The Ronald O. Perelman Center for Political Science and Economics (PCPSE)
133 South 36th Street
Philadelphia, PA 19104-6297

pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper
24-032

Behavioral Foundations of Model Misspecification

J. AISLINN BOHREN
University of Pennsylvania

DANIEL N. HAUSER
Aalto University

August 6, 2024

Behavioral Foundations of Model Misspecification*

J. Aislinn Bohren† Daniel N. Hauser‡

August 6, 2024

We link two approaches to biased belief formation: non-Bayesian updating and misspecified models. The former parameterizes a bias with an updating rule mapping signals to posterior beliefs or a belief forecast describing anticipated beliefs; the latter is an incorrect model of the signal generating process. Our main result derives necessary and sufficient conditions for an updating rule and belief forecast to have a misspecified model representation, shows that these two components uniquely pin down a representation, and constructs it. This clarifies the belief restrictions implicit in the misspecified model approach. It also allows leveraging of the distinct advantages of each approach by decomposing a model into empirically identifiable components, showing these components isolate the two forms of bias that the model encodes—the retrospective bias after information arrives and the prospective bias beforehand, and rendering off-the-shelf tools to characterize asymptotic learning and equilibrium predictions in misspecified models applicable to non-Bayesian updating.

KEYWORDS: Model misspecification, belief formation, learning, non-Bayesian updating, heuristics

*We thank Nageeb Ali, Cuimin Ba, Renee Bowen, Sylvain Chassang, In-Koo Cho, Wouter Dessein, Hanming Fang, Mira Frick, Kevin He, Ryota Iijima, Alex Imas, Navin Kartik, Qingmin Liu, Jawaad Noor, Ran Spiegler, and seminar and conference participants at various institutions for helpful comments and suggestions. Cuimin Ba, Matt Murphy and Marcus Tomaino provided excellent research assistance. Bohren gratefully acknowledges financial support from NSF grant SES-1851629 and PIER. Both authors would like to thank the hospitality of the Study Center Gerzensee where part of this research was conducted.

†University of Pennsylvania; Email: abohren@sas.upenn.edu

‡Aalto University School of Business and Helsinki GSE; Email: daniel.hauser@aalto.fi

1 Introduction

Extensive work in economics and psychology has documented systematic biases and errors individuals exhibit when interpreting information and forming beliefs. A rich literature has explored how to model such biased updating. Two modeling approaches are commonly used: the 'non-Bayesian' approach and the 'misspecified model' approach. The former parameterizes a particular bias with an updating rule that maps signal realizations to posterior beliefs (e.g., under- and overreaction in Epstein, Noor, and Sandroni (2010)). In the latter, a subjective model of the signal generating process describes how an individual interprets signals; the individual applies Bayes rule to this model to form beliefs, but the model may be wrong.

Each approach has distinct advantages. The misspecified model approach can capture a variety of behavioral biases without departing too far from the standard framework. It is therefore relatively easy to incorporate this approach in existing economic models. Moreover, a misspecified model also pins down an individual's anticipated beliefs before observing information, which can be relevant for ex-ante decisions, strategic interaction, and social learning. Finally, the approach is amenable to analysis in a general context. A large literature establishes general learning characterizations for misspecified models (e.g., Bohren and Hauser (2021); Fudenberg, Lanzani, and Strack (2021); Frick, Iijima, and Ishii (2023)) and develops a general solution concept—Berk-Nash equilibrium (Esponda and Pouzo 2016).

In contrast, the non-Bayesian approach provides a transparent link between the conceptual form of the bias (e.g., overprecision, partisan bias) and the resulting belief distortion, highlighting the specific way in which an agent distorts information. For example, the agent may miscode certain signal realizations, double-count signals, or slant beliefs in a particular direction. This connection to the underlying psychological friction allows for empirically validated modeling choices (e.g., the updating rules used in Woodford (2020); Ba, Bohren, and Imas (2024)). Additionally, this is the approach often used in empirical work, as an updating rule can be identified from belief data (Benjamin 2019). Importantly, however, the approach is incomplete in that it does not pin down anticipated beliefs. The analysis is also typically conducted on a case-by-case basis to understand how a specific updating rule impacts learning, to determine which solution concept to pair with an updating rule, or to pin down antici-

pated beliefs. For example, Rabin and Schrag (1999); Epstein et al. (2010) study how confirmation bias and over/underreaction, respectively, impact asymptotic beliefs, Eyster and Rabin (2010) define a solution concept for naive learning, and Benjamin, Bodoh-Creed, and Rabin (2019) outline an assumption to pin down anticipated beliefs in a setting with base-rate neglect. This contrasts with the misspecified model approach, which provides a general and complete framework for studying biases but little guidance on how to capture a specific bias.

The goal of this paper is to link these two approaches in order to leverage the advantages of each. We first determine when it is possible to represent an updating rule as a misspecified model, in the sense that the model prescribes the same posterior belief as the updating rule following each signal realization. While we show that such a representation exists for many commonly used updating rules, in general, this representation is not unique. We next show that an agent's *belief forecast*—her prediction of her future beliefs—is the other component needed to pin down a unique representation. Importantly for empirical work, a belief forecast is also identifiable from belief data.[1] Bringing these pieces together, our main result establishes necessary and sufficient conditions for a given updating rule and belief forecast to be jointly represented by a misspecified model and constructs this unique representation.

From the perspective of the misspecified model approach, this result clarifies its belief formation restrictions, decomposes the model into empirically identifiable components, and highlights how these components isolate the two forms of bias that a given model encodes—the induced updating rule captures the retrospective bias that emerges after information arrives, while the induced belief forecast captures the prospective bias that emerges before information arrives. It also provides a powerful tool to construct misspecified models that capture a given psychological friction. From the perspective of the non-Bayesian approach, this result provides guidance on how to incorporate a given updating rule into economic settings that also require ex-ante beliefs, and yields a set of off-the-shelf tools that can be used to immediately establish important results such as belief convergence and equilibrium characterization.

