{-2}/\sigma_{2F}^{-2}}{1 + \sigma_{1F}^{-2}/\sigma_{2F}^{-2}},
$$

where $\sigma_{1F}^{-2}/\sigma_{2F}^{-2}$ is the ratio of her *estimated* private signal precision to her *estimated* public signal precision. Note that if the forecaster inefficiently combines the public information, i.e., $\alpha^F \neq \alpha$, she will have effectively underestimated the precision of the public signal, leading to $\sigma_{2F}^{-2} < \sigma_2^{-2}$. At the same time, it is possible that $\gamma^F = \gamma$ even if $\alpha^F \neq \alpha$ if the forecaster underestimates the precision of her private signal by exactly the right amount.

Now consider the machine forecast $\mathbb{E}$ of y, which is based on both the public information x and the survey forecast $\mathbb{F}$:

$$\mathbb{E} = \widehat{\beta}\mathbb{F} + \widehat{B}x = \widehat{\beta}\left[\gamma^F z + \left(1-\gamma^F\right)\alpha^F x\right] + \widehat{B}x,$$

where $\widehat{\beta}$ and $\widehat{B}$ are coefficient estimates. Although the machine cannot directly observe the private signal z, it can still learn about the weight assigned to it by the forecaster from observing $\mathbb{F}$. The machine estimates the coefficients $b \equiv (\beta, B)'$ from a regression of y on $(\mathbb{F}, x)'$. The online Appendix proves that this estimator results in the values:

$$\widehat{\beta} = \frac{\gamma}{\gamma^F}, \quad \widehat{B} = (1-\gamma)\alpha - \left(\frac{\gamma}{\gamma^F} - \gamma\right)\alpha^F.$$

The machine will set $\widehat{\beta} < 1$ ($\widehat{\beta} > 1$) if and only if the forecaster gives more (less) weight γ^F to her private signal z than the correct weight based on its true relative precision $\sigma_1^{-2}/\sigma_2^{-2}$. Even though the private signal component of the survey respondent's forecast is not directly observable by the machine, an artificial intelligence algorithm can effectively control for that intangible information by conditioning on $\mathbb{F}_t$. The machine can then correct for inefficiencies in the survey forecast by setting $\widehat{\beta} \neq 1$ and/or $\widehat{B} \neq 0$.

In the framework above, the case of $\gamma^F/\gamma > 1$ resulting in $\widehat{\beta} < 1$, could happen for two reasons. First, the forecaster might overestimate the precision of her private signal, i.e., $\sigma_{1F}^{-2} > \sigma_1^{-2}$, a circumstance often referred to as "overconfidence" in the behavioral economics literature. Second, the forecaster might inefficiently combine the public information, i.e., $\alpha^F \neq \alpha$, so that she effectively underestimates the precision of her public signal, i.e., $\sigma_{2F}^{-2} < \sigma_2^{-2}$. Either way, $\widehat{\beta} < 1$ indicates an overreliance, in relative terms, on the private or judgmental component of her forecast.[6] We return to a discussion of this case below when we present the machine parameter estimates.

D. *Estimator*

We now describe the machine estimation used to quantify any belief distortions. To do so, let us simplify notation by collecting all the independent variables and

[6] In principle $\hat{\beta} < 1$ could also occur if there is idiosyncratic measurement error in the survey responses. While possible, we argue that this is unlikely to be a plausible explanation for the findings reported below, for two reasons. First, such an interpretation is implausible for professional forecasters where $\hat{\beta}_{\mathbb{F},t}^{(i)}$ is nonetheless quite substantially below unity on average. Second, measurement errors should wash out in the mean survey forecast, yet estimates of this parameter for the mean, and of the average bias, are similar to those for the median.

coefficients on the right-hand side of (1) into a single matrix and vector and writing the machine model as

$$y_{j,t+h} = \mathcal{X}_t'\beta_{jh}^{(i)} + \epsilon_{jt+h}, \tag{5}$$

where $\mathcal{X}_t = \left(1, \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right], \mathcal{Z}_{jt}\right)'$ and $\beta_{jh}^{(i)} \equiv \left(\alpha_{jh}^{(i)}, \beta_{jh\mathbb{F}}^{(i)}, \left(\mathbf{B}_{jh\mathcal{Z}}^{(i)}\right)\right)'$.

Let $\mathbf{X}_T = \left(y_{j,1}, \ldots y_{j,T}, \ldots \mathcal{X}_1', \ldots, \mathcal{X}_T'\right)'$ be the vector containing all observations in a sample of size T. We consider estimators of $\beta_{jh}^{(i)}$ that take the form

$$\hat{\beta}_{jh}^{(i)} = m\left(\mathbf{X}_T, \boldsymbol{\lambda}^{(i)}\right),$$

where $m\left(\mathbf{X}_T, \boldsymbol{\lambda}^{(i)}\right)$ defines an estimator as a function of the data $\mathbf{X}_T$ and a nonnegative regularization or "tuning" parameter vector $\boldsymbol{\lambda}^{(i)}$ estimated using cross-validation. The values of $\boldsymbol{\lambda}^{(i)}$, which will be estimated dynamically over time, determine the optimal shrinkage and sparsity of the time t machine specification. Denote this latter estimator $\hat{\boldsymbol{\lambda}}_t^{(i)}$ and denote the combined final estimator $\hat{\beta}_{jh}^{(i)}\left(\mathbf{X}_T, \hat{\boldsymbol{\lambda}}_t^{(i)}\right)$. Our main approach uses the elastic net (EN) estimator, where $\boldsymbol{\lambda}^{(i)}$ is a bivariate vector that uses dual lasso and ridge penalties to achieve both shrinkage and sparsity.[7]

The estimation of (5) is repeated sequentially in rolling subsamples, with parameters estimated from information known at time t used to predict variables $y_{j,t+h}$ in *subsequent* periods. This leads to a sequence of machine efficient beliefs about $y_{j,t+h}$. Denote the coefficients and regularization parameters obtained from an estimation conducted with information through time t as $\hat{\beta}_{jh,t}^{(i)}$ and $\hat{\boldsymbol{\lambda}}_t^{(i)}$, respectively. Note that the time t subscripts on $\hat{\beta}_{jh,t}^{(i)}$ and $\hat{\boldsymbol{\lambda}}_t^{(i)}$ are used to denote one in a sequence of time-invariant parameter estimates obtained from rolling subsamples, rather than estimates that vary over time within a sample. Likewise, we shall denote the time t machine belief about $y_{j,t+h}$ as $\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right]$, defined by

$$\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right] \equiv \mathcal{X}_t'\hat{\beta}_{jh,t}^{(i)}\left(\mathbf{X}_T, \hat{\boldsymbol{\lambda}}_t^{(i)}\right).$$

Forecast errors are differentially denoted for the survey and machine

$$\text{survey error}_{t+h}^{(i)} = \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] - y_{j,t+h},$$

$$\text{machine error}_{t+h}^{(i)} = \mathbb{E}_t^{(i)}\left[y_{j,t+h}\right] - y_{j,t+h}.$$

Survey and machine mean squared errors (MSEs) are denoted with $\mathbb{F}$ and $\mathbb{E}$ subscripts, i.e.,

$$\text{survey MSE} \equiv MSE_{\mathbb{F}} = (1/P)\sum_{i=1}^{P}\left(\text{survey error}_{t+h}^{(i)}\right)^2 \tag{6}$$

$$\text{machine MSE} \equiv MSE_{\mathbb{E}} = (1/P)\sum_{i=1}^{P}\left(\text{machine error}_{t+h}^{(i)}\right)^2, \tag{7}$$

[7] We have also implemented the approach in simulated data and hold-out samples for lasso and ridge separately, for random forest, and for empirical Bayes linear regression. The EN estimator was the best performing, followed by lasso, while random forest and Bayesian regression performed poorly.

where P is the length of the forecast evaluation sample. To reduce notation clutter, we leave off superscripts "(i)" in the definitions above, but the reader is reminded that these statistics also depend on the respondent-type. Distortions in survey responses are quantified by the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ over an extended forecast evaluation sample of size P. To uncover any belief distortions, the machine needs to solve a high-dimensional, dynamic, out-of-sample learning problem. We discuss our algorithm for doing so next.

E. *Machine Learning Algorithm*

This section discusses a novel machine learning algorithm developed to detect demonstrable, ex ante expectational errors in real time. This algorithm is explicitly designed to combat overfitting and cope with structural change in a dynamic setting. The full estimation and evaluation procedure involves iterating on the following steps, which are described in greater detail in the online Appendix.

(i) **Sample Partitioning:** At time t, a prior training sample of size $\widetilde{T}$ is partitioned into two subsample windows: an "estimation" subsample consisting of the first T_E observations, and a hold-out "validation" subsample of T_V subsequent observations, i.e., $\widetilde{T} = T_E + T_V$.

(ii) **In-Sample Estimation:** Initial estimates of $\beta^{(i)}$ are obtained with the EN estimator using observations $1, \ldots, T_E$, given an arbitrary fixed (nonrandom) starting value for $\boldsymbol{\lambda}_t^{(i)}$. Denote this initial estimate $\beta_{T_E}^{*(i)}\left(\mathbf{X}_{T_E}, \boldsymbol{\lambda}_t^{(i)}\right)$, where "$*$" denotes the value of the estimator given an arbitrary $\boldsymbol{\lambda}_t^{(i)}$.

(iii) **Cross-Validation:** The regularization parameter $\boldsymbol{\lambda}_t^{(i)}$ is estimated by minimizing mean squared loss $\mathcal{L}\left(\boldsymbol{\lambda}_t^{(i)}, T_E, T_V\right)$ over *pseudo* out-of-sample forecast errors generated from rolling regressions through the validation sample, where

$$\mathcal{L}\left(\boldsymbol{\lambda}_t^{(i)}, T_E, T_V\right) \equiv \frac{1}{T_V - h} \sum_{\tau=T_E}^{T_E+T_V-h} \left(\mathcal{X}_\tau' \beta_{jh,\tau}^{*(i)}\left(\mathbf{X}_{T_E}, \boldsymbol{\lambda}_t^{(i)}\right) - y_{j,\tau+h}\right)^2, \tag{8}$$

and where $\beta_{jh,\tau}^{*(i)}\left(\mathbf{X}_{T_E}, \boldsymbol{\lambda}_t^{(i)}\right)$ is the time τ EN estimate of $\beta_{jh}^{(i)}$ given $\boldsymbol{\lambda}_t^{(i)}$ and data through time τ in a sample of size T_E.

(iv) Steps (i)–(iii) are repeated for new values of $T_E \in \left\{\underline{T}_E, \ldots, \overline{T}_E\right\}$ and $T_V \in \left\{\underline{T}_V, \ldots, \overline{T}_V\right\}$ such that alternative partitions satisfy $T_E + T_V \leq \widetilde{T}$, where shorter window lengths remove consecutive observations at the start of the prior sample. The final machine estimator of $\beta_{jh,t}^{(i)}\left(\mathbf{X}_{T_E}, \boldsymbol{\lambda}_t^{(i)}\right)$ is based on the most recent $\widehat{T}_E$ observations where $\left\{\hat{\boldsymbol{\lambda}}_t^{(i)}, \widehat{T}_E, \widehat{T}_V\right\} = \arg\min_{\lambda^{(i)}, T_E, T_V} \mathcal{L}\left(\boldsymbol{\lambda}_t^{(i)}, T_E, T_V\right)$ and is denoted $\hat{\beta}_{jh,t}^{(i)}\left(\mathbf{X}_{\widehat{T}_E}, \hat{\boldsymbol{\lambda}}_t^{(i)}\right)$.

(v) **Out-of-Sample Prediction:** The values of the regressors at time t are used to make a *true* out-of-sample prediction of y_{t+h}, using $\hat{\beta}_{jh,t}^{(i)}\left(\mathbf{X}_{\widehat{T}_E}, \hat{\lambda}_t^{(i)}\right)$ and the machine forecast error $y_{t+h} - \mathcal{X}_t'\hat{\beta}_{jh,t}^{(i)}\left(\mathbf{X}_{\widehat{T}_E}, \hat{\lambda}_t^{(i)}\right)$ stored.

(vi) **Roll Forward and Repeat:** The prior sample of data is rolled forward one period, and steps (i)–(v) are repeated. This continues until the last out-of-sample forecast is made for $y_{j,T}$, where T is the last period of our sample.

Referring back to the notation in (6) and (7), $MSE_{\mathbb{E}}$ is computed by averaging across the sequence of squared forecast errors in the true out-of-sample forecasting step (v) for periods $t = \left(\widetilde{T} + h\right), \ldots, T$. We refer to this subperiod as the external forecast *evaluation sample*.

Several points about the procedure above bear emphasizing. First, the algorithm ensures that the machine forecast selected from step (iv) can only differ from the survey forecast if it demonstrably improves pseudo out-of-sample prediction in the rolling training samples *prior* to making a true out-of-sample forecast in step (v). Otherwise, the machine adopts the survey forecast. It follows that the true out-of-sample forecasts of the machine recorded in step (v) can differ from those of the survey only if demonstrable, ex ante biases are detected. The resulting measure of belief distortion therefore explicitly excludes *ex post mistakes* that the machine algorithm could only have understood with hindsight. An implication of this ex ante approach is that more than one type can show no bias if the machine is unable to detect patterns in extraneous economic data that can be exploited in real time to improve forecasts. We quantify the overall magnitude of forecaster bias with the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ taken over the evaluation sample.

Second, the machine algorithm is repeated for each i and for each t in the evaluation sample. This can be important for all the parameter estimates but especially so for the estimate of the intercept, which functions as a latent time-varying mean.

Third, each new training renews the optimized selection of in-sample estimation and validation sample windows lengths, innovating on traditional machine learning approaches by paying close attention to the time structure of data. This is important because, in a dynamic setting subject to possible structural change, no single set of window lengths is likely to work best in all time periods.[8] The algorithm developed here instead asks the machine to choose both the optimal estimation window and the optimal validation window for determining shrinkage and sparsity, dynamically. We discuss this further below.

Fourth, the cumulation of true out-of-sample forecast errors from step (v) serves as an external validation step that exists outside of the optimization loop. It is crucial that the number of forecast error observations aggregated from this step be large enough so that the evidence on relative forecast accuracy is not the result of a few random outliers. For this reason, we require that our external evaluation samples from this step be at least 84 quarters long for all machine specifications.