We now describe our setting in more detail. An agent learns about a hidden state

[1] See Chambers and Lambert (2021); Karni (2020) for methods to elicit an agent's prediction of her own belief.

from a signal with a fixed true distribution. The non-Bayesian approach consists of an updating rule mapping each signal realization to a posterior belief and a belief forecast describing the agent's anticipated distribution of her posterior belief after observing the signal. This set-up draws a distinction between the *prospective* bias of the agent—how the agent reasons about information yet to be realized via the belief forecast—and the *retrospective* bias—how the agent reasons about realized information via the updating rule. The misspecified model approach consists of a family of subjective distributions over the signal space, one for each state. A model is misspecified when it differs from the true (objective) signal distribution. We say a misspecified model *represents* an updating rule when the posterior belief prescribed by the updating rule is equal to the posterior belief derived from Bayesian updating with respect to the misspecified model, and it represents a belief forecast when the predicted distribution of the posterior belief derived from the misspecified model is equal to the forecast.

We first derive necessary and sufficient conditions for an updating rule or a forecast to be individually represented. The condition for the updating rule is quite mild: it must be *Bayes-feasible*, in that the prior belief is contained in the relative interior of the convex hull of the set of posterior beliefs prescribed by the updating rule. This rules out updating rules that, for example, move beliefs towards the same state following all signal realizations and is satisfied by many updating rules commonly used in the literature.[2] The condition for a belief forecast is more restrictive: it must be *plausible*, in that its expectation is equal to the prior. This is a misspecified analogue of Bayes plausibility (Kamenica and Gentzkow 2011). In both cases, it is straightforward to see that these conditions are necessary implications of Bayesian updating, as required in a misspecified model; it turns out that they are also sufficient.

Theorem 1 brings together these results to establish necessary and sufficient conditions for an updating rule and a forecast to be *jointly* represented by the same misspecified model. In addition to the two conditions described above, a third—*no unexpected beliefs*—is needed. It requires any set of posterior beliefs that the agent anticipates with positive probability to arise with positive probability given her updating rule and vice versa. Together, these conditions clarify the belief formation

[2]For example, overreaction (Epstein et al. 2010), partisan bias (Bohren and Hauser 2021) and confirmation bias (Rabin and Schrag 1999) all satisfy this condition.

restrictions implicit in using the misspecified model approach: (i) Bayes-feasible updating rules, (ii) plausible belief forecasts, and (iii) no unexpected beliefs. We show that the second and third condition imply the first, so (i) is redundant. Therefore, any updating rule and belief forecast satisfying (ii) and (iii) have a misspecified model representation. In a sufficiently rich signal space condition (iii) is mild, so a given updating rule is compatible with many different forecasts and similarly for a given forecast. This flexibility means that the prospective bias of a misspecified model places little restriction on the retrospective bias of the model (and vice versa).

Importantly, Theorem 1 also shows that such a representation is unique and easy to construct. This establishes that a misspecified model can be uniquely decomposed into the prospective and retrospective biases that it encodes. The prospective bias reflected in the belief forecast captures errors in anticipating belief formation, while the retrospective bias reflected in the updating rule captures errors in interpreting information after it arrives. Every misspecified model is uniquely identified by these two components, and they jointly describe all bias that the model encodes. This provides a convenient formulation for a misspecified model in terms of the resulting biases—and also, in terms of components that can be identified from belief data. Moreover, it establishes that the induced updating rule and belief forecast together pin down all behavioral implications of a misspecified model; the model imposes no further belief distortions beyond those reflected in these two components. From the perspective of the non-Bayesian approach, this result establishes that for any given updating rule, selecting a belief forecast uniquely pins down a misspecified model that can be used for analysis. Finally, Theorem 1 constructs the misspecified model representation of a given updating rule and belief forecast. This provides a simple formula that can be easily used in applications.

We next derive classes of models that feature only prospective bias or only retrospective bias. The sophisticated-prospectively correct (sophisticated-PC) and naive-prospectively correct (naive-PC) models both shut down prospective bias—based on whether an agent is sophisticated or naive about her retrospective bias—in order to isolate the implications of retrospective bias. Given that updating rules are more frequently studied in the literature, this provides a natural way to choose a belief fore-

cast when retrospective bias is the primary focus.[3] Analogously, the retrospectively correct (RC) model shuts down retrospective bias in order to isolate the implications of prospective bias. Our results derive necessary and sufficient conditions for each of these models to exist and be unique. Taken together, these classes of models serve dual purposes. When a researcher starts with an updating rule or belief forecast and wants to use the misspecified model approach for analysis, they highlight which representation to select in order to avoid introducing any additional bias. When a researcher starts with a misspecified model that generates both retrospective and prospective bias, they provide natural benchmarks that shut down each bias in turn, thereby isolating the effect of the other and providing insight into their interaction.

We next develop two applications to demonstrate how our results yield novel insights in specific economic settings. The first shows the power of our representation for conducting a general analysis of a behavioral bias. We apply results from the misspecified learning literature to characterize the long-run learning outcomes of a general version of the confirmation bias updating rule from Rabin and Schrag (1999). We show that the prediction of incorrect learning from Rabin and Schrag (1999) robustly emerges in an individual learning setting, but the belief forecast plays a crucial role in determining learning outcomes in a social learning setting. The second explores a search decision when a firm has a misspecified model of the signal variance. We decompose this model into the retrospective and prospective biases it encodes, and then use the prospectively and retrospectively correct representations to isolate the impact of each bias on search behavior. We show that whether bias emerges ex-ante versus ex-post has important implications for behavior, and the extent of the misspecification determines whether the prospective and retrospective bias offset or amplify each other. This has important implications for choosing policy interventions, including which bias is more important to target and whether targeting one bias but not the other will lead to further inefficiency.

We close with several extensions of our setting, including allowing for the possibility that an agent has a misspecified prior and how time inconsistency can emerge in a dynamic version of our framework.

[3]For example, Benjamin et al. (2019) pair an updating rule that features base rate neglect with a forecast that corresponds to our notion of naive-PC.

Literature Review. Model misspecification is a popular approach for capturing behavioral biases. In a variety of general settings, recent work has developed the solution concept 'Berk-Nash equilibrium' (Esponda and Pouzo 2016), characterized the asymptotic beliefs of misspecified learning (Bohren and Hauser 2021; Fudenberg et al. 2021; Frick et al. 2023; Esponda, Pouzo, and Yamamoto 2021), and explored robustness to perturbations of the model (Frick, Iijima, and Ishii 2020; Bohren and Hauser 2021).[4] Papers have also studied the implications of misspecified learning for a variety of specific biases, including overconfidence (Heidhues, Koszegi, and Strack 2018), gambler's fallacy (He 2022), selective attention (Schwartzstein 2014) and omitted variable bias (Mailath and Samuelson 2020; Levy, Razin, and Young 2022). Our paper shows how a non-Bayesian updating rule can be translated to a misspecified model, allowing for analysis using these general results.

Another strand of literature seeks to provide a foundation for which misspecified models arise and persist (Ba 2024; Fudenberg and Lanzani 2023; He and Libgober 2024; Frick, Iijima, and Ishii 2024; Lanzani 2024). One of the classes of models we consider—prospectively correct models—are naturally robust to many of these criteria. In such models, the misspecified agent correctly anticipates the unconditional distribution of signals. This is analogous to conditions used to correct misspecified models in Espitia (2021); Spiegler (2020); Mailath and Samuelson (2020) and solution concepts such as cursed equilibrium (Eyster and Rabin 2005), behavioral equilibrium (Esponda 2008), and analogy-based expectation equilibrium (Jehiel 2005).

The misspecified model approach assumes that an agent updates using Bayes rule. A number of papers characterize properties of posteriors that arise from Bayesian updating. Shmaya and Yariv (2016) derive a similar result to our Lemma 1 on how the set of posteriors that can arise from an information structure relate to the prior. Similarly, the belief forecast in our framework is analogous to the unconditional distribution over posterior beliefs in the Bayesian persuasion (Kamenica and Gentzkow 2011). Molavi (2024) shows that any distribution over posteriors satisfying very mild assumptions can be induced via Bayes rule with respect to a misspecified model. His condition is weaker than ours, as he allows the misspecified model to put positive probability on signals outside the support of the true model.

[4]Early papers in this literature include Arrow and Green (1973); Nyarko (1991).

There is also a recent literature that provides a foundation for general classes of non-Bayesian updating rules and draws parallels between the structure of non-Bayesian and Bayesian updating (Epstein et al. 2010; Cripps 2018; Chauvin 2020; Jakobsen 2023; de Clippel and Zhang 2022). In contrast, we characterize the properties of updating rules that emerge from Bayesian updating with respect to a misspecified model. Other work characterizes properties of specific non-Bayesian updating rules. He and Xiao (2017) describe a class of updating rules in which sequential and simultaneous processing of multiple signals lead to the same posterior. Benjamin et al. (2019) study an updating rule that features base rate neglect and pin down prospective beliefs by assuming an agent believes she will use Bayes rule to update in the future. This is similar in spirit to our naive-PC model.

Benjamin, Rabin, and Raymond (2016) first highlighted the need to distinguish between how an agent retrospectively processes information versus prospectively predicts she will process information in models of non-Bayesian updating. They draw this distinction in relation to how an agent groups multiple signals for processing and highlight how different perceived versus actual groupings can lead to time-inconsistency. In contrast, our distinction separates prospective versus retrospective bias that emerges with respect to a single signal (or more generally, a fixed grouping of signals): specifically, it distinguishes between the anticipation of how a signal will be interpreted versus how it is actually interpreted. This does not necessarily lead to time-inconsistency, as we further discuss in Section 6.

Much of the literature on biased belief formation focuses on prospective or retrospective bias in isolation. The work on misspecified causal graphs (e.g., Spiegler (2016)) and Berk-Nash equilibrium (Esponda and Pouzo 2016) take a prospective perspective, focusing on how an agent (incorrectly) predicts what will happen after her decision. Papers such as Heidhues et al. (2018); Levy et al. (2022) and most of the behavioral literature modelling and empirically documenting specific updating biases (see Benjamin (2019)) focus on retrospective bias.[5] But in many economic settings, such as the search application in Section 5.2 and the strategic interactions studied

[5]A notable exception is Le Yaouanq and Schwardmann (2022), which measures both retrospective and prospective bias when learning from past behavior in an experimental real-effort task. They find that participants are retrospectively unbiased but underestimate their future learning.

in Bohren and Hauser (2021); He (2022); Frick et al. (2024), both prospective and retrospective bias play a key role in determining beliefs and behavior. For example, using our decomposition, Bohren and Hauser (2023) show how retrospective and prospective bias differentially impact an optimal lending contract.

Recent work on the wisdom of the crowd explores how higher order beliefs impact identification. Prelec and McCoy (2022); Libgober (2024) show that if many agents draw signals from the same model (i.e., information structure), then knowing an agent's posterior belief and her belief about the distribution of others' beliefs identifies the model. This relates to our insight that eliciting an updating rule must be paired with a component describing the distribution over beliefs to identify a unique misspecified model.

2 Model

2.1 States, Priors, and Signals

Nature selects one of N *states* of the world $\omega \in \Omega \equiv \{\omega_1, \omega_2, \ldots, \omega_N\}$ according to full support *prior* $p \equiv (p_1, ..., p_N) \in \Delta(\Omega)$. An agent learns about the state by observing a *signal* z, which is drawn from measurable space $(\mathcal{Z}, \mathcal{F})$, where $\mathcal{Z}$ is the set of signal realizations and $\mathcal{F}$ is a σ-algebra over $\mathcal{Z}$. Let $\mu_i(\cdot; p) \in \Delta(\mathcal{Z})$ denote the *true signal distribution* conditional on state ω_i at prior p, $\mu(\cdot; p) \equiv \sum_{i=1}^{N} p_i \mu_i(\cdot; p)$ denote the *true unconditional signal distribution* at p, and refer to family of signal distributions $\{\mu_i(\cdot; p)\}_{\omega_i \in \Omega}$ as the *true model at* p. The distribution of the signal can depend on the prior to capture settings where information is endogenously generated by an ex-ante action choice. To ensure that no signal perfectly rules out a state, assume $\mu_i(\cdot; p)$ and $\mu_j(\cdot; p)$ are mutually absolutely continuous for each $\omega_i, \omega_j \in \Omega$ and all $p \in \Delta(\Omega)$. For technical reasons, assume that there exists a σ-finite reference measure ν on $(\mathcal{Z}, \mathcal{F})$ such that $\mu_i(\cdot; p)$ is absolutely continuous with respect to ν for all $\omega_i \in \Omega$ and $p \in \Delta(\Omega)$.[6] This environment captures many common signal structures used in the literature, including real-valued continuous signals, finite signals, multidimensional signals, and causal graphs (Spiegler (2016)).

[6] Defining a reference measure that dominates the other measures allows for the consideration of multiple types of signal spaces within the same framework (e.g., discrete and continuous).