[8] See Giacomini and White (2006) and Pesaran and Timmermann (2007) for extensive evidence related to these themes.

At the same time, we must balance this imperative against two others: the need to reserve a minimum number of observations to do estimation and training, and, because there are differences in data availability across the surveys, the need to compare relative forecast using roughly comparable time frames. For the SPF, data on inflation and GDP growth forecasts are available from 1969:III to 2018:III. Thus the machine can partition a prior sample up to size $\widetilde{T} = 98$ quarters in every step of the recursion and still produce an external evaluation sample with 97 quarterly observations that spans 1995:I to 2018:II. In this case, the first true out-of-sample forecast of four-quarter-ahead outcomes is recorded for the period 1994:I to 1995:I. For SOC, both inflation and GDP growth forecast data are available from 1978:I to 2018:II, while for BC, inflation forecasts are available from 1986:I to 2018:II, and GDP forecasts are available from 1984:III to 2018:II. Since these surveys extend less far backward in time, estimation and training must be accomplished on smaller prior sample sizes $\widetilde{T}$ in the early recursions in order to ensure an external evaluation sample of at least 84 observations. But it is still possible to allow for reasonable minimal prior sample sizes at each step in these recursions, while nonetheless ending up with external evaluation samples of at least 84 quarters that cover roughly comparable time periods. Our external evaluation sample for the SOC surveys consists of 97 quarterly observations and span 1995:I to 2018:II. For the BC surveys the external evaluation sample for inflation forecasts consists of 84 quarterly observations and spans 1997:III to 2018:II, while that for the GDP growth forecasts consists of 89 quarterly observations and spans 1996:I to 2018:II.

Our approach of dynamically selecting estimation and cross-validation window lengths to minimize the pseudo MSEs merits further discussion. As noted previously, it is crucial for our investigation that both of these specification choices be made on an ex ante basis. This presents offsetting considerations. On the one hand, in a dynamic setting subject to structural change, no single value for T_E or T_V is likely to work best for all time, so it seems important to allow for some flexibility. On the other hand, allowing completely free reign in the choice of window lengths could cause the ranking of the forecasting specifications based on pseudo MSEs to be so affected by random variations that you'd often choose a specification that breaks down out of sample. We address these trade-offs in two ways. First, we allow for a limited degree of choice across five different window lengths for the validation step and ten different window lengths for estimation. The precise grids (in quarters) are $T_V = \{12, 16, 20, 24, 28\}$ and $T_E = \{24, 28, 32, 36, 40, 44, 48, 52, 56, 60\}$. Second, we always use the last T_V observations to do the pseudo out-of-sample forecast evaluation, implying that each of our validation subsamples has the same end point $\widetilde{T}$ immediately preceding the date over which the machine makes a true out-of-sample forecast. This means that there is substantial overlap in the observations that make up the different validation subsamples of length T_V. This overlap further limits the potential for highly random variations that can arise from choosing T_V to minimize the pseudo MSE.[9] In important earlier work, Pesaran

[9] This may be contrasted with traditional "K-fold" cross-validation techniques, which partition a sample randomly, leading to tuning parameters that are chosen partly on the basis of how well the future predicts the past. Table A2 of the online Appendix shows that the machine performs poorly when it is trained using traditional K-fold cross-validation techniques.

and Timmermann (2007) propose choosing the estimation window length T_E by minimizing the pseudo MSE over the validation sample, as we do, but unlike our approach they fix T_V. Unfortunately, neither econometric theory nor Pesaran and Timmermann (2007) offer any practical guidance on how to choose T_V on an ex ante basis, and in econometric practice the actual choices appear to be somewhat arbitrary. The procedure we propose, of allowing a limited degree of flexibility in the selection of both windows, can be executed in practice by exploiting prior samples of data to determine judicious grids for the window lengths. This training could be implemented repeatedly over time as it evolves.

F. *Switching Model*

Since at least the groundbreaking contribution of Hamilton (1989), it has been well known that aggregate output growth is well described by a process that evolves differently across distinct economic states associated with recessions and expansions. A large subsequent literature treats models with regime changes as part of the standard forecasting toolbox for output growth (e.g., Chauvet and Potter 2013). Furthermore, by the mid-1990s, a large body of evidence had accumulated that the slope of the term structure of interest rates had strong predictive power for the US business cycle. Specifically, inversions or a flattening of the yield curve typically anticipate a sharp downturn in economic growth. By the mid-1990s it had become a well established practice of many professional forecasters to switch to simpler forecasting specifications for economic growth that focused almost exclusively on yield spread information whenever the term structure was flat or downward sloping.[10] By contrast, there was much less evidence that term spreads were useful in forecasting inflation at any time in the business cycle.

Putting this all together, it is clear that a forecaster operating in the mid-1990s would have had access to a large body of evidence indicating that (i) output growth behaves differently in recessions than in expansions, and (ii) turning points are often anticipated by a flat or inverted yield curve. Based on this prior knowledge, the machine efficient benchmark is specified to follow a simple switching model for output growth—but not inflation—for forecasts starting in 1995:I.

To implement this idea—for forecasting GDP growth only—we combine the notion of distinct regimes with the predictive power of the term structure slope using a threshold model. This feature allows the machine to choose in real time whether to switch to a simpler, recession specification. The threshold aspect of the GDP growth forecasting specifications works as follows:

$$\begin{aligned}
\Delta gdp_{j,t+h} &= \alpha_{jh}^{(i)} + \beta_{jh\mathbb{F}}^{(i)} \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] + \mathbf{B}_{j\mathcal{Z}}^{(i)'} \mathcal{Z}_{jt} + \epsilon_{jt+h} && \text{if } slope_{kt} > \widehat{tr}_{kt}, \\
\Delta gdp_{j,t+h} &= B_{kt} I_{kt} && \text{if } slope_{kt} \leq \widehat{tr}_{kt},
\end{aligned}$$

where B_{kt} is a parameter and I_{kt} is a dummy variable that depends on a yield curve measure at time t.

[10] See Harvey (1989); Estrella and Hardouvelis (1990, 1991); Plosser and Rouwenhorst (1994); Haubrich and Dombrosky (1996); Kozicki (1997); Dotsey (1998); and Estrella and Mishkin (1998).

At each point in time, the machine chooses whether to use the "normal-times" specification (first row) or the "recession" specification (second row). The normal-times specification is based on the machine algorithm discussed in the previous section. The recession specification is chosen whenever $slope_{kt} < \widehat{tr}_{kt}$, where $slope_{kt}$ is a yield spread measure at time t and $\widehat{tr}_{kt}$ is a threshold. We consider three different yield curve slope indicators, indexed by k: the 10-year minus 3-month Treasury spread $(10y3m)$, the 10-year minus 1-year Treasury spread $(10y1y)$, and the 10-year minus 2-year Treasury spread $(10y2y)$. The machine uses past data to run a regression of GDP growth h quarters ahead (Δgdp_{t+h}) on a dummy variable I_{kt} that equals one when $slope_{kt} \leq tr_{kt}$ and zero otherwise. The machine searches in real time for the specific threshold $\widehat{tr}_{kt}$ and the yield spread indicator $slope_{kt}$ that maximizes the R^2 of this regression. The machine repeatedly reoptimizes the choice of both $\widehat{tr}_{kt}$, and the specific measure of the term spread $(10y3m$, $10y1y$, or $10y2y)$ based on real-time forecasting regressions of GDP growth on I_{kt} using expanding windows of data up to time t and beginning in 1976:III, when the data on the 2-year Treasury bill rate is first available. The time t recession specification forecast of GDP growth in $t + h$ is then simply the average GDP growth over all periods where $I_t = 1$ in a sample spanning 1976:III to t.

Figure 2 reports the real-time R^2s from the expanding-window regressions of GDP growth at $t + h$ on I_{kt} using the different measures of the term spread. The figure shows that regressions using the $10y2y$ dummies are almost always chosen by the machine because that specification delivers the highest real-time predictive power for GDP growth in almost all periods of our evaluation sample, including those just prior to the two recessions in the sample that occurred in 2001 and 2007–2009.

It is worth noting, however, that the only time over the evaluation samples that the recession specification is triggered is just prior to the 2001 recession. In particular, there is no switch triggered prior to the Great Recession. This happens because all term spreads exhibited a secular decline between the two recessions, so that by 2007 the threshold values commensurate with a forecast of negative economic growth had also declined. At the same time, the declines in yield spreads prior to the 2007–2009 recession were relatively modest by historical standards and thus never fell below the (lower) real-time $\widehat{tr}_{kt}$ thresholds. In contrast to the 2001 recession, yield spreads generally failed to signal the Great Recession, a structural shift that shows up in Figure 2 as a sharp drop in real-time R^2 statistics right after the recession.

III. Data

The data used for this study fall into several categories. For each category the sources and details are left to the online Appendix.

Survey Data.—The first data category is the survey data.

The SPF is a quarterly survey. Respondents provide both nowcasts and quarterly forecasts from one to four quarters ahead. We focus on the survey questions about the level of the GDP deflator (PGDP) and the level of real GDP. We use these data to construct forecasts of GDP growth, as explained in the online Appendix. We also use SPF forecasts of ten-year-ahead consumer price index (CPI) inflation as information variables.

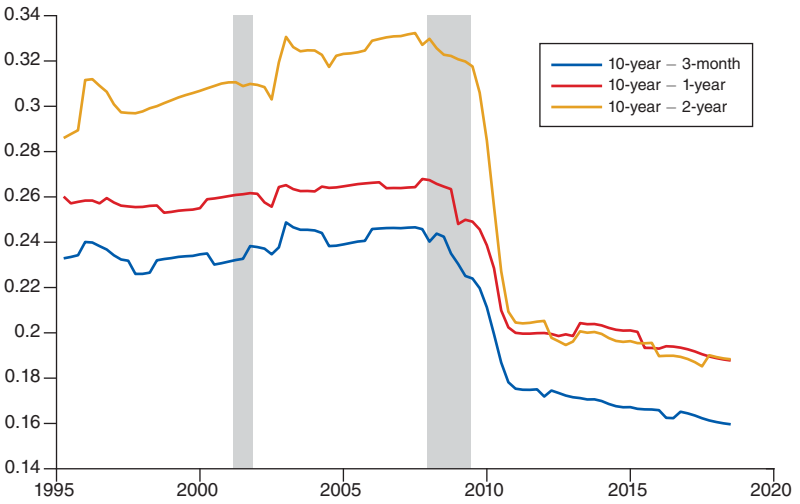

FIGURE 2. REAL-TIME R^2 WITH OPTIMAL THRESHOLDS

Notes: This figure reports the R^2s from expanding-window regressions of GDP growth at t on a dummy variable I_{t-h}. The dummy variable I_t equals one if the term spread is below its real-time N_tth percentile at time t. The threshold N_t maximizes the R^2 from the regression estimated at time t. The real-time percentile of the term spread at time t is computed using the data from 1976:III to time t. The yellow line reports the results using term spread defined as 10-year Treasury bond rate minus 2-year rate. The red line uses 10-year rate minus 1-year rate. The blue line uses 10-year rate minus 3-month rate. NBER recessions are shown with gray shaded bars. The sample is 1995:I–2018:II.

The SOC asks households directly about inflation, and we use the questions on whether households expect prices to go up or down during the next 12 months, and by how much, to gauge their expectations about inflation. Following Curtin (2019), we take these forecasts to be most relevant for annual CPI inflation, and therefore compare SOC forecasts to actual outcomes for CPI inflation. Since the SOC doesn't directly ask about GDP growth, we take the approach discussed in Curtin (2019) which is based on responses to question A7 in the SOC: "About a year from now, do you expect that in the country as a whole business conditions will be better, or worse than they are at present, or just about the same?" This qualitative economic forecast is converted to a point forecast for GDP growth by fitting a regression of future GDP growth data to the balance score for question A7 (percent respondents expect economy to improve – percent expect worsen + 100) using rolling regressions and real-time GDP data.

For the BC survey, we use questions in which forecasters are asked to predict the average quarter-over-quarter percentage change in Real GDP and the GDP deflator, beginning with the current quarter and extending four to five quarters into the future.

For all surveys, we align the timing of survey response deadlines with real-time data, so that the machine can only use data available in real time before the survey deadline.

Real-Time Macro Data.—A real-time macro dataset provides observations on the left-hand-side variables on which forecasts are formed obtained from the

Federal Reserve Bank of Philadelphia's Real-Time Dataset. Following Coibion and Gorodnichenko (2015), to construct forecasts and forecast errors, we use the vintage of inflation and GDP growth data that are available four quarters after the period being forecast. We also use the real-time macro data to form real-time quarterly macro factors from a constructed dataset of real-time quarterly macro variables observed on or before the day of the survey deadline at each date t. The resulting real-time macro dataset, contains observations on 92 real-time macro variables. Our real time macro variable dataset also include data on home and energy prices, which are not revised and so do not have multiple vintages. The complete list of macro variables is given in the online Appendix.

Monthly Financial Data.—To take into account financial market data, we form factors from a panel dataset of 147 monthly financial indicators that include valuation ratios, growth rates of aggregate dividends and prices, default and term spreads, yields on corporate bonds of different ratings grades, yields on Treasuries and yield spreads, and a broad cross section of industry equity returns. We convert the monthly factors formed from the dataset into quarterly factors by using the first month's observation for each quarter.

Daily Financial Data.—"Up-to-the-forecast" financial market information is accounted for by using daily data on financial indicators up to one day before the survey respondents forecasts are due. The daily financial dataset includes series from five broad classes of financial assets: (i) commodities prices, (ii) corporate risk variables including a number of different credit spreads measuring default risk, (iii) equities, (iv) foreign exchange, and (v) government securities. In total, we use 87 such series, 39 commodity and futures prices, 16 corporate risk series, 9 equity series plus implied volatility, 16 government securities, and 7 foreign exchange variables), with the complete set of variables reported in the online Appendix. In order to use both daily and quarterly data in our estimation, we combine diffusion index estimation of daily financial factors with mixed data sampling frequency techniques, described in detail in the online Appendix.

Additional Nonfactor Data.—A number of other nonfactor variables are also included in the machine model in $\mathbf{W}_{jt}'$. These include the ith percentile's own nowcast for the variable being forecast, lags of the ith percentile's own forecasts and those of other percentiles, higher-order cross-sectional moments of the lagged forecast distributions, several autoregressive lags of the left-hand-side variables, long-term trend inflation measures, and measures of detrended employment and GDP (Hamilton 2018).