2.2 Modeling Bias in Belief Updating

We introduce two approaches used to model bias in interpreting the signal: (i) the non-Bayesian approach, where an updating rule maps each signal realization to a posterior belief and a belief forecast specifies a prediction of future beliefs; and (ii) the misspecified model approach, where beliefs are derived from Bayesian updating with respect to an incorrect model of the signal process. In both approaches we assume that the agent has a correct prior belief; the analysis immediately extends to a misspecified prior $\hat{p} \neq p$ (see Section 3.4).

The Non-Bayesian Approach. This approach is often used in the behavioral learning literature (see Benjamin (2019) for review). An *updating rule* describes an agent's posterior belief after each possible signal realization.

Definition 1 (Updating Rule)**.** *An* updating rule $h : \mathcal{Z} \times \Delta(\Omega) \to \Delta(\Omega)$ *is a measurable function that maps each signal realization and prior belief to a posterior belief.*

Given prior belief p and signal realization $z \in \mathcal{Z}$, updating rule $h(z, p)$ assigns probability $h(z, p)_i$ to state ω_i. The requirement that h is measurable rules out randomness in updating conditional on the signal. We restrict attention to updating rules that do not (incorrectly) interpret any signals as perfectly ruling out a state and map certainty to certainty: $h(z, p)_i = 0$ iff $p_i = 0$ and $h(z, p)_i = 1$ iff $p_i = 1$. A special case is Bayesian updating with respect to the true signal distribution:

$$h_B(z, p)_i \equiv \frac{p_i \frac{d\mu_i}{d\nu}(z; p)}{\sum_{j=1}^{N} p_j \frac{d\mu_j}{d\nu}(z; p)}, \tag{1}$$

with $0/0 = 0$ by convention. An updating rule is *biased* at p if it differs from Bayesian updating, $h(\cdot; p) \neq h_B(\cdot; p)$ on a $\mu_1(\cdot; p)$-positive measure set of signals. We refer to bias that stems from the updating rule as *retrospective bias*, since it arises after the signal is realized. Definition 1 nests many biased updating rules used in the literature, including settings where an updating rule distorts the Bayesian posterior or true signal likelihood (in the former case, the updating rule can be written with h_B as an argument, as is sometimes done in the literature). The following example illustrates several such updating rules (see Section 5.1 for a confirmation bias example).[7]

[7]These examples echo updating rules used in the following papers: (a) Epstein et al. (2010) (b) & (f) Grether (1980); Benjamin (2019) (c) Thaler (2021); Benjamin (2019); Bohren and Hauser

Example 1 (Common Updating Rules). *Suppose* $\Omega = \{\omega_1, \omega_2\}$ *and* $\mathcal{Z} \subset [0,1]$. *In the binary state case, the posterior belief is pinned down by the belief* $h(z,p)_2$ *that the state is* ω_2. *Normalize the signal to be the Bayesian posterior belief of* ω_2 *following a flat prior,* $z = h_B(z, 0.5)_2$. *This normalization implies* $h_B(z,p)_2 = p_2 z/(p_2 z + p_1(1-z))$.

(a) Linear under/overreaction: $h(z,p)_2 = \alpha h_B(z,p)_2 + (1-\alpha)p_2$ *with* $\alpha \in (0,1)$ *for underreaction to the signal and* $\alpha > 1$ *for overreaction.*

(b) Geometric under/overreaction: $\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_2}{p_1}\left(\frac{z}{1-z}\right)^{\alpha}$ *with* $\alpha \in (0,1)$ *for underreaction to the signal and* $\alpha > 1$ *for overreaction.*

(c) Partisan bias: $h(z,p)_2 = h_B(z,p)_2^{\alpha}$ *(distort posterior) or* $\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_2}{p_1}\left(\frac{z^{\alpha}}{1-z^{\alpha}}\right)$ *(distort signal likelihood) with* $\alpha \in (0,1)$ *for slanting the posterior belief towards* ω_2 *and* $\alpha > 1$ *for slanting towards* ω_1.

(d) Cognitive noise: $h(z,p)_2 = \alpha h_B(z,p)_2 + (1-\alpha)p_d$ *for* $\alpha \in (0,1)$ *and cognitive default* $p_d \in (0,1)$ *(typically,* p_d *is chosen to be uniform).*

(e) Base rate neglect: $\frac{h(z,p)_2}{h(z,p)_1} = \left(\frac{p_2}{p_1}\right)^{\alpha}\left(\frac{z}{1-z}\right)$ *with* $\alpha \in (0,1)$ *underweighs the prior.*

(f) Complexity reduction: given finite interval partition $\{Z_1, Z_2, ..., Z_K\}$ *of* $\mathcal{Z}$ *and set of posteriors* $\{x_1, ..., x_K\} \subset [0,1]$, $h(z,p)_2 = x_j$ *for all* $z \in Z_j$ *is a coarse updating rule where intervals of signals are mapped to the same posterior.*

While an updating rule captures how the agent interprets the signal after it arrives, it does not specify her ex-ante beliefs about the signal and how she will interpret it (e.g., is she naive or sophisticated about her retrospective bias; does she exhibit other forms of bias ex-ante). Such prospective beliefs are a crucial component of many economic settings, including settings with a decision before information arrives (e.g., what information to acquire or pay attention to, whether to pursue a new project before learning about its quality), strategic interaction (e.g., expectations about how an opponent's action will vary with the signal), and social learning (e.g., expectations about how an opponent's action reflects his private signal). In the latter two cases, the relevant prospective beliefs capture ex-ante beliefs about the signal distribution and how *others* interpret it. We define a *belief forecast* to capture this component, which specifies the agent's subjective distribution over posterior beliefs.

Definition 2 (Belief Forecast). *A* belief forecast $\hat{\rho}(\cdot; p) \in \Delta(\Delta(\Omega))$ *specifies a Borel*

(2021) (d) Woodford (2020) (e) Jakobsen (2023); Mullainathan (2002)

probability measure over the posterior at prior p, such that there exists a measurable $\phi : \mathcal{Z} \to \Delta(\Omega)$ with $\mu_1(\phi^{-1}(\cdot); p)$ and $\hat{\rho}(\cdot; p)$ mutually absolutely continuous.