In all, once factors are formed the machine model entertains a total of 68 predictor variables for inflation and 72 predictor variables for GDP growth, before the machine chooses sparsity. We refer below to estimated factors with an economic name. The economic name makes use of group classifications for individual series and output from time series regressions of individual series onto estimated factors, for each time period in our evaluation sample. For example, if regressions of nonfarm payrolls onto the first common macro factor from the real-time macro panel dataset exhibits the highest average (across all time periods of our evaluation

sample) marginal R^2, then that factor is labeled an "employment" factor and normalized so that it increases when nonfarm payrolls increase.

IV. Results

This section reports results using our estimates of belief distortions across different respondent-types, surveys, and variables. In all cases, we focus on $h = 4$ quarter-ahead forecasts.

A. *Forecast Comparison*

We present a comparison of the accuracy of forecasts made by the machine benchmark and the survey respondents over the external evaluation samples. Table 1 reports the ratio of the machine $MSE_{\mathbb{E}}$ to the survey $MSE_{\mathbb{F}}$ for inflation and GDP growth for all three surveys over their respective external evaluation samples, along with several other results. We discuss these in turn.

First consider the average predictive accuracy of the machine versus the forecaster-type over the external evaluation sample. The top panel of Table 1 shows that the machine model performs better than the survey forecasts of inflation for all surveys and all respondent-types as measured by the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$, which is less than one in all cases, sometimes by large amounts. To put this ratio in the same units as an in-sample R^2, the table also reports an out-of-sample R^2 for the machine vis-à-vis the survey as $R^2_{OOS} \equiv 1 - MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$. The overall magnitude by which the machine model improves on the survey forecasts is in most cases sizable, which is notable since survey forecasts of inflation are known to be difficult to beat or even match by statistical models out-of-sample, as discussed above. For example, the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ for the median SPF forecast is 0.85. These ratios are similar for the median BC survey, as shown in the last panel, where in this case $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ is 0.84. In general, the magnitude of measured belief distortions about future inflation is much larger for SOC respondents than for the SPF and BC respondents, as shown in the middle panel. The SOC median $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ ratio is 0.62, respectively, implying large out-of-sample R^2 statistics.[11]

For GDP growth, the lower panel of Table 1 shows that machine model is again always more accurate than the survey respondent-type no matter which respondent-type or survey is studied. The $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ ratios for the median SPF and BC forecasts of GDP growth are 0.88 and 0.87, respectively. For the SOC, there is only a single forecast, denoted as if it corresponds to the "median" household, since the SOC forecast is constructed from the balance score for business conditions expectations, eliminating the heterogeneity (see above). The $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ for this single SOC forecast of GDP growth is 0.78.

Table 2 shows that the biggest gains in forecast accuracy afforded by the machine algorithm over the human forecasters occurs in more recent data, toward the end of

[11] Table A1 in the online Appendix shows that the machine also improves over the mean of the survey forecasts. We do not report those results here because the mean is always an amalgam that does not correspond to the belief of any single respondent-type in the survey and would not be known to any individual. It is arguably less relevant to the study of what, if any, systematic errors individuals may make when forming macroeconomic expectations.

TABLE 1—MACHINE LEARNING VERSUS SURVEY FORECASTS

ML: $y_{j,t+h} = \alpha_{jh}^{(i)} + \beta_{jh\mathbb{F}}^{(i)} \mathbb{F}_t^{(i)}[y_{j,t+h}] + \mathbf{B}_{jh\mathcal{Z}}^{(i)} \mathcal{Z}_{jt} + \epsilon_{jt+h}$

Inflation forecasts

Percentile:	Median	5th	10th	20th	25th	30th	40th
Survey of Professional Forecasters (SPF)							
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.85	0.56	0.74	0.83	0.90	0.88	0.89
OOS R^2	0.15	0.44	0.26	0.17	0.10	0.12	0.11
w^*	0.68	0.82	0.78	0.74	0.63	0.66	0.66
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.82	0.41	0.64	0.77	0.81	0.82	0.84
	60th	70th	75th	80th	90th	95th	
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.74	0.70	0.67	0.59	0.55	0.47	
OOS R^2	0.26	0.30	0.33	0.41	0.45	0.53	
w^*	0.78	0.76	0.74	0.80	0.79	0.83	
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.76	0.66	0.61	0.53	0.39	0.27	

Percentile:	Median	5th	10th	20th	25th	30th	40th
Michigan Survey of Consumers (SOC)							
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.62	0.22	0.27	0.45	0.58	0.69	0.70
OOS R^2	0.38	0.78	0.73	0.55	0.42	0.31	0.30
w^*	1.00	0.96	0.93	0.92	0.92	0.95	1.00
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.62	0.11	0.20	0.42	0.52	0.64	0.76
	60th	70th	75th	80th	90th	95th	
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.40	0.22	0.16	0.13	0.05	0.03	
OOS R^2	0.60	0.78	0.84	0.87	0.95	0.97	
w^*	1.00	1.00	1.00	1.00	1.00	1.00	
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.37	0.21	0.16	0.11	0.04	0.02	

Percentile:	Median	5th	10th	20th	25th	30th	40th
Blue Chip Financial Forecasts (BC)							
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.84	0.58	0.60	0.85	0.85	0.86	0.91
OOS R^2	0.16	0.42	0.40	0.15	0.15	0.14	0.09
w^*	0.65	0.73	0.76	0.62	0.63	0.62	0.58
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.76	0.45	0.55	0.72	0.76	0.78	0.78
	60th	70th	75th	80th	90th	95th	
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.78	0.69	0.65	0.59	0.48	0.38	
OOS R^2	0.22	0.31	0.35	0.41	0.52	0.62	
w^*	0.70	0.79	0.82	0.86	0.94	0.92	
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.73	0.66	0.62	0.57	0.43	0.33	

(*continued*)

our forecast sample. The table compares the accuracy of the machine forecast to that of the median forecast for each survey over the last five years of our external forecast sample, from 2013:II to 2018:II. For GDP growth, the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ for the median SPF forecast is 0.82 over this subperiod, while it is 0.67 for median BC forecast. For inflation, the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ over this same subperiod is 0.63 for the median SPF forecast, 0.67 for the median BC forecast, and 0.37 for the median SOC forecast. That the machine does better at the end rather than the beginning of the sample is of interest, since it suggests that bounded rationality in the form of

Table 1—Machine Learning versus Survey Forecasts (*continued*)

ML: $y_{j,t+h} = \alpha_{jh}^{(i)} + \beta_{jh\mathbb{F}}^{(i)} \mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] + \mathbf{B}_{jh\mathcal{Z}}^{(i)} \mathcal{Z}_{jt} + \epsilon_{jt+h}$

GDP forecasts

Percentile:	Median	5th	10th	20th	25th	30th	40th
Survey of Professional Forecasters (SPF)							
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.88	0.70	0.81	0.80	0.84	0.87	0.88
OOS R^2	0.12	0.30	0.19	0.20	0.16	0.13	0.12
w^*	0.83	0.96	1.00	1.00	1.00	0.94	0.87
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.87	0.74	0.83	0.88	0.89	0.89	0.88

	60th	70th	75th	80th	90th	95th
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.85	0.81	0.79	0.80	0.69	0.64
OOS R^2	0.15	0.19	0.21	0.20	0.31	0.36
w^*	0.85	0.88	0.87	0.83	0.85	0.84
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.84	0.81	0.79	0.77	0.67	0.58

Percentile:	Median
Michigan Survey of Consumers (SOC)	
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.78
OOS R	0.22
w^*	0.81

Percentile:	Median	5th	10th	20th	25th	30th	40th
Blue Chip Financial Forecasts (BC)							
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.88	0.77	0.74	0.89	0.83	0.82	0.78
OOS R^2	0.13	0.23	0.26	0.11	0.17	0.18	0.22
w^*	0.64	0.68	0.75	0.60	0.73	0.71	0.76
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.75	0.70	0.76	0.79	0.79	0.78	0.76

	60th	70th	75th	80th	90th	95th
$MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$	0.78	0.79	0.75	0.71	0.67	0.67
OOS R^2	0.22	0.21	0.25	0.29	0.33	0.33
w^*	0.72	0.71	0.72	0.77	0.77	0.74
$MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$	0.73	0.70	0.69	0.66	0.60	0.55

Notes: The machine and survey mean squared forecast errors for four-quarter-ahead forecasts, averaged over the evaluation sample are denoted by $\text{MSE}_{\mathbb{E}}$ and $\text{MSE}_{\mathbb{F}}$, respectively. The out-of-sample R^2, OOS R^2, is defined as $1 - \text{MSE}_{\mathbb{E}}/\text{MSE}_{\mathbb{F}}$. The weight on the machine forecast for the hybrid forecast described in the text is denoted by w^*. The MSE from the full information, rational expectations (FIRE) benchmark described in the text is denoted by $\text{MSE}_{\mathbb{R}}$. The vintage of observations on the variable being forecast is the one available four quarters after the period being forecast. The evaluation period for the Survey of Professional Forecasters (SPF) and the Michigan Survey of Consumers (SOC) is 1995:I to 2018:II; and for the Blue Chip (BC) survey is 1997:III to 2018:II (inflation) and 1996:I to 2018:II (GDP).

limitations on the human capacity for collecting and processing large amounts of information are unlikely to fully explain our findings. By 2013, at least professional forecasters would have had both the resources and the capacity to take advantage of advances in information-processing technology and computing power.

While a comparison of mean squared forecast errors is one sensible way to evaluate predictive accuracy, researchers sometimes consider other aspects of the forecast environment. For example, there is often interest in characterizing uncertainty around the predictive accuracy of two models in statistical terms, and frequentist

TABLE 2—MACHINE LEARNING VERSUS SURVEY FORECASTS

ML: $y_{j,t+h} = \alpha_{jh}^{(i)} + \beta_{jh\mathbb{F}}^{(i)} \mathbb{F}_t^{(i)}[y_{j,t+h}] + \mathbf{B}_{jh\mathcal{Z}}^{(i)} \mathcal{Z}_{jt} + \epsilon_{jt+h}$

Median inflation forecasts, $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$

SPF		SOC		BC	
1995:I–2018:II	2013:II–2018:II	1995:I–2018:II	2013:II–2018:II	1997:III–2018:II	2013:II–2018:II
0.85	0.63	0.62	0.37	0.84	0.67

Median GDP forecasts, $\text{MSE}_{\mathbb{E}}/\text{MSE}_{\mathbb{F}}$

SPF		SOC		BC	
1995:I–2018:II	2013:II–2018:II	1995:I–2018:II	2013:II–2018:II	1996:I–2018:II	2013:II–2018:II
0.88	0.82	0.78	0.71	0.87	0.67

Notes: Machine versus survey mean squared forecast errors. The machine and survey mean squared forecast errors, for four-quarter-ahead forecasts, averaged over the evaluation sample are denoted by $\text{MSE}_{\mathbb{E}}$ and $\text{MSE}_{\mathbb{F}}$, respectively. The vintage of observations on the variable being forecast is the one available four quarters after the period being forecast.

econometric tests are sometimes employed in this service. While this is a laudable goal for some questions, we argue that such tests are less useful or relevant when the objective is to measure belief distortions in survey *point* forecasts, as here. In this case, forecaster bias is revealed by any instance in which forecasters choose point forecasts that—on the basis of ex ante information—demonstrably fail to minimize the loss function. For example, a professional forecaster might have two models that produce economically large differences in mean squared forecast error, while statistical tests can often indicate that they are the "same." The question then becomes, what should an unbiased point forecaster do? If mean squared loss is the objective, there is only one optimal choice, and this is unaffected by the amount of sampling noise around her two model forecasts.

Of course, measures of uncertainty about forecast point forecasts are inherently interesting for reasons other than forecaster bias, such as when we want to characterize our overall confidence in any prediction or predictions. But even for this purpose, tests is that they return merely a binary answer on whether a null hypothesis is rejected or not—while being silent on the practical quantitative question of by how much one model is more accurate than another—would seem to be of limited utility.

For these reasons we take an alternative approach to characterizing uncertainty, one that allows us to quantify any gains in forecast accuracy on a continuum from low to high, without being affected by the overall amount of statistical noise in the environment that both the machine and the survey forecast are subject to. The approach is motivated by the work of Amisano and Geweke (2017), who consider the properties of weighted linear combinations of prediction models. The key idea is that even if one model has superior predictive power over others, an optimal linear combination typically includes several models with positive weights, since being better on average is not synonymous with being always better. Amisano and Geweke (2017) focus on density forecasts, but given that survey responses are point forecasts rather than density forecasts, we adapt their idea by solving for the optimal linear combination of the machine and survey forecasts that minimizes the mean square forecast error over our evaluation sample. We refer to this linear combination as the optimal "hybrid" forecast.

Specifically, consider a hybrid forecast of $y_{j,t+h}$, denoted $\mathbb{EF}_t^{(i)}\left[y_{j,t+h}\right]$, obtained as a weighted average of the machine and the survey forecasts:

$$\mathbb{EF}_t^{(i)}\left[y_{j,t+h}\right] \equiv w\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right] + (1-w)\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right],$$

where $w \in [0,1]$. Conceptually we can ask, given the average performance of these two forecasts over our sample, how much weight w would one want to place on one versus the other in a hybrid forecast if we faced an identical sample in the future? To answer this question, note that the hybrid forecast errors are a linear combination of the machine and survey forecast errors:

$$\begin{aligned}\text{hybrid error}_{t+h}^{(i)} &= \mathbb{EF}_t^{(i)}\left[y_{j,t+h}\right] - y_{j,t+h} \\ &= w\mathbb{E}_t^{(i)}\left[y_{j,t+h}\right] + (1-w)\mathbb{F}_t^{(i)}\left[y_{j,t+h}\right] - y_{j,t+h} \\ &= w\left(\text{machine error}_{t+h}^{(i)}\right) + (1-w)\left(\text{survey error}_{t+h}^{(i)}\right),\end{aligned}$$

with the hybrid mean squared forecast error given by

$$\text{hybrid MSE} \equiv MSE_{\mathbb{EF}} = (1/P)\sum_{i=1}^{P}\left(\text{hybrid error}_{t+h}^{(i)}\right)^2.$$

The optimal weight w^* placed on the machine forecast is defined as the one that minimizes the hybrid MSE over our evaluation samples, i.e.,

$$w^* = \arg\min MSE_{\mathbb{EF}} = \arg\min(1/P)\sum_{i=1}^{P}\left(\text{hybrid error}_{t+h}^{(i)}\right)^2,$$

where P is the length of the evaluation sample.

The weights w^* are reported in the third row of each subpanel in Table 1. To interpret these numbers, note that if the machine were always better than the survey, w^* would be 1. This happens with many percentiles of the SOC inflation forecasts, and in several percentiles of the SPF GDP growth forecasts. If instead the machine is only marginally better than the survey, w^* would be close to 0.5. For the median GDP growth forecasts, the weights are well above 0.5, equal to 0.83, 0.81 and 0.64 for the SPF, SOC and BC median forecasts, respectively. The analogous weights for the median inflation forecasts are 0.68, 1, and 0.65, and typically higher for the other percentile types. These relatively large numbers close to unity imply that the machine produced economically meaningful gains in forecast accuracy over the survey responses during the historical sample over which the two forecasts were separately evaluated.

Finally, in the last rows of Table 1 we report the results of a different type of model comparison. Specifically, we compare the accuracy of survey forecasts with that from an alternative machine specification that differs from our baseline machine specification along only one dimension: it uses every percentile-type's time t forecast rather than just the ith percentile's. This alternative benchmark is motivated by certain imperfect information models in which every agent in the economy receives a private signal, but other agents' private signals are not publicly known. In such

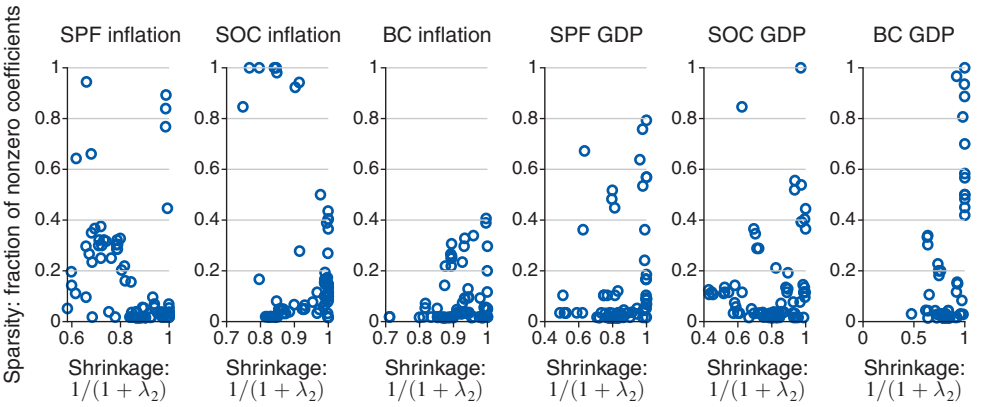

FIGURE 3. DEGREE OF SPARSITY AND SHRINKAGE

Notes: The figure displays a scatterplot of the strength of the ridge and least absolute shrinkage and selection operator (LASSO) penalties estimated from training samples over time for predicting median inflation or real GDP growth. For each observation in the evaluation sample from 1995:I–2018:II (94 observations), the *y*-axis displays the degree of sparsity implied by the estimated L_1 penalty, λ_1, in units of the fraction of nonzero regression coefficients, and the *x*-axis displays the degree of shrinkage implied by the estimated L_2 penalty, λ_2 in units of $1/(1+\lambda_2)$.

models, the full information, rational expectations (FIRE) benchmark against which any distortion is measured is based on the union of everyone's information at time t.[12] The last rows of each panel in Table 1 report the ratio of the mean squared forecast errors under this FIRE benchmark, denoted $MSE_{\mathbb{R}}$, to the survey $MSE_{\mathbb{F}}$. Comparing the ratio $MSE_{\mathbb{R}}/MSE_{\mathbb{F}}$ with those in the first row showing the baseline $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$, we see that the numbers are quite similar, and in some cases the latter ratio is a bit smaller than the former. This shows that the improvement in forecast accuracy afforded by allowing the machine benchmark to observe everyone's time t prediction—where it exists at all—is minimal. This finding is relevant because it suggests that information frictions based on noisy "dispersed information" are unlikely to be the most relevant source of belief distortion in our data.

Returning to the baseline specifications, it is of interest to consider the nature of the empirical specifications chosen by the machine that produce gains in forecast accuracy. Figure 3 reports a scatterplot that quantifies the strength of the estimated ridge and LASSO penalties, with each point representing a combination of the two penalties chosen for one time period of the evaluation sample. The *y*-axis displays the degree of sparsity implied by the L^1 (LASSO) penalty, as measured by the fraction of nonzero coefficients. The *x*-axis displays the degree of shrinkage implied by the L^2 (ridge) penalty, as measured by $1/\left(1+\widehat{\lambda}_{2,t}\right)$, where $\widehat{\lambda}_{2,t}$ is the estimated ridge penalty parameter for period t. The right border of the plot is the case where there is no ridge penalty at all, while the top edge of the plot is the case where there is no LASSO penalty. We see that the machine algorithm often results in a sparse specification. In many time periods the fraction of non-zero coefficients hovers around 10 percent or less, though in some periods the machine chooses little if any

[12] We thank an anonymous referee for suggesting this alternative comparison.

sparsity, but much greater L^2 shrinkage. Occasionally, the machine chooses minimal sparsity and minimal L^2 shrinkage. This implies that achieving the efficiency gains of the machine over the extended evaluation sample requires entertaining large datasets in every period, even though much of that information is associated with a coefficient that is shrunk all the way to zero most of the time.

B. *Dynamics of Belief Distortions*

We now turn to investigate the dynamics of systematic expectational errors, by reporting the median bias over time, i.e., $bias_{j,t}^{(50)} \equiv \mathbb{F}_t^{(50)}\left[y_{j,t+h}\right] - \mathbb{E}_t^{(50)}\left[y_{j,t+h}\right]$ over our evaluation sample. Note that the units of $bias_{j,t}^{(50)}$ are the same as the forecasts themselves and are in annual percentage points. Figure 4 shows biases associated with the mean and median respondents for all three surveys.

Figure 4 shows that systematic errors in the median forecasts vary substantially over time and can range between 50 percent and 400 percent of the average annual inflation or GDP growth, depending on the survey. Survey forecasts for GDP growth oscillate between "optimism" and "pessimism," a finding reminiscent of learning models that feature extended waves of optimism and pessimism (e.g., Eusepi and Preston 2011). For GDP growth the figure shows extended periods of overoptimism that are especially prevalent for professional forecasters in the post-Great Recession part of our subsample. From 2010:I to 2018:II, the median SPF forecast of GDP growth is biased upward by 0.88 percent at an annual rate, or 39 percent of actual GDP growth during this period. This large upward bias since 2010 contributes heftily to the upward bias over the full evaluation sample (1995:I–2018:II), which is also sizable and amounts to 19 percent of observed GDP growth. These distortions are quite similar for the median BC expectation of GDP growth. For the SOC, the *average* bias is close to zero even though the SOC forecast is less accurate than the SPF or BC forecasts. This happens because the SOC forecast makes systematic errors of greater magnitude that fluctuate more wildly between optimism and pessimism. For all surveys, there are large spikes in the biases at the cusp of the 2000–2001 recession, a finding we discuss further below.

For inflation, Figure 4 shows that the median expectations are biased upward (a direction we defined above as "pessimistic") over most of the sample for the SPF and the SOC, while the BC survey exhibits an average bias that is close to zero.[13] Despite being upwardly biased on average over the full sample, median inflation forecasts exhibit a downward bias from 2011 to 2014 that ranges across surveys from −0.08 percent to −0.82 percent at an annual rate, or −4.3 percent to −47 percent of actual inflation during this period. Given that inflation has been declining over time, this could be interpreted as evidence of a learning process.

[13] Whether an upward bias in inflation expectations should be viewed as pessimism or optimism may depend on the time period. Bhandari, Borovicka, and Ho (2019) argue that a general interpretation of higher expected inflation as optimism is at odds with surveys of inflation attitudes, but others have argued that a downward bias in inflation expectations could be interpreted as pessimism during specific episodes, such as when nominal interest rates are at the zero lower bound (Masolo and Monti 2021). We use "pessimistic" as a shorthand labeling device for upwardly biased inflation expectations, regarding the interpretation as roughly right for households in most time periods.

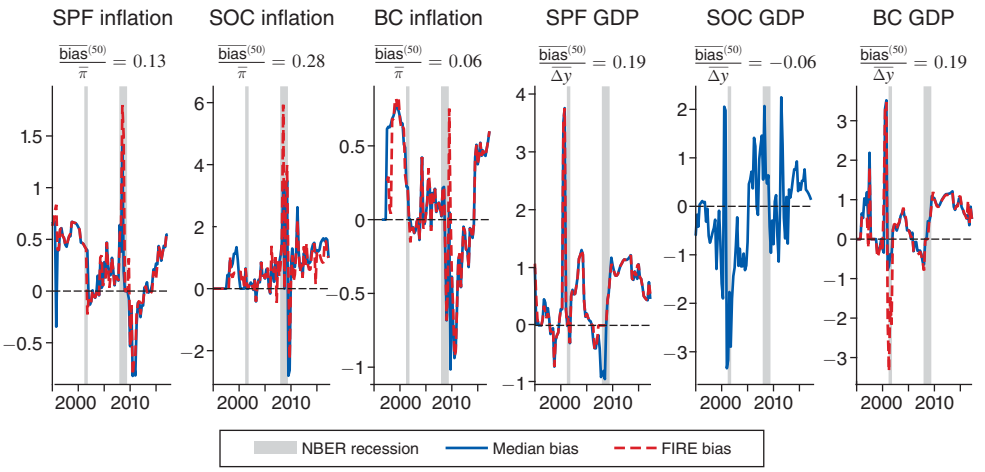

FIGURE 4. BIASES IN THE MEDIAN SURVEY FORECAST

Notes: The figure reports the time series $bias^{(i)}_{j,t+h} = \mathbb{F}^{(i)}_t[y_{j,t+h}] - \mathbb{E}^{(i)}_t[y_{j,t+h}]$ for $i = 50$; *mean*. NBER recessions are shown with gray shaded bars. The sample spans the period 1995:I–2018:II.

Figure 4 also shows clearly the reason for the big relative gains in forecast accuracy afforded by the machine algorithm during the last five years of our forecast sample, from 2013:II to 2018:II, as documented in Table 2. Professional forecasters and households alike underperformed over this subperiod because they repeatedly overpredicted both economic growth and inflation.

Finally, Figure 4 plots the estimated biases in the median survey forecasts as measured against the machine FIRE benchmark discussed above, which uses the union of everyone's information at time t, in addition to the extensive public information (red dashed line). With the exception of a few outlier observations, deviations of the median forecast from this FIRE benchmark track closely $bias^{(50)}_{j,t}$, reinforcing the conclusion that noisy, dispersed information is unlikely to be an important driver of our measured belief distortions.

Figure 5 contrasts the common and heterogeneous components of these belief distortions over time, breaking them out by survey. The common component is measured as the first principle component (PC) of $bias^{(i)}_{j,t}$ across all percentiles i, with heterogeneity exhibited by the distribution of $bias^{(i)}_{j,t}$ across i.[14] For all surveys, we observe substantial variation in belief distortions over time that is common across SPF respondents. For SPF and BC, the optimism about economic growth in the immediate aftermath of the Great Recession is present in the common component, as is a downward bias to inflation expectations for this same time period. At the same time, there is substantial heterogeneity across responses that varies over time, with greater dispersion observed in recessions. For the SPF, the most optimistic

[14] Since the PCs and their factor loadings Λ are not separately identifiable, the loadings are normalized by $(\Lambda'\Lambda)/N = \mathbf{I}_q$ where N is the number of $bias^{(i)}$ series over which common factors are formed and q is the number of common factors. This implies that the units for these series have no straight-forward interpretation in terms of the raw data.

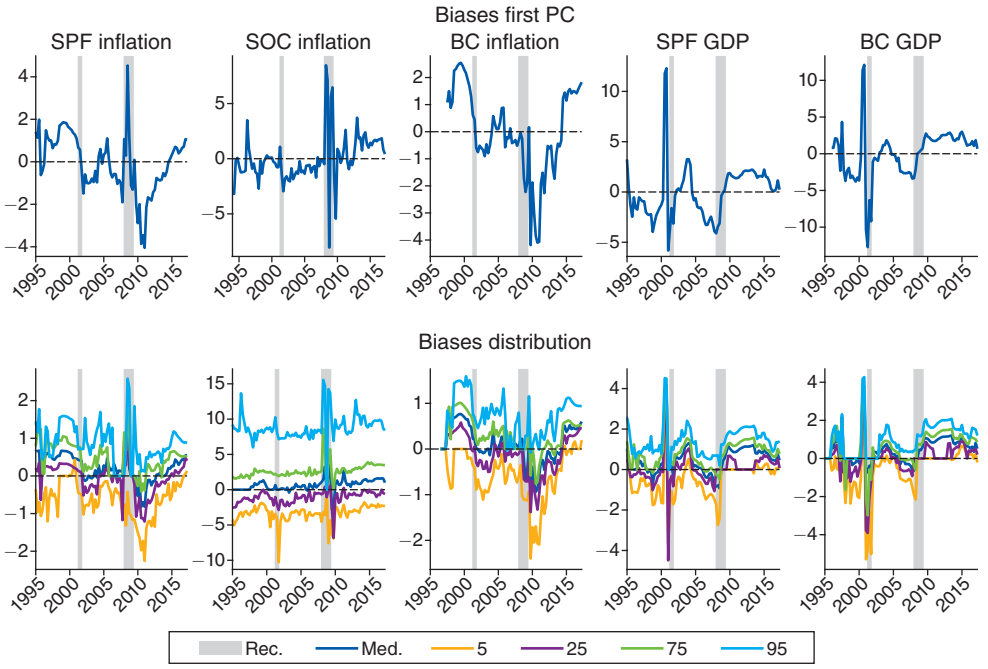

FIGURE 5. COMMON AND HETEROGENEOUS DISTORTIONS

Notes: The first row reports the first principal component of the biases across different surveys. For each respondent type, the second row reports the time series $bias^{(i)}_{j,t+h} = \mathbb{F}^{(i)}_t[y_{j,t+h}] - \mathbb{E}^{(i)}_t[y_{j,t+h}]$. The figure does not report the SOC GDP bias because only one series is available in that case. NBER recessions are shown with gray shaded bars. The sample is 1995:I–2018:II.

and pessimistic responses differ in some recession periods by more than 4 percent for GDP growth and by more than 2 percent for inflation, similarly for the BC survey. The high degree of disagreement among professional forecasters resulting in substantial heterogeneity in biases is an example of what Kahneman, Sibony, and Sunstein (2021) call "noise." For households in the SOC, the heterogeneity in measured belief distortions about inflation is enormous, especially immediately after the Great Recession, where the forecast of annual inflation from the respondent-type at the ninetieth percentile is almost 15 percent, while that for the respondent-type at the fifth percentile is less than -5 percent.