The second part of the definition is analogous to the assumption that h is measurable: it rules out perceived randomness in the posterior beyond that stemming from the signal by requiring that the support of the belief forecast is no "larger" than the support of the signal.[8]

A special case is the *accurate* belief forecast $\rho_h(\cdot; p)$ for updating rule h at prior p, which corresponds to the objective probability measure over the posterior when the agent uses updating rule h. Given a (Borel) set of posteriors $X \in \Delta(\Omega)$, it is equal to the probability of the set of signals that h maps to a posterior in X, where the probability is taken with respect to the ex-ante true signal distribution $\mu(\cdot; p)$:

$$\rho_h(X; p) = \mu(\{z : h(z, p) \in X\}; p). \tag{2}$$

Let $\rho_B(X; p)$ denote the accurate forecast with respect to Bayesian updating rule h_B—this is the true distribution of posteriors when the agent has no retrospective bias. Bias can also enter through the belief forecast. A belief forecast is biased if, given updating rule h, it differs from the accurate forecast, $\hat{\rho} \neq \rho_h$. We refer to such bias as *prospective bias*, since it arises before the signal is realized. The following example illustrates several biased belief forecasts.

Example 2 (Biased Belief Forecasts)**.** *Suppose $\Omega = \{\omega_1, \omega_2\}$. In the binary state case, the belief forecast is pinned down by a distribution over the posterior belief that the state is ω_2. Suppose the accurate forecast is uniform on $[0, 1]$.*

(a) Over/underprecision: the beta distribution $\beta(\alpha, \alpha)$ with $\alpha \in (0, 1)$ ($\alpha > 1$) overweighs (underweighs) the likelihood of precise beliefs.

(b) Partisan bias: the beta distribution $\beta(\alpha, \beta)$ with $\alpha > \beta$ ($\alpha < \beta$) places higher weight on beliefs favoring state ω_2 (ω_1).

(c) Complexity reduction: given finite interval partition $\{X_1, ..., X_K\}$ of $[0, 1]$ and

[8] With a finite support signal, this requires that the support of the forecast is no larger than the number of signal realizations. With an infinite support signal, the condition is more nuanced; it uses mutual absolute continuity to relate the measure-zero sets of the forecast and signal. The condition is defined with respect to μ_1, but it implies that $\mu_i(\phi^{-1}(\cdot); p)$ and $\hat{\rho}(\cdot; p)$ are m.a.c. for all $\omega_i \in \Omega$, given the assumption that $\mu_1(\cdot; p)$ and $\mu_i(\cdot; p)$ are m.a.c. for all $\omega_i \in \Omega$.

set of posteriors $\{x_1, ..., x_K\} \subset [0,1]$, $\hat{\rho}(x;p) > 0$ *iff* $x \in \{x_1, ..., x_K\}$ *is a coarse belief forecast where an agent entertains a finite number of posterior beliefs.*

The Misspecified Model Approach. In this approach, posterior beliefs and belief forecasts are pinned down by applying Bayes' rule to an agent's subjective model of the signal process. Let $\hat{\mu}_i(\cdot;p) \in \Delta(\mathcal{Z})$ denote the *perceived signal distribution* in state ω_i at prior p, $\hat{\mu}(\cdot;p) \equiv \sum_{i=1}^{N} p_i \hat{\mu}_i(\cdot;p)$ denote the perceived unconditional signal distribution, and refer to the family of signal distributions $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i \in \Omega}$ as the *subjective model at* p. An agent's model is *misspecified at* p if there exists an $\omega_i \in \Omega$ such that $\hat{\mu}_i(\cdot;p) \neq \mu_i(\cdot;p)$. Assume each perceived signal distribution $\hat{\mu}_i(\cdot;p)$ is mutually absolutely continuous with the true signal distribution $\mu_i(\cdot;p)$ for $\omega_i \in \Omega$ and $p \in \Delta(\Omega)$. This rules out 'unexpected signals' where a model assigns zero probability to a set of signal realizations that occur with positive probability and 'fake signals' where a model assigns positive probability to a set of signal realizations that occur with zero probability. It also implies that $\hat{\mu}_i(\cdot;p)$ and $\hat{\mu}_j(\cdot;p)$ are mutually absolutely continuous and $\hat{\mu}_i(\cdot;p)$ is absolutely continuous with respect to ν for each $\omega_i, \omega_j \in \Omega$ and $p \in \Delta(\Omega)$. This assumption is primarily technical, given that the subjective model can place arbitrarily small probability on sets of signals that the true model assigns positive probability to and vice versa (see Appendix D.3 for further discussion). Let $\mathcal{M}(p) \subset \Delta(\mathcal{Z})^N$ denote the set of subjective models $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i \in \Omega}$ such that $\hat{\mu}_i(\cdot;p)$ is mutually absolutely continuous with $\mu_i(\cdot;p)$ for all $\omega_i \in \Omega$. We refer to models in this set as *admissible at* p. The agent uses Bayes rule to form her posterior belief with respect to her subjective model. It follows from Bayes rule and mutual absolute continuity that prior p and model $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i \in \Omega}$ induce posterior belief $\hat{Pr}(\omega_i|z) = \frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z;p)}{\sum_{j=1}^{N} p_j \frac{d\hat{\mu}_j}{d\nu}(z;p)}$ that the state is ω_i following signal realization z, and belief forecast

$$\hat{Pr}(x \in X) = \hat{\mu}\left(\left\{z : \left(\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z;p)}{\sum_{j=1}^{N} p_j \frac{d\hat{\mu}_j}{d\nu}(z;p)}\right)_{\omega_i \in \Omega} \in X\right\}; p\right) \tag{3}$$

that the posterior belief x is in Borel set $X \subset \Delta(\Omega)$. The following example illustrates several misspecified models.

Example 3 (Examples of Misspecified Models)**.**

(a) Misperceived mean/variance: given true signal distribution $N(\omega_i, 1)$ *in state* ω_i*,*

believing signals are normally distributed with mean $\hat{\mu}_i \neq \omega_i$ or variance $\hat{\sigma} \neq 1$.

(b) Correlation neglect: given signal $z = (z_1, z_2)$ with correlation $\rho_i \neq 0$ in state ω_i, believing z_1 and z_2 are independent.

(c) Parametric approximation: using a parametric model for a nonparametric distribution (e.g., assuming normality).

(d) Complexity reduction: a model that assumes the same signal distribution for similar states when each state has a distinct distribution in the true model.

2.3 Defining a Representation

To connect the non-Bayesian and misspecified model approaches, we next define what it means for an updating rules or belief forecast to be *represented* as a subjective model, in that the subjective model induces the same posterior beliefs as the updating rule or the same distribution over posterior beliefs as the forecast.