Figure 6 compares forecasted and actual values over time. The figure displays the median forecast of four-quarter-ahead inflation or GDP growth over our evaluation sample along with the actual inflation or GDP growth rate during the corresponding four quarter period being forecast. For all surveys, the machine has been more accurate not just on average over the long evaluation samples, but also consistently over the last five years of these samples, from 2013:II to 2018:II, and by even larger magnitudes. For GDP growth, the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ for the median SPF forecast is 0.82 over this subperiod, while it is 0.67 for median BC forecast. For inflation, the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ over this same subperiod is 0.63 for the median

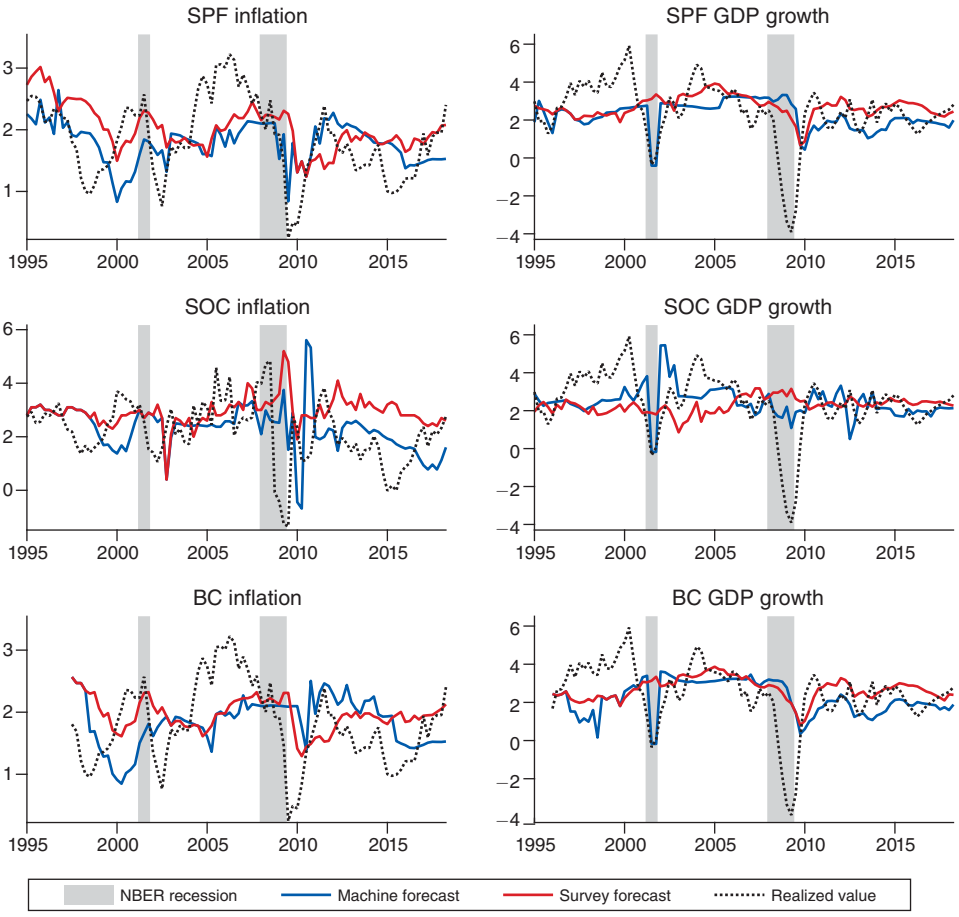

FIGURE 6. FORECASTED VERSUS ACTUAL INFLATION, GDP GROWTH

Notes: For each variable and survey, the figure reports the median survey forecast of inflation or GDP growth over the next four quarters, the corresponding machine forecast, and the realized inflation or GDP growth values during this period. Realized values are measured in real-time data as the vintage available four quarters after the period being forecast. NBER recessions are shown with gray shaded bars. The sample is 1995:I–2018:II.

SPF forecast, 0.67 for the median BC forecast, and 0.37 for the median SOC forecast. In both cases, professional forecasters and households alike underperformed in this subperiod because they systematically overpredicted both economic growth and inflation. That the machine does better at the end rather than the beginning of the evaluation sample is noteworthy, since it suggests that bounded rationality in the form of limitations on the human capacity for collecting and processing large amounts of information are unlikely to fully explain these findings. At least professional forecasters would have been capable by 2013 of taking advantage of advances in information-processing technology.

Unsurprisingly, the machine does not perform well in every time period of our sample. Figure 6 shows that professional forecasters made large forecast errors that were overly optimistic about GDP growth at the onset of the Great Recession, as

noted in Gennaioli and Shleifer (2018). This pattern is likewise evident in Figure 6 for all surveys studied here. The figure shows that large forecast errors were made during this episode by the machine as well, with the machine algorithm doing somewhat better than the SOC forecast, only slightly better than the BC forecast, and about the same but if anything slightly worse than the SPF forecast. This occurs despite the fact that the machine algorithm takes into account hundreds of pieces of real-time information including that encoded in numerous financial series and dozens of credit spreads, recorded at daily, monthly, and quarterly sampling intervals. Such large ex post forecast errors during the Great Recession are arguably understandable when placed in the broader context of the time. As noted above, in contrast to many past recessions, yield spreads generally failed to signal the Great Recession to come. Moreover, the period leading up to the recession was characterized by unusually elevated objective uncertainty about the macroeconomy (Jurado, Ludvigson, and Ng 2015 and Ludvigson, Ma, and Ng 2019). We argue that this episode underscores the role of largely unforeseen events in generating occasionally large prediction errors, not all of which can be attributed to a systematic bias in expectations.

Of course, with hindsight we now know that the Great Recession was preceded by a global financial crisis, itself triggered by a collapse in the value of residential real estate. It is thus tempting to consider feeding the machine a different switching indicator just prior to the Great Recession, for example, one based on credit spreads or indicators of balance sheet health for firms and households. Our view is that this approach would be hard to defend, however. Unlike the case for yield spreads, where by the mid-1990s there existed a large body of public evidence showing their unique predictive power for recessions, there is no analogous body of empirical evidence for credit spreads and/or balance sheet indicators *before* 2007. Indeed, several of the empirical studies cited above and published in the early to mid 1990s explicitly compared credit spreads to yield spreads for forecasting output growth. Such studies universally found that credit spreads were comparatively weak predictors of recessions. Moreover, many of the balance sheet indicators now understood to be predictive for the global financial crisis (e.g., Greenwood et al. 2020) would not have been available in real time prior to the crisis, given the substantial data collection and data processing lags for such indicators. In short, the focus today on credit spread and balance sheet indicators as key predictors of recessions appears largely motivated by our ex post understanding of the 2007–2008 global financial crisis, rather than by a large body of prior knowledge that these indicators were the most robust predictors of economic contractions before the crisis. Since our approach is explicitly designed to exclude from the measure of belief distortions ex post mistakes that could only be understood with hindsight, we take the conservative approach of restricting our recession indicators to those that can be clearly defended on the basis of a prior body of publicly available empirical evidence.

C. *Bias Decomposition*

If the machine algorithm generates better forecasts, the survey respondents must be misweighting pertinent economic information. This raises the question: What kind of errors in judgment are the respondent-types making? To address this question, recall

that the time t bias is defined as the difference between the survey respondent-type and machine forecasts:

$$(9)\quad bias^{(i)}_{j,t+h} \equiv \mathbb{F}^{(i)}_{j,t+h|t} - \mathbb{E}^{(i)}_{j,t+h|t} = \mathbb{F}^{(i)}_{t}\left[y_{j,t+h}\right] - \hat{\alpha}_{jh} - \hat{\beta}^{(i)}_{jh\mathbb{F}}\mathbb{F}^{(i)}_{t}\left[y_{j,t+h}\right] - \hat{\mathbf{B}}^{(i)'}_{jh\mathcal{Z}}\mathcal{Z}_{jt}$$

$$= \underbrace{\left[-\hat{\alpha}^{(i)}_{jh}\right]}_{\text{Intercept}} + \underbrace{\left[\left(1 - \hat{\beta}^{(i)}_{j\mathbb{F}}\right)\mathbb{F}^{(i)}_{t}\left[y_{j,t+h}\right]\right]}_{\text{Survey}} + \underbrace{\left[-\hat{\mathbf{B}}^{(i)'}_{j\mathcal{Z}}\mathcal{Z}_{jt}\right]}_{\text{Info variables}}.$$

We are interested in the contribution of the three terms on the right-hand side of (9), shown in large square brackets, the sum of which equals 100 percent of $bias^{(i)}_{j,t+h}$. This decomposition gives an indication of which information is most misweighted by the survey respondent-type, and by how much. The intercept term $\hat{\alpha}^{(i)}_{jh}$ changes over the evaluation sample through the dynamic estimation algorithm and is akin to a time-varying latent conditional mean applied to the most recent rolling subsample window. We refer to this parameter as a "rolling mean" and denote it with a t subscript, i.e., $\hat{\alpha}^{(i)}_{jh,t}$. The estimates $\hat{\beta}^{(i)}_{j\mathbb{F}}$ and $\hat{\mathbf{B}}^{(i)'}_{j\mathcal{Z}}$ also vary over the evaluation sample and are likewise denoted with a t subscript.

It is useful to consider the magnitude and signs of the coefficients in the components above. First consider the coefficient on the survey forecast. If $\hat{\beta}^{(i)}_{j\mathbb{F},t} < 1$, this implies that the machine improves forecasts by downweighting the survey response in favor of giving greater absolute weight to publicly available information. Thus an estimate of $\hat{\beta}^{(i)}_{j\mathbb{F},t} < 1$ implies that the respondent-type *over*weighted the marginal information in her own forecast relative to an efficient weighting of publicly available information. Conversely, if $\hat{\beta}^{(i)}_{j\mathbb{F},t} > 1$, the machine improved forecasts by giving greater weight to the survey forecast than the implicit weight given by the respondent-type to her own forecast. For the information variables and the rolling mean, any estimate of $\hat{\mathbf{B}}^{(i)'}_{j\mathcal{Z},t} \neq 0$ or $\hat{\alpha}^{(i)}_{jh,t} \neq 0$ indicates that the machine improved forecasts by giving greater absolute weight to $\mathcal{Z}_{j,k,t}$ or $\hat{\alpha}^{(i)}_{jh,t}$ compared to the respondent-type's implicit weight of zero conditional on her own forecast. Thus we refer to any estimate with $\hat{\mathbf{B}}^{(i)'}_{j\mathcal{Z},t} \neq 0$ or $\hat{\alpha}^{(i)}_{jh,t} \neq 0$ as *under*weighting of these sources of information.

Figure 7 reports, for each survey and each variable, the contribution to the bias in the median forecast of the three terms in square brackets in (9) at each point in time over our forecast evaluation samples. The solid lines in each subfigure of Figure 7 report the total median bias, $bias^{(50)}_{j,t+h}$, while the contributions of the three terms in square brackets in (9) are reported as bar charts, with the height of the bar showing the absolute magnitude by which that component contributed to the bias. Any above (below) zero bar indicates that the term contributed positively (negatively) to the overall bias. Since there are many terms in the information variable term, the figure reports contributions only for the most quantitatively important information variable contributors to the bias at each time t. In the case of the survey contribution, we further indicate with color coded bars whether a contribution to the bias was created by the respondent-type having over- or underweighted her own forecast. A

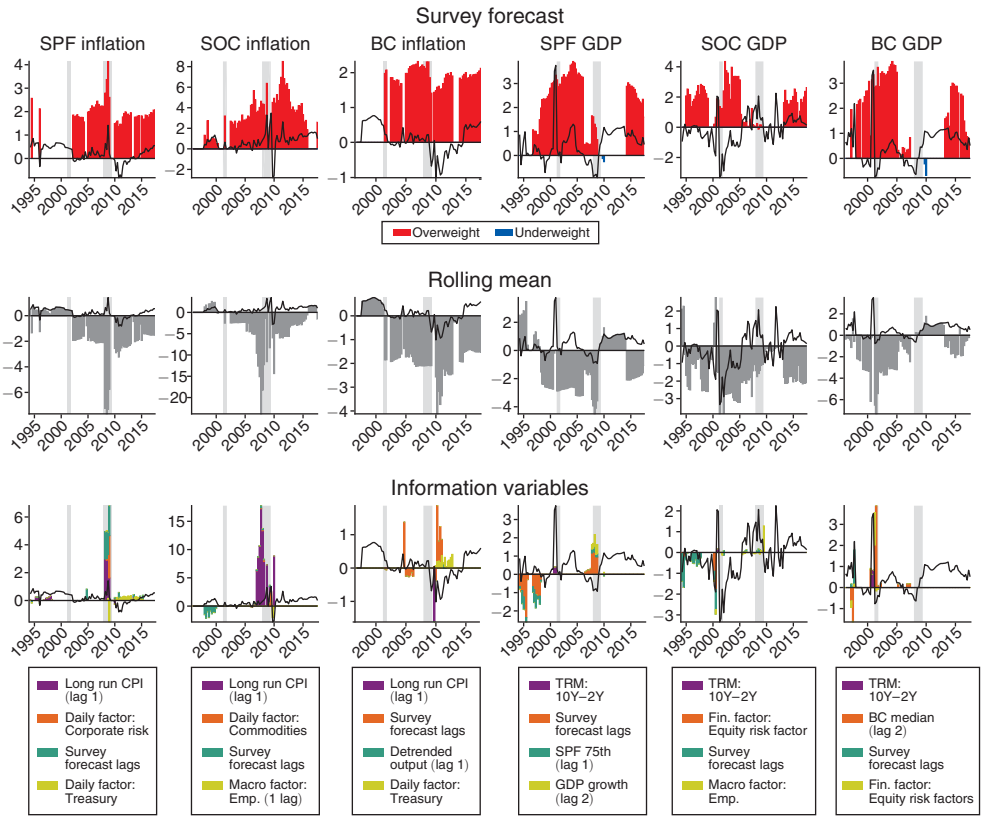

FIGURE 7. BIAS DECOMPOSITION: MEDIAN FORECAST

Notes: The figure plots contributors to the median bias $\mathbb{F}_t^{(50)}[y_{j,t+h}] - \mathbb{E}_t^{(50)}[y_{j,t+h}] = -\hat{\alpha}_{jh}^{(50)} + \left(1 - \hat{\beta}_{j\mathbb{F}}^{(50)}\right)\mathbb{F}_t^{(50)}[y_{j,t+h}] - \beta_{j\mathcal{Z}}^{(50)'}\mathcal{Z}_{jt}$ at each time t. The solid black lines in each subpanel plot the median bias, $F_t^{(50)}[y_{j,t+h}] - E_t^{(50)}[y_{j,t+h}]$. The bar charts in the first row panel report $\left(1 - \hat{\beta}_{j\mathbb{F}}^{(50)}\right)\mathbb{F}_t^{(50)}[y_{j,t+h}]$; those in the second row report $-\hat{\alpha}_{jh}^{(50)}$; those in the third row report $-\beta_{j\mathcal{Z}}^{(50)'}\mathcal{Z}_{jt}$ for the most important predictor contributors to the time t bias. Red bars indicate that the survey forecast was given too much weight relative to the machine efficient forecast, corresponding to $\left(1 - \hat{\beta}_{j\mathbb{F}}^{(50)}\right) > 0$. Blue bars indicate that the survey forecast was given too little weight relative to the machine efficient forecast, corresponding to $\left(1 - \hat{\beta}_{j\mathbb{F}}^{(50)}\right) < 0$. NBER recessions are shown with gray shaded bars.

red bar indicates that the median respondent-type *over*weighted her own forecast (i.e., $\hat{\beta}_{j\mathbb{F},t}^{(i)} < 1$), while a blue bar indicates that she *under*weighted. For the intercept and information variables terms, any bar with a nonzero height indicates that the respondent-type gave too little absolute weight to that information. Recessions are shown in the figure by light gray shaded areas.

A key finding exhibited in Figure 7 that is robust across all surveys and all variables is that $\hat{\beta}_{j\mathbb{F},t}^{(50)}$ is very often substantially less than one. This happens not only for all surveys and for both inflation and GDP growth expectations, but also for most time periods in the evaluation sample. The mean (across time) values of $\hat{\beta}_{j\mathbb{F},t}^{(50)}$ for inflation and GDP growth are 0.40 and 0.45, respectively, for SPF, 0.33

and 0.54 for BC, and 0.14 and 0.41 for SOC. Thinking back to the model of public and private signals, we can interpret a finding of $\hat{\beta}^{(i)}_{j\mathbb{F},t} > 0$ as indicating that the marginal information contained in the survey response, capturing information intangible to the machine, such as that provided by a private signal or judgment, is in fact often valuable. But the finding that $\hat{\beta}^{(i)}_{j\mathbb{F},t} < 1$ indicates that such information is less valuable than the implicit weight placed on it by the survey respondent. The model of private and public signals implies that estimates $\hat{\beta}^{(i)}_{j\mathbb{F},t} < 1$ occur when the respondent-type overrelies on her private information or judgment, either because she overestimates the precision of her private signal or because she inefficiently combines the public information, thereby underestimating the precision of the public signal. Many instances of such an apparent overreliance are exhibited in Figure 7 by the frequent, tall red bars in the survey forecast panels of the first row. The length of the bars indicates that this factor contributes in most cases to quantitatively large distortions in macro expectations. For example, the first panel in Figure 7 indicates that this factor contributed 4 percent to the upward bias in the median SPF forecast of inflation—accounting for more than 100 percent of the bias—during several periods at the end of the Great Recession.

If the median forecaster typically placed too much weight on her own forecast, then by definition she placed too little absolute weight on other information. The bottom two rows of Figure 7 gives an indication of the type of other objective economic information that was misweighted by the median forecaster over time. A key finding here is that the type of information is not static but instead changes over time. For example, in forming inflation expectations, the third rows shows that, during the Great Recession, too little attention was paid by the median SPF respondent to daily data on corporate credit spreads and to monthly data on long-run survey inflation forecasts, while for the years immediately after the Great Recession, between 2010 and 2015, the median respondent paid too little attention to daily information on Treasury yields and lagged values of the SPF forecasts. The type of information that was underweighted varies also across surveys. For the SOC, underweighting of long-run CPI survey forecasts shows up right before the Great Recession, but not elsewhere in the sample, while we find that the BC median forecast underweighted this information *after* the Great Recession while subsequently giving too little weight to lagged survey forecasts.

Turning to expectations of economic growth, Figure 7 shows that the overoptimism displayed by professional forecasters (both SPF and BC) in the post-Great Recession period was largely driven, at first, by paying too little attention to the predictable slowing of average economic growth captured by the rolling mean, and then subsequently by repeated instances of overweighting the marginal information in the survey response relative to what would be optimal under an efficient weighting of public information. Evidently, the median professional forecaster placed too much weight on a mistaken belief that economic growth would accelerate more than it did, a factor that accounts for more than 100 percent of the bias in the last five years of the sample. Overall the results are suggestive of a substantial overreliance by professional forecasters on the private or judgmental component of their predictions.

Taken together, the findings in Figure 7 underscore the crucial role of considering extensive and varied information in reducing forecaster bias. Although our

machine learning algorithm often chooses sparse specifications, the findings in this figure show that *different* sparse information sets are relevant at different points in time.[15] Since it is virtually impossible for a human to know with certainty which information may be relevant ex ante, algorithmic "openness" to wide-ranging and rich sources of information are vital for improving forecast accuracy over extended periods of time.

D. *Some Comparisons with the Literature*

With these results in hand, we now revisit some results in the prior literature that help illuminate the role played by key elements of our machine learning approach for establishing whether and by how much beliefs embedded in human judgments are distorted.

One key element pertains to the basic principle of out-of-sample versus in-sample forecasting, a principle illustrated by contrasting results from ex ante and ex post econometric analyses, bearing in mind that survey respondents are asked to make genuine out-of-sample forecasts based on information known in real time. To illustrate the potential importance of this for the measurement of belief distortions, we revisit the *in-sample* regressions run in Coibion and Gorodnichenko (2015)—henceforth, CG. CG found that mean survey forecast errors are positively predicted by ex ante mean forecast revisions in in-sample regressions. We reproduce their findings for the SPF on updated data and report the results in panel A of Table 3. Consistent with CG, we find strong evidence that lagged forecast revisions predict next period's forecast error in these regressions. Moreover, other information, e.g., lagged inflation, is found to be unimportant in predicting forecast errors once the information in forecast revisions is taken into account.[16] CG observe that these findings are consistent with the implications of theories that feature information frictions and underreaction to aggregate news.[17]

The bottom panel of Table 3 reports results from the same regression forecasts, but this time run out of sample rather than in sample. (Details on the standard out-of-sample estimation procedure can be found in the online Appendix.) Table 3 shows that over a range of forecast evaluation subsamples and using either rolling or recursive regressions, the mean SPF survey forecast generates much lower prediction error than a specification that attempts to exploit information in the lagged revision of the mean forecast. In contrast to the in-sample findings, the inclusion of information on lagged forecast revisions substantially *worsens* predictions of mean survey forecast errors when these predictions are made on an out-of-sample basis. This result recalls a body of prior econometric evidence finding that survey forecast

[15] One possible sparse specification is the random walk model for inflation found previously in d'Agostino, Giannone, and Surico (2006) to perform better than professional survey forecasts over the 1985:I–1994:IV subperiod. The specifications entertained by the machine nest the random walk model, but the dynamic algorithm chooses that specification infrequently over our forecast sample.

[16] We include one lag of the *quarterly* inflation rate as an additional control variable, consistent with the procedure implemented in CG. There is a typo in the published version of CG that erroneously indicates their procedure controlled for one lag of annual rather than quarterly inflation.

[17] As an aside, we note that the machine forecast errors do not exhibit a correlation with lagged machine forecast revisions, even in in-sample regressions. These results are reported in the online Appendix.

TABLE 3—CG REGRESSIONS OF FORECAST ERRORS ON FORECAST REVISIONS

Panel A. In-sample regressions (CG sample)

Regression: $\pi_{t+3} - \mathbb{F}_t^{(\mu)}[\pi_{t+3}] = \alpha^{(\mu)} + \beta^{(\mu)}\left(\mathbb{F}_t^{(\mu)}[\pi_{t+3}] - \mathbb{F}_{t-1}^{(\mu)}[\pi_{t+3}]\right) + \delta\pi_{t-1,t-2} + \epsilon_t$

Constant	0.001	−0.077
t-stat	(0.005)	(−0.442)
$\mathbb{F}_t[\pi_{t+3,t}] - \mathbb{F}_{t-1}[\pi_{t+3,t}]$	1.194	1.141
t-stat	(2.496)	(2.560)
$\pi_{t-1,t-2}$		0.021
t-stat		(0.435)
$\bar{R}^2$	0.195	0.197

Panel B. Out-of-sample regressions

Regression: $\pi_{t+3} - \mathbb{F}_t^{(\mu)}[\pi_{t+3}] = \alpha^{(\mu)} + \beta^{(\mu)}\left(\mathbb{F}_t^{(\mu)}[\pi_{t+3}] - \mathbb{F}_{t-1}^{(\mu)}[\pi_{t+3}]\right) + \epsilon_{t+3}$

Method	Forecast sample	$\text{MSE}_{CG}/\text{MSE}_{\mathbb{F}}$
Rolling 5 years	1975:IV–2018:II	1.38
Rolling 10 years	1980:IV–2018:II	1.29
Rolling 20 years	1990:IV–2018:II	1.31
Recursive 5 years	1975:IV–2018:II	1.69
Recursive 10 years	1980:IV–2018:II	1.60
Recursive 20 years	1990:IV–2018:II	1.33

Notes: Panel A reports the in-sample results over the sample used in Coibion and Gorodnichenko (2015) (CG), 1969:I to 2014:IV. Newey-West corrected *t*-statistics with lags = 4 are reported in parentheses. Panel B reports the ratio of out-of sample MSE of the CG model forecast to that for the survey forecast computed using different rolling or recursive estimation windows. The MSE for the CG model averages the (square of the) forecast errors $\pi_{t+3} - \widehat{\pi}_{t+3}^{(\mu)}$, where $\widehat{\pi}_{t+3}^{(\mu)} = \widehat{\alpha}_t^{(\mu)} + \left(1 + \widehat{\beta}_t^{(\mu)}\right)\mathbb{F}_t^{(\mu)}[\pi_{t+3}] - \widehat{\beta}_t^{(\mu)}\,\mathbb{F}_{t-1}^{(\mu)}[\pi_{t+3}]$. In both panels, the regression estimation uses the latest vintage of inflation in real time and, following CG, computes forecast errors with real-time data available four quarters after the period being forecast. Annual inflation is defined as $\pi_{t+3,t} = \frac{P_t}{P_{t-1}} \times \frac{P_{t+1}}{P_t} \times \frac{P_{t+2}}{P_{t+1}} \times \frac{P_{t+3}}{P_{t+2}}$, and $\mathbb{F}_t[\pi_{t+3,t}]$ is the mean forecast of annual inflation as of time *t* from the SPF. The sample of panel B spans the period 1969:I–2018:II.

of inflation are hard to beat or even match with statistical models when forecasts are conducted out-of-sample.[18]

How can we reconcile the contradictory in-sample and out-of-sample evidence? One possibility is that the empirical relationship between forecast errors and lagged forecast revisions is unstable, as suggested by results below. Such an instability can create a high degree of sampling error so that what is revealed to be important ex post is simply not apparent ex ante. Whatever the reason for the poor performance of the specification out-of-sample, we've argued here that it is impossible to establish the extent to which beliefs are distorted due to information frictions or any other cause, unless the statistical model used to measure distortions adheres to the same forecasting context that survey respondents faced at the time they made their predictions. After all, even agents such as our machine who possess vast information processing capacity will optimally downweight information that might appear relevant ex post if it systematically failed to improve forecasts ex ante. It would not be correct to interpret this type of downweighting as underreaction to economic news or as evidence of a systematic bias in expectations.

[18] For example, Ang, Bekaert, and Wei (2007); Del Negro and Eusepi (2011); Aiolfi, Capistrán, and Timmermann (2011); Genre et al. (2013); and Faust and Wright (2013).

While this shows that lagged forecast revisions are not reliable out-of-sample predictors of mean forecast errors in the simple regression specification CG considered, it is reasonable to ask whether they contain any valuable predictive information in our machine specifications. We run the following machine version of the CG regressions, which use the mean SPF forecast $\mathbb{F}_t^{(\mu)}$ and again places observations on forecast errors on the left-hand side:

$$\pi_{j,t+3} - \mathbb{F}_t^{(\mu)}\left[\pi_{j,t+3}\right] = \alpha_{\pi}^{(\mu)} + \beta_{\pi\text{FR}}^{(\mu)}\left(\mathbb{F}_t^{(\mu)}\left[\pi_{t+3}\right] - \mathbb{F}_{t-1}^{(\mu)}\left[\pi_{t+3}\right]\right) + \mathbf{B}_{\pi\mathcal{Z}}^{(\mu)'}\mathcal{Z}_{\pi t} + \epsilon_{\pi t+h}. \tag{10}$$

This machine estimation differs from the CG estimation in three ways. First, the machine forecasts are made out of sample rather than in sample. Second, the machine entertains the large-scale information set $\mathcal{Z}_{\pi t}$ as additional predictor variables. Third, the machine uses the EN estimator and dynamic cross-validation algorithm described above, while CG use least squares. We denote the estimate of the coefficient on forecast revisions from this machine specification with $\beta_{\pi\text{FR}}^{(\mu)}$ and that from the univariate, in-sample least squares regression of CG as $\beta_{\pi\text{CG}}^{(\mu)}$.