Definition 3 (Subjective Model Representation)**.**

1. *Updating rule h is represented by admissible subjective model $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ at prior p if, for every signal realization $z \in \mathcal{Z}$, updating via Bayes rule with respect to this model results in the same posterior belief as $h(\cdot,p)$ $\mu_1(\cdot;p)$-almost everywhere:*[9]

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z;p)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z;p)} = h(z,p)_i. \tag{4}$$

2. *Belief forecast $\hat{\rho}$ is represented by admissible subjective model $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ at prior p if the belief forecast induced by this model at p is equal to $\hat{\rho}(\cdot;p)$:*

$$\hat{\mu}\left(\left\{z : \left(\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z(\cdot;p))}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z(\cdot;p))}\right)_{\omega_i\in\Omega} \in X\right\};p\right) = \hat{\rho}(X;p). \tag{5}$$

for every Borel set $X \subset \Delta(\Omega)$.

In characterizing the uniqueness of a representation, we focus on uniqueness with respect to beliefs, and hence, behavior. An updating rule has an essentially unique

[9]Note that, by the mutual absolute continuity assumption, a property that holds $\mu_1(\cdot;p)$-almost everywhere also holds $\mu_i(\cdot;p)$-almost everywhere for all $\omega_i \in \Omega$. As a convention, we use $\mu_1(\cdot;p)$ to keep track of measure zero sets, but the statements can be equivalently stated with respect to $\mu_i(\cdot;p)$ or any admissible $\hat{\mu}_i$ for any $\omega_i \in \Omega$.

representation when any representation is equivalent on the sets of signal realizations that map a given prior to the same posterior.

Definition 4 (Essentially Unique Representation)**.** *An updating rule h has an* essentially unique representation *at prior p if all admissible subjective models representing h at p are equivalent when restricted to the σ-algebra generated by $h(\cdot, p)$, i.e., $\mathcal{F}_h(p) \equiv \{Z \in \mathcal{F} : Z = h^{-1}(X, p)$ for some Borel set $X \subset \Delta(\Omega)\}$, where in a slight abuse of notation h^{-1} is taken with respect to the first argument.*

This notion rules out trivial multiplicities when an updating rule maps multiple signal realizations to the same posterior, as any model that represents this updating rule can also be represented by other models that differ only by shifting mass between these signal realizations. The difference between these models is economically trivial, as they prescribe the same actual and perceived distributions over posteriors and induce the same updating rule. See Appendix C.1 for an illustration of this concept.

2.4 Discussion

Retrospective versus Prospective Bias. A fundamental aspect of behavioral learning models is the distinction between "prospective" and "retrospective" belief formation (see, e.g., Benjamin et al. (2016, 2019)), a distinction that does not arise in rational models. The way a behavioral agent forecasts her beliefs may differ from how she actually forms beliefs. The two components of our non-Bayesian set-up capture this distinction: we formalize retrospective bias in the form of an updating rule and prospective bias in the form of a belief forecast. A misspecified model also allows for such inconsistency with respect to predicted versus actual beliefs. In particular, the distribution an agent expects her future beliefs to be drawn from can differ from the distribution her past beliefs are drawn from. This distinction is similar in spirit to the wedge between an agent's prediction of her future actions and her actual actions in the behavioral literature (e.g., the literatures on time consistency, projection bias, reference dependence, and self-control). See Section 6 for a discussion of time consistency in misspecified models.

Comparison of Approaches. The non-Bayesian updating rule approach is often used to model a specific form of bias or belief-updating error. In general, papers using this approach choose a reasonable parameterization for the bias of interest and study

how this parameterization impacts beliefs and behavior. In contrast, the misspecified model approach is often used in a general, parametric-free way to capture a range of biases within the same framework. For example, recent work in this literature establishes general convergence results for a large class of misspecified models (Bohren and Hauser 2021; Frick et al. 2023; Fudenberg et al. 2021). Establishing a connection between these approaches will makes it straightforward to use general tools from the misspecified learning literature to extend results from the non-Bayesian updating literature. For instance, Section 5.1 uses convergence results from the misspecified model literature to generalize the learning results from Rabin and Schrag (1999) to a larger set of updating rules featuring confirmation bias. This establishes that the qualitative insights of Rabin and Schrag (1999) do not rely on their specific choice of updating rule or information structure (binary signals).

To a large extent, the behavioral literature on biased beliefs—both theoretical and empirical—has focused on updating rules, which are a simple way to define and express biases. But updating rules do not pin down all aspects of belief formation required for economic analysis. Since a misspecified model of belief formation does, mapping updating rules into misspecified models makes it possible to study the implications of a given bias in a richer set of economic environments and clarifies the additional bias—or lack thereof—when doing so.

Dynamics. While we outline our framework for a single signal realization and fixed prior p, it is straightforward to map this set-up into a dynamic environment. Consider a sequence of signals $z_1, z_2, ..., z_T$ (where $T = \infty$ captures an infinite sequence) and let p_t denote the prior belief in period t for each $t = 1, ..., T$. In addition to the assumptions outlined in Section 2.1, assume the sequence of signals are independently drawn, conditional on the state. As in Section 2.1, $\mu_i(\cdot; p)$ denotes the *true signal distribution* conditional on state ω_i at prior p. Fixing p_1, for each $t \geq 1$, the prior in period $t+1$ is equal to the posterior from period t. In the non-Bayesian updating rule approach, this corresponds to $p_{t+1} = h(z_t, p_t)$ and in the misspecified model approach, this corresponds to $p_{t+1,i} = \frac{p_{t,i} \frac{d\hat{\mu}_i}{d\nu}(z_t; p_t)}{\sum_{j=1}^{N} p_{t,j} \frac{d\hat{\mu}_j}{d\nu}(z_t; p_t)}$. The forecast $\hat{\rho}(x; p_t)$ corresponds to the period t forecast of the posterior belief p_{t+1} before observing z_t.

3 Representing Updating Rules and Belief Forecasts

This section derives the main representation result. We establish a necessary and sufficient condition for an updating rule to be represented by a subjective model, and analogously for a belief forecast. In general, these representations are not unique. We then establish necessary and sufficient conditions for an updating rule and belief forecast to be jointly represented by a subjective model and show that this model is essentially unique.