Figure 8 reports the coefficients $\beta_{j\text{FR}}^{(\mu)}$ obtained from estimating (10) using the machine algorithm. Since the machine estimation is repeated on rolling samples using real-time information up to time t, the figure reports the entire time series of estimates $\widehat{\beta}_{j\text{FR},t}^{(\mu)}$ using a bar chart, where the height of the bar indicates the magnitude of $\widehat{\beta}_{j\text{FR},t}^{(\mu)}$ and the time period t refers to the period of the external evaluation sample 1995:I–2018:II, which is given on the x-axis. Time periods t for which there is no bar displayed indicate $\widehat{\beta}_{j\text{FR},t}^{(\mu)} = 0$. For comparison, the fixed in-sample estimates $\widehat{\beta}_{\pi\text{CG}}^{(\mu)}$ from the CG least squares regressions are shown as separate horizontal lines, one for each of three estimation samples: 1969:I–2014:IV (CG sample), 1969:I–2018:II (our full sample) and 1995:I–2018:II (our machine external evaluation sample).

Figure 8 shows that the horizontal lines indicating $\widehat{\beta}_{\pi\text{CG}}^{(\mu)}$ over the first two samples are both close to 1.2, while that for the shorter recent sample are smaller by half. By contrast, the machine estimates $\widehat{\beta}_{j\text{FR},t}^{(\mu)}$ are always much smaller than the in-sample least squares estimates $\widehat{\beta}_{\pi\text{CG}}^{(\mu)}$ when those are obtained using the two longer subsamples, and they only match or exceed the half-as-large value in the shorter recent sample in one time period. Instead, the machine estimates of $\widehat{\beta}_{j\text{FR},t}^{(\mu)}$ are shrunk all the way to zero in 88 out of 94 quarters in favor of placing greater absolute weight on other pieces of information contained in $\mathcal{Z}_{\pi t}$ or $\widehat{\alpha}_{\pi,t}^{(\mu)}$. These findings do not point to an important role for ex ante revisions in predicting average ex post forecast errors.

A second key element of our machine learning problem pertains to the data-rich environment that survey respondents operate in. To illustrate the importance of this, we revisit an exercise in the spirit of Chauvet and Potter (2013), who considered a wide range of low dimensional statistical models for predicting GDP growth, finding that a second-order autoregression performed best for one-quarter ahead predictions when evaluated in a hold-out sample. Table 4 shows the estimated autoregressive coefficients from rolling, one-quarter-ahead, out-of-sample forecasting regressions

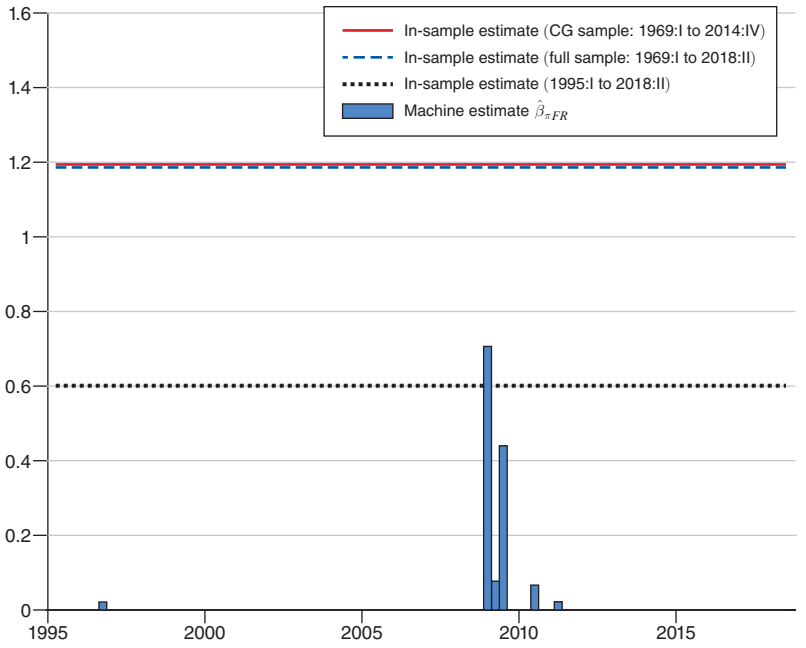

FIGURE 8. COEFFICIENT ON FORECAST REVISIONS

Notes: The blue bar plots the estimated coefficient on the forecast revision from regressions of forecast errors on forecast revisions and additional regressors for the mean of the SPF inflation forecast: $\underbrace{\pi_{t+3} - F_t^{(\mu)}[\pi_{t+3}]}_{\text{Forecast Error}} = \alpha_{jh}^{(\mu)} + \beta_{j\text{FR}}^{(\mu)}\left(\underbrace{F_t^{(\mu)}[\pi_{t+3}] - F_{t-1}^{(\mu)}[\pi_{t+3}]}_{\text{Forecast Revisions}}\right) + B_{jZ}^{(\mu)\prime} Z_{jt} + \epsilon_{jt+h}$. The sample is 1995:I–2018:II. The red solid line shows the estimated in-sample coefficient over the CG sample 1969:I–2014:IV. The blue dashed line shows the estimated in-sample coefficient over the full sample 1969:I–2018:II. The black dotted line shows the estimated in-sample coefficient over the evaluation sample 1995:I–2018:II.

of GDP growth on predictors, in two specifications. A high-dimensional specification entertains very large numbers of potential predictor variables, as in baseline machine specification. The two autoregressive lags are always among these predictors. A low dimensional specification uses the two autoregressive lags and only two additional predictors: the SPF median forecast of GDP growth and its current nowcast, both of which are also included in the high dimensional model. We find that the coefficient on the first autoregressive lag, large and positive in the low-dimensional setting, is zero in the high dimensional setting. As we have seen, this result does not imply that sparse specifications are rarely optimal. What it points to is the difficulty with knowing *which* small number of predictor variables are likely to be informative over time, when one does not have the benefit of hindsight afforded by an academic study of a single hold-out sample. The challenge for real-time decision-making is that different pieces of information become relevant at different points in time.

E. *Belief Distortions over the Business Cycle*

For our last set of results, we investigate the implications of our estimates for over- and underreaction by survey respondents, a subject intense interest in the behavioral

Table 4—Average Coefficients on the First Two AR lags

	High dimensional	Low dimensional
β_1	0.000	0.022
β_2	−0.002	−0.013

Notes: This table reports average autoregressive coefficients from one-year-ahead rolling regressions of real GDP growth on predictors. The average coefficient on the first AR lag is β_1; the average coefficient on the second is β_2. The high dimension estimation entertains very large numbers of potential predictors, in addition to the autoregressive lags, while the low dimension setting uses only two additional predictors. The sample spans 1995:I–2018:II.

economics literature. We do so in a dynamic context in the wake of cyclical shocks, using the approach of Angeletos, Huo, and Sastry (2020)—henceforth, AHS. Specifically, AHS estimate the dynamic responses of inflation or real GDP growth, as well as survey forecasts of those variables, to two cyclical shocks identified in Angeletos, Collard, and Dellas (2018a).[19] The cyclical shocks are the "inflation-targeted shock," ϵ_t^{π}, and the "GDP-targeted shock," ϵ_t^{GDP}. By construction, these shocks account for most of the business cycle variation in inflation and GDP growth, respectively.[20] Due to limitations of space, we restrict our analysis to the SPF median forecasts of four-quarter-ahead inflation or GDP growth.

Figure 9 reports dynamic responses of the machine forecast, the median SPF survey forecast, and the relevant outcome variable, to innovations in ϵ_t^{π} and ϵ_t^{GDP}, estimated using local projections (Jorda 2005).[21] The plots "align" the forecast responses so that, at a given vertical slice of the plot, the outcome and forecast responses are measured over the same time horizon and the difference between the two is the forecast error. For example, given a shock at time t, the first response plotted for the survey forecast is $\mathbb{F}_t^{(50)}[y_{t+4}]$, which is aligned vertically with the response of y at time $t + 4$. Following AHS, we set $H = 20$ quarters as the maximum period for tracing out impulse responses.

Although the outcome variable is shown in Figure 9 for context, our measure of dynamic under- and overreaction of the survey respondent's belief is taken vis-à-vis the *machine* forecast, not the ex post outcome. The figure shows that, in general, survey respondents initially underreact to a shock (but more so in response to the output shock) but later overreact (especially to the inflation shock). However, comparing the survey forecast to the realized value of the outcome variable greatly overstates the degree of over- or underreaction that can be attributed to belief distortions. This can be observed by noting that the survey forecasts recorded after the shock undershoot the realized outcome by much more than they undershoot the machine forecasts and they subsequently overshoot the realized outcome by more than they

[19] We are grateful to the authors for providing us their data on these shocks.

[20] These shocks are identified using a ten-variable macro vector autoregression (VAR) as the structural shock that maximizes the volatility of the outcome variable (i.e., inflation, GDP growth) at frequencies corresponding to cycles between 6 and 32 quarters.

[21] The online Appendix gives the details of this estimation. We use a four-quarter forecast horizon, in contrast to AHS who use a three-quarter horizon. Our sample is also shorter than that used in AHS. The online Appendix shows that we reproduce the results in AHS for the same forecast horizon and sample size that they use, and that the results are similar using the shorter sample of this paper.

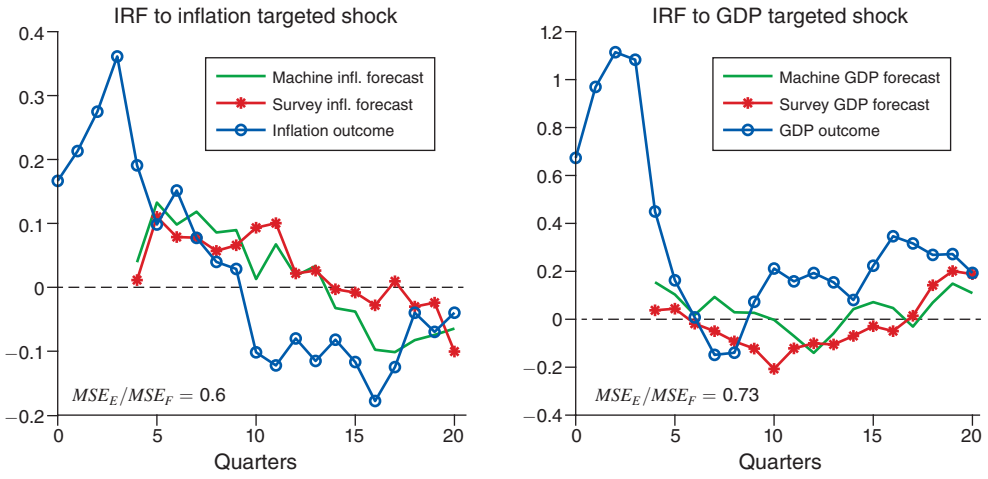

FIGURE 9. DYNAMIC RESPONSES TO CYCLICAL SHOCKS

Notes: The figure plots dynamic responses of the machine and survey beliefs $\mathbb{F}_t^{(50)}[\cdot]$ and $\mathbb{E}_t^{(50)}[\cdot]$ for the median respondent of the SPF to cyclical shocks from Angeletos et al. (2018a) (AHS). The AHS inflation and GDP growth "targeted" cyclical shocks are those from a ten-variable VAR that maximize the volatility in inflation and GDP growth at business cycle frequencies, respectively. The figure aligns the forecast responses such that, at a given vertical slice, the outcome and forecast responses are measured over the same horizon, and the difference between the two is the forecast error. "$\text{MSE}_E/\text{MSE}_F$" is the ratio of the machine to survey mean squared forecast error averaged over the response time periods in the plot. The vintage of observations on the outcome variable is the one available four quarters after the period being forecast. The sample is 1995:I–2018:II.

overshoot the machine forecast. AHS have interpreted the difference between the survey forecast and the realized value of the outcome variable as a measure of nonrational expectations. By contrast, we interpret the difference between the survey and *machine* forecasts as a measure of systematic expectational error, and the difference between the machine forecast and the outcome variable as pure random forecast error, rather than bias. The discrepancy between the two suggests that the cyclical shocks ϵ_t^{π} and ϵ_t^{GDP} are not well observed in real time, even by a machine with a high degree of information processing capacity. This may be because ϵ_t^{π} and ϵ_t^{GDP} are constructed from an in-sample estimation using fully revised, final-release historical data, while both the survey and machine forecasts are by contrast forced to rely entirely on real-time information, including that about the outcome variables being forecast.[22]

In the wake of both cyclical shocks, the machine produces more accurate forecasts than the median SPF survey respondent. The gains in forecast accuracy are larger for inflation where the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ is 0.60, but even for GDP growth the ratio $MSE_{\mathbb{E}}/MSE_{\mathbb{F}}$ is 0.73. That the machine improves forecasts in this context is noteworthy because it was not trained to optimize out-of-sample prediction at the

[22] It is not obvious that these estimated cyclical shocks can be observed in real time. The SPF collects survey responses in February on the outlook for GDP in the second quarter of the year, but the *advance* estimate of Q2 GDP is not released until the end of July. The *final-release* data used to construct the shocks are subject to further revision subsequently over the course of two months. And while some information pointing toward a large business cycle shock may be available in real time, such as that contained in financial markets, that is already accounted for by the machine.

specific business cycle frequencies that, by construction, dominate variation in the outcome variables in Figure 9.

V. Conclusion

This paper provides new measures of belief distortions in survey responses and relates them to macroeconomic activity. Our measures are based on a novel dynamic machine learning algorithm explicitly designed to combat overfitting and detect demonstrable, ex ante errors in macroeconomic expectations. For the median respondent from all surveys, expectations about both inflation and GDP growth are biased upward on average, with overoptimism about GDP growth especially prevalent among professional forecasters in the period after the Great Recession up to the end of our sample in 2018:II. These averages mask large variation over time in the median respondent's bias, as well across respondents at any given point in time. A pervasive finding across all surveys is that respondents place too much weight on the marginal information embedded in their own belief and too little weight on other publicly available information. In response to cyclical shocks, we find that *under*reaction preponderates in survey expectations of economic growth, while inflation expectations show greater delayed *over*reaction. The results suggest that artificial intelligence algorithms can be productively deployed to correct errors in human judgment and improve predictive accuracy.

REFERENCES

Adam, Klaus, Albert Marcet, and Johannes Beutel. 2017. "Stock Price Booms and Expected Capital Gains." *American Economic Review* 107 (8): 2352–2408.

Afrouzi, Hassan, and Laura Veldkamp. 2019. "Biased Inflation Forecasts." Paper presented at the Annual Meeting of the Society for Economic Dynamics, St. Louis, MO, June 29.