3.1 Representing Updating Rules

An important feature of Bayesian updating is that the expectation of the posterior belief is equal to the prior, i.e., the posterior is a martingale. Therefore, given the set of posterior beliefs that arise under an updating rule, it must be possible to find a subjective model that satisfies this property. We define the relevant set of posteriors that arise from updating rule $h(\cdot, p)$ with respect to the support of the accurate forecast, $\mathcal{X}(h,p) \equiv \operatorname{supp} \rho_h(\cdot; p)$. This is the smallest set of distributions over the state space such that, when at prior p and updating according to h, the posterior is in this set with probability one. An updating rule is *Bayes-feasible* if the prior p lies inside the relative interior of the convex hull of this support, $S(h,p) \equiv \operatorname{rel\,int}(\operatorname{Conv}(\mathcal{X}(h,p)))$.[10]

Definition 5 (Bayes-Feasible Updating Rule)**.** *An updating rule h is* Bayes-feasible at p *if* $p \in S(h,p)$.

It is straightforward to see that prior p must fall within $S(h,p)$ in order for the martingale property to hold. It turns out that this condition is also sufficient for prior p to be the center of mass for *some* distribution over posterior beliefs, which we can then map back into a family of signal distributions, and hence, model.

Lemma 1 (Updating Rule Representation)**.** *There exists an admissible model $\{\hat{\mu}_i(\cdot; p)\}_{\omega_i \in \Omega} \in \mathcal{M}(p)$ that represents updating rule h at prior p if and only if h is Bayes-feasible at p.*

This result extends Lemma 1 from Shmaya and Yariv (2016) to a more general signal space and the class of updating rules we consider.[11] Some care must be taken here,

[10] The relative interior of a set A is the set of points on the interior of A within its affine hull.

[11] In Shmaya and Yariv (2016), the analogue of $S(h,p)$ is the relative interior of the convex hull spanned by posteriors. Our set $S(h,p)$ is the analogue of this set with the additional measurability

both due to the lack of structure on the signal space and the requirements that a misspecified model is absolutely continuous with respect to the true model and has non-zero Radon-Nikodym derivatives. The space of posterior beliefs has more structure than the signal space, which we leverage via $S(h, p)$ for the characterization.

The Bayes-feasibility condition is relatively weak: it holds for many of the non-Bayesian updating rules that have been considered in the literature, including most in Example 1. An example of a violation is an updating rule in which beliefs place more weight on the same state following all signal realizations. It is also violated for certain parameters and signal distributions in updating rules that model base rate neglect, cognitive noise, and partisan bias as parameterized in Example 1.[12] Finally, it is restrictive when the state space is larger than the signal space.

Therefore, this result establishes that many common non-Bayesian updating rules can be represented by a subjective model. This is good news if one would like to use a misspecified model to fill in the gaps left by an 'incomplete' updating rule. However, in general, the representation is not essentially unique: there are often many distinct misspecified models that represent a given updating rule. Each representation induces a different belief forecast. Thus, the choice of representation determines the prospective bias; different choices can lead to different predictions precisely when a belief forecast is needed to close the model. See Appendices C.1 and C.2 for examples of updating rules that satisfy Bayes-feasibility and an illustration of the multiplicity.

3.2 Representing Belief Forecasts

We next develop an analogous result to Lemma 1 for belief forecasts. Again, the property that the posterior belief is equal to the prior in expectation plays a key role. In this case, since the forecast is a distribution over posterior beliefs, the property applies to the forecast directly. This motivates the following definition.

Definition 6 (Plausible Belief Forecast)**.** *A belief forecast $\hat{\rho}$ is* plausible at prior p *if* $\int_{\Delta(\Omega)} x_i d\hat{\rho}(x; p) = p_i$ *for each* $\omega_i \in \Omega$.

In other words, a forecast is *plausible* if the expected posterior, taken with respect to

restrictions necessary for this to be well-defined on an infinite signal space.

[12] The key feature of base rate neglect and cognitive noise that leads to a violation is that the bias manipulates the prior belief. If this manipulated prior is in fact the agent's subjective prior, then these updating rules satisfy Bayes-feasibility with respect to the misspecified prior (see Section 3.4).

the agent's forecast, is equal to the prior. Plausibility ensures that the agent believes that their prior captures all current uncertainty about the state.

Plausibility is a necessary property of Bayesian updating: a Bayesian agent always believes that, on average, her posterior will be equal to her prior. Therefore, in order for the forecast to be represented by a subjective model, it must be plausible—a misspecified agent does not believe that she is systematically biased. We show that this condition is also sufficient for a representation to exist. In other words, for any plausible forecast, it is possible to find a subjective model that induces it.

Lemma 2 (Existence of a Belief Forecast Representation)**.** *There exists an admissible model* $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ *that represents belief forecast* $\hat{\rho}$ *at prior* p *if and only if* $\hat{\rho}$ *is plausible at* p*.*

This is a misspecified analogue of the result in Kamenica and Gentzkow (2011) showing that a distribution over posteriors can be induced by some information structure if and only if it satisfies the martingale property.

Plausibility is relatively strong compared to Bayes-feasibility. Unlike the updating rule, which needs very little structure to be consistent with a misspecified model, a belief forecast must satisfy a strong requirement of Bayesian learning. However, while plausibility rules out many forms of prospective bias for a given prior (e.g., belief forecasts that systematically slant posteriors towards one state under an accurate prior), it still allows for belief forecasts that capture a broad set of prospective biases.

As in the case of updating rules, a belief forecast on its own generally does not have a unique representation. In fact, a continuum of misspecified models can represent a given forecast. Each model induces a different updating rule, and hence, can lead to very different predictions depending on the retrospective bias it encodes. See Appendices C.1 and C.4 for examples of belief forecasts that satisfy plausibility and an illustration of the multiplicity.

3.3 Decomposition

As shown above, an updating rule or belief forecast on its own does not identify a unique misspecified model. This multiplicity gives rise to several important questions. First, given an updating rule, what (if any) restrictions does compatibility with it place on the set of belief forecasts? In other words, does fixing a retrospective bias

restrict the set of feasible prospective biases that a misspecified model can feature, or vice versa? Second, given an updating rule and belief forecast that can be jointly represented, are these two components sufficient to pin down a unique representation, or does a subjective model contain additional restrictions on belief formation? Our next result answers these questions.

A necessary condition for a belief forecast to be compatible with a given updating rule, in that the pair can be jointly represented by a subjective model, is that they have the same support. Recall from Section 3.1 that we define the 'support' of an updating rule with respect to its accurate forecast.

Definition 7 (No unexpected beliefs)**.** *An updating rule h and a belief forecast $\hat{\rho}$ satisfy* no unexpected beliefs *at prior p if $\hat{\rho}(\cdot; p)$ has the same support as the accurate forecast for $h(\cdot, p)$,* $\operatorname{supp} \hat{\rho}(\cdot; p) = \mathcal{X}(h, p)$.