Aiolfi, Marco, Carlos Capistrán, and Allan Timmermann. 2011. "Forecast Combinations." In *Oxford Handbook of Economic Forecasting*, edited by Michael P. Clements and David F. Hendry, 355–89. New York: Oxford University Press.

Amisano, Gianni, and John Geweke. 2017. "Prediction Using Several Macroeconomic Models." *Review of Economics and Statistics* 99 (5): 912–25.

Amromin, Gene, and Steven A. Sharpe. 2014. "From the Horse's Mouth: Economic Conditions and Investor Expectations of Risk and Return." *Management Science* 60 (4): 845–66.

Ang, Andrew, Geert Bekaert, and Min Wei. 2007. "Do Macro Variables, Asset Markets, or Surveys Forecast Inflation Better?" *Journal of Monetary Economics* 54 (4): 1163–1212.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas. 2018a. "Business Cycle Anatomy." NBER Working Paper 24875.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas. 2018b. "Quantifying Confidence." *Econometrica* 86 (5): 1689–1726.

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry. 2020. "Imperfect Macroeconomic Expectations: Evidence and Theory." NBER Working Paper 27308.

Angeletos, George-Marios, and Jennifer La'O. 2013. "Sentiments." *Econometrica* 81 (2): 739–79.

Bacchetta, Philippe, Elmar Mertens, and Eric Van Wincoop. 2009. "Predictability in Financial Markets: What Do Survey Expectations Tell Us?" *Journal of International Money and Finance* 28 (3): 406–26.

Barber, Brad M., and Terrance Odean. 2000. "Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors." *Journal of Finance* 55 (2): 773–806.

Barberis, Nicholas, Robin Greenwood, Lawrence Jin, and Andrei Shleifer. 2015. "X-CAPM: An Extrapolative Capital Asset Pricing Model." *Journal of Financial Economics* 115 (1): 1–24.

Barberis, Nicholas, Andrei Shleifer, and Robert W. Vishny. 1998. "A Model of Investor Sentiment." *Journal of Financial Economics* 49 (3): 307–43.

Ben-David, Itzhak, John R. Graham, and Campbell R. Harvey. 2013. "Managerial Miscalibration." *Quarterly Journal of Economics* 128 (4): 1547–84.

Bhandari, Anmol, Jaroslav Borovicka, and Paul Ho. 2019. "Survey Data and Subjective Beliefs in Business Cycle Models." Federal Reserve Bank of Richmond Working Paper 19–14.

Bianchi, Francesco, Cosmin Ilut, and Martin Schneider. 2018. "Uncertainty Shocks, Asset Supply and Pricing over the Business Cycle." *Review of Economic Studies* 85 (2): 810–54.

Bianchi, Francesco, Sydney C. Ludvigson, and Sai Ma. 2022. "Replication Data for: Belief Distortions and Macroeconomic Fluctuations." American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/E162101V1.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer. 2019. "Diagnostic Expectations and Stock Returns." *Journal of Finance* 74 (6): 2839–74.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. 2018. "Over-reaction in Macroeconomic Expectations." NBER Working Paper 24932.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2018. "Diagnostic Expectations and Credit Cycles." *Journal of Finance* 73 (1): 199–227.

Bouchaud, Jean-Philippe, Philipp Krueger, Augustin Landier, and David Thesmar. 2019. "Sticky Expectations and the Profitability Anomaly." *Journal of Finance* 74 (2): 639–74.

Chauvet, Marcelle, and Simon Potter. 2013. "Forecasting Output." In *Handbook of Economic Forecasting*, Vol. 2, edited by Graham Elliott and Allan Timmermann, 141–94. Amsterdam: Elsevier.

Coibion, Olivier, and Yuriy Gorodnichenko. 2015. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (8): 2644–78.

Coibion, Olivier, and Yuriy Gorodnichenko. 2012. "What can survey forecasts tell us about information rigidities?" *Journal of Political Economy* 120 (1): 116–59.

Curtin, Richard Thomas. 2019. *Consumer Expectations: Micro Foundations and Macro Impact.* New York: Cambridge University Press.

d'Agostino, Antonello, Domenico Giannone, and Paolo Surico. 2006. "(Un) Predictability and Macroeconomic Stability." Unpublished.

Daniel, Kent, and David Hirshleifer. 2015. "Overconfident Investors, Predictable Returns, and Excessive Trading." *Journal of Economic Perspectives* 29 (4): 61–88.

Daniel, Kent, David Hirshleifer, and Avanidhar Subrahmanyam. 1998. "Investor Psychology and Security Market Under-and Overreactions." *Journal of Finance* 53 (6): 1839–85.

Daniel, Kent D., David Hirshleifer, and Avanidhar Subrahmanyam. 2001. "Overconfidence, Arbitrage, and Equilibrium Asset Pricing." *Journal of Finance* 56 (3): 921–65.

De Bondt, Werner F. M., and Richard H. Thaler. 1990. "Do Security Analysts Overreact?" *American Economic Review* 80 (2): 52–57.

De Long, J. Bradford, Andrei Shleifer, Lawrence H. Summers, and Robert J. Waldmann. 1990. "Positive Feedback Investment Strategies and Destabilizing Rational Speculation." *Journal of Finance* 45 (2): 379–95.

Del Negro, Marco, and Stefano Eusepi. 2011. "Fitting Observed Inflation Expectations." *Journal of Economic Dynamics and Control* 35 (12): 2105–31.

Dotsey, Michael. 1998. "The Predictive Content of the Interest Rate Term Spread for Future Economic Growth." *Federal Reserve Bank of Richmond Economic Quarterly* 84 (3): 31–51.

Epstein, Larry G., and Martin Schneider. 2010. "Ambiguity and Asset Markets." *Annual Review of Financial Economics* 2 (1): 315–46.

Estrella, Arturo, and Gikas Hardouvelis. 1990. "Possible Roles of the Yield Curve in Monetary Analysis." In *Intermediate Targets and Indicators for Monetary Policy: A Critical Survey*, 339–62. New York: Federal Reserve Bank of New York.

Estrella, Arturo, and Gikas Hardouvelis. 1991. "The Term Structure as a Predictor of Real Economic Activity." *Journal of Finance* 46 (2): 555–76.

Estrella, Arturo, and Frederic S. Mishkin. 1998. "Predicting U.S. Recessions: Financial Variables as Leading Indicators." *Review of Economics and Statistics* 80 (1): 45–61.

Eusepi, Stefano, and Bruce Preston. 2011. "Expectations, Learning, and Business Cycle Fluctuations." *American Economic Review* 101 (6): 2844–72.

Faust, Jon, and Jonathan H. Wright. 2013. "Forecasting Inflation." In *Handbook of Economic Forecasting*, Vol. 2, edited by Graham Elliott and Allan Timmermann, 2–56. Amsterdam: Elsevier.

Fuhrer, Jeffrey C. 2018. "Intrinsic Expectations Persistence: Evidence from Professional and Household Survey Expectations." Federal Reserve Bank of Boston Working Paper 18-9.

Gabaix, Xavier. 2014. "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 129 (4): 1661–1710.

Gabaix, Xavier. 2020. "A Behavioral New Keynesian Model." *American Economic Review* 110 (8): 2271–2327.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer. 2016. "Expectations and Investment." *NBER Macroeconomics Annual* 30 (1): 379–431.
Gennaioli, Nicola, and Andrei Shleifer. 2018. *A Crisis of Beliefs: Investor Psychology and Financial Fragility.* Princeton: Princeton University Press.
Genre, Véronique, Geoff Kenny, Aidan Meyler, and Allan Timmermann. 2013. "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting* 29 (1): 108–21.
Giacomini, Raffaella, and Halbert White. 2006. "Tests of Conditional Predictive Ability." *Econometrica* 74 (6): 1545–78.
Greenwood, Robin, and Samuel G. Hanson. 2015. "Waves in Ship Prices and Investment." *Quarterly Journal of Economic*s 130 (1): 55–109.
Greenwood, Robin, Samuel G. Hanson, Andrei Shleifer, and Jakob Ahm Sørensen. 2020. "Predictable Financial Crises." NBER Working Paper 27396.
Greenwood, Robin, and Andrei Shleifer. 2014. "Expectations of Returns and Expected Returns." *Review of Financial Studies* 27 (3): 714–46.
Hamilton, James D. 1989. "A New Approach to the Economic Analysis of Nonstationary Returns and the Business Cycle." *Econometrica* 57 (2): 357–84.
Hamilton, James D. 2018. "Why You Should Never Use the Hodrick-Prescott Filter." *Review of Economics and Statistics* 100 (5): 831–43.
Hansen, Lars Peter, and Thomas J. Sargent. 2008. *Robustness.* Princeton: Princeton University Press.
Harvey, Campbell R. 1989. "Forecasts of Economic Growth from the Bond and Stock Markets." *Financial Analysts Journal* 45 (5): 38–45.
Haubrich, Joseph G., and Ann M. Dombrosky. 1996. "Predicting Real Growth Using the Yield Curve." *Economic Review* 32 (1): 26–35.
Ilut, Cosmin L., and Hikaru Saijo. 2021. "Learning, Confidence, and Business Cycles." *Journal of Monetary Economics* 117: 354–76.
Ilut, Cosmin L., and Martin Schneider. 2014. "Ambiguous Business Cycles." *American Economic Review* 104 (8): 2368–99.
Jorda, Oscar. 2005. "Estimation and Inference of Impulse Responses by Local Projections." *American Economic Review* 95 (1): 161–82.
Jurado, Kyle, Sydney C. Ludvigson, and Serena Ng. 2015. "Measuring Uncertainty." *American Economic Review* 105 (3): 1177–1216.
Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.* New York: Little, Brown Spark.
Khaw, Mel Win, Luminita Stevens, and Michael Woodford. 2017. "Discrete Adjustment to a Changing Environment: Experimental Evidence." *Journal of Monetary Economics* 91: 88–103.
Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh. 2020. "Shrinking the Cross-Section." *Journal of Financial Economics* 135 (2): 271–92.
Kozicki, Sharon. 1997. "Predicting Real Growth and Inflation with the Yield Spread." *Economic Review, Federal Reserve Bank of Kansas City* 82 (4): 39–57.
Ludvigson, Sydney C., Sai Ma, and Serena Ng. 2021. "Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response?" *American Economic Journal: Macroeconomics* 13 (4): 369–410.
Ludvigson, Sydney C., and Serena Ng. 2007. "The Empirical Risk-Return Relation: A Factor Analysis Approach." *Journal of Financial Economics* 83 (1): 171–222.
Ludvigson, Sydney C., and Serena Ng. 2009. "Macro Factors in Bond Risk Premia." *Review of Financial Studies* 22 (12): 5027–67.
Ludvigson, Sydney C., and Serena Ng. 2010. "A Factor Analysis of Bond Risk Premia." In *Handbook of Empirical Economics and Finance*, Vol. 1, edited by Aman Ulah and David E. A. Giles, 313–72. Boca Raton, FL: Chapman and Hall.
Malmendier, Ulrike, and Stefan Nagel. 2011. "Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?" *Quarterly Journal of Economics* 126 (1): 373–416.
Malmendier, Ulrike, and Stefan Nagel. 2016. "Learning from Inflation Experiences." *Quarterly Journal of Economics* 131 (1): 53–87.
Mankiw, N. Gregory, and Ricardo Reis. 2002. "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117 (4): 1295–1328.
Martin, Ian, and Stefan Nagel. 2019. "Market Efficiency in the Age of Big Data." Unpublished.
Masolo, Riccardo M., and Francesca Monti. 2021. "Ambiguity, Monetary Policy, and Trend Inflation." *Journal of the European Economic Association* 19 (2): 839–71.
Milani, Fabio. 2011. "Expectation Shocks and Learning as Drivers of the Business Cycle." *Economic Journal* 121 (552): 379–401.

Milani, Fabio. 2017. "Sentiment and the U.S. Business Cycle." *Journal of Economic Dynamics and Control* 82: 289–311.

Morris, Stephen, Hyun Song Shin, and Muhamet Yildiz. 2016. "Common Belief Foundations of Global Games." *Journal of Economic Theory* 163: 826–48.

Odean, Terrance. 1998. "Volume, Volatility, Price, and Profit When All Traders Are above Average." *Journal of Finance* 53 (6): 1887–1934.

Pesaran, M. Hashem, and Allan Timmermann. 2007. "Selection of Estimation Window in the Presence of Breaks." *Journal of Econometrics* 137 (1): 134–61.

Plosser, Charles I., and K. Geert Rouwenhorst. 1994. "International Term Structures and Real Economic Growth." *Journal of Monetary Economics* 33 (1): 133–55.

Reis, Ricardo. 2006a. "Inattentive Consumers." *Journal of Monetary Economics* 53 (8): 1761–1800.

Reis, Ricardo. 2006b. "Inattentive Producers." *Review of Economic Studies* 73 (3): 793–821.

Sims, Christopher A. 2003. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (3): 665–90.

Stock, James H., and Mark W. Watson. 1989. "New Indexes of Coincident and Leading Economic Indicators." In *NBER Macroeconomics Annual: 1989*, Vol. 4, edited by Olivier J. Blanchard and Stanley Fischer, 351–94. Cambridge, MA: MIT Press.

Stock, James H., and Mark W. Watson. 1991. "A Probability Model of the Coincident Economic Indicators." In *Leading Economic Indicators: New Approaches and Forecasting Records*, edited by G. Moore and K. Lahiri, 63–90. Cambridge, MA: Cambridge University Press.

Stock, James H., and Mark W. Watson. 2002a. "Forecasting Using Principal Components from a Large Number of Predictors." *Journal of the American Statistical Association* 97 (460): 1167–79.

Stock, James H., and Mark W. Watson. 2002b. "Macroeconomic Forecasting Using Diffusion Indexes." *Journal of Business and Economic Statistics* 20 (2): 147–62.

Stock, James H., and Mark W. Watson. 2006. "Forecasting with Many Predictors." In *Handbook of Economic Forecasting*, Vol. 1, edited by M. Hashem Pesaran and Martin Weale, 515–54. Oxford, United Kingdom: Elsevier.

Woodford, Michael. 2002. "Imperfect Common Knowledge and the Effects of Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by Phillippe Aghion, Roman Frydman, Joseph Stiglitz, and Michael Woodford, 25–58. Princeton: Princeton University Press.

Woodford, Michael. 2013. "Macroeconomic Analysis without the Rational Expectations Hypothesis." *Annual Review of Economics.* 5 (1): 303–46.