This condition rules out arriving at an entirely unexpected posterior or assigning positive probability to a set of posteriors that will never eventuate given the updating rule. Importantly, this does not rule out the possibility of an incorrect belief forecast: the predicted and actual probabilities of holding a given set of posteriors can differ, and indeed do whenever the agent has prospective bias.

Our main result shows that 'no unexpected beliefs', together with the conditions for an updating rule and the belief forecast to be individually represented—Bayes-feasibility and plausibility—are necessary and sufficient for determining whether they can be jointly represented. Moreover, plausibility and 'no unexpected beliefs' imply Bayes-feasibility—and hence, this latter condition is redundant. We also show that the representation is unique and construct it.

Theorem 1 (Decomposition)**.** *Consider an updating rule h and a belief forecast $\hat{\rho}$. There exists an admissible model $\{\hat{\mu}_i(\cdot; p)\}_{\omega_i \in \Omega} \in \mathcal{M}(p)$ that represents h and $\hat{\rho}$ at prior p if and only if (i) $\hat{\rho}$ is plausible at p and (ii) h and $\hat{\rho}$ satisfy no unexpected beliefs at p. When such a representation exists, it is essentially unique and, for each $\omega_i \in \Omega$, satisfies*

$$\hat{\mu}_i(Z; p) = \frac{1}{p_i} \int_Z h(z, p)_i d\hat{\rho}(h(z, p); p) \tag{6}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h(p)$. This model is misspecified

unless $h(\cdot, p) = h_B(\cdot, p)$ $\mu_1(\cdot; p)$*-almost everywhere and* $\hat{\rho}(\cdot; p) = \rho_B(\cdot; p)$.[13]

It also follows from this result that if an updating rule and a belief forecast are induced by a subjective model, then the belief forecast must be plausible and the pair must satisfy no unexpected beliefs. Thus, not only are plausibility and no unexpected beliefs necessary consequences of the misspecified model approach, they encompass *all* of the belief formation restrictions implicit in using this approach. See Appendix C.1 for an example illustrating how to construct this unique representation.

Discussion. This result has several important theoretical and empirical implications. First, it shows that the updating rule and the belief forecast are the "essential" components of a misspecified model: they fully capture how the model impacts behavior (e.g., how it differs from that of a correctly specified agent). Thus, a misspecified model can be decomposed into the two forms of bias it encodes: the prospective bias through the belief forecast and the retrospective bias through the updating rule. Section 5.2 demonstrates this in a misspecified model of the signal variance.

Second, it establishes that the forms of prospective and retrospective bias encoded in a misspecified model are largely independent from each other: aside from 'no unexpected beliefs', the belief forecast places no further restrictions on which updating rules it can be paired with and vice versa. Thus, many forms of retrospective bias do not place very strong restrictions on the form of prospective bias that a misspecified model can encode. For instance, optimistic updating does not necessarily imply optimistic forecasting. This implies the misspecified model approach can be used to capture the interaction between different (and possibly conflicting) natural biases. A notable exception is updating rules or belief forecasts that simplify the learning environment (e.g., Example 1(f) and Example 2(c)). If one component does so, then to satisfy no unexpected beliefs, the other component must as well. Another exception is whether retrospective bias can be paired with no prospective bias: many updating rules cannot be paired with the accurate forecast for this updating rule, and hence,

[13] In (6), we use $d\hat{\rho}(h(z,p);p)$ for the integral with respect to the measure $\hat{\rho} \circ h(Z,p)$ on $\mathcal{F}_h$. This construction is on the σ-algebra generated by $h(\cdot,p)$, i.e. $\mathcal{F}_h(p)$, since the belief forecast does not place structure on how mass is allocated between signal realizations that induce the same posterior. Lemma 4 in Appendix A constructs one representation on the underlying σ-algebra $\mathcal{F}$. While other constructions are possible, they are all equivalent on $\mathcal{F}_h(p)$, as required for essential uniqueness.

can only be represented by a misspecified model that also encodes some form of prospective bias. We further explore this case in Section 4.1.

Third, the result provides a powerful tool to construct models of biased belief formation. Rather than needing to specify a family of conditional probability distributions—a potentially complicated process that is removed from the conceptual biases of interest—a researcher can simply write down a reasonable parameterization of the desired retrospective and prospective biases and construct a model from these components. Section 5.1 illustrates this construction for the case of confirmation bias; it shows how such a representation can be used to apply the rich asymptotic learning results from the misspecified learning literature (e.g., Bohren and Hauser (2021); Frick et al. (2023)) to updating rules.

Finally, on the empirical side, the updating rule and the belief forecast can both be identified from belief data (see e.g. Benjamin (2019) for updating rules and Chambers and Lambert (2021); Karni (2020) for forecasts). Therefore, the result provides a method to empirically identify a misspecified model via these two components. Relatively simple parameterizations of updating rules or belief forecasts are often used in empirical analysis. To connect the estimates from such analyses with a misspecified model—for instance, to utilize the rich set of theoretical results about misspecified models—one simply needs to ensure that the desired parameterization satisfies the given conditions.

3.4 Misspecified Prior

Recent work on biased learning also allows for a misspecified prior (e.g., Fudenberg, Romanyuk, and Strack (2017)). Our framework easily extends to allow for this. There is a direct analogue of Theorem 1, substituting the misspecified prior for the correct prior. A wider range of prospective biases are possible when the prior is misspecified, as the belief forecast does not need to be correct on average (i.e., average to p). Unlike the case of retrospective and prospective bias, the misspecified model approach does impose a link between prior and prospective bias. For example, optimism bias in the prior (e.g., overweighing the likelihood of the high state) must be accompanied by optimism bias in the forecast (e.g., overweighing the likelihood of posterior beliefs that place high weight on the high state). See Appendix D.1 for the details.

4 Properties of Representations

We next derive classes of models with certain properties. We derive two representations of a biased updating rule with no prospective bias—based on whether an agent is sophisticated or naive about her retrospective bias—and a representation of a biased belief forecast with no retrospective bias. Taken together, these representations serve dual purposes. First, when a researcher starts with an updating rule or belief forecast and wants to use the misspecified model approach for analysis, they highlight which representation to select in order to avoid introducing any additional bias. Second, when a researcher starts with a misspecified model that encodes both retrospective and prospective bias, they provide a way to shut down each bias in turn, thereby isolating the impact of the other and providing insight into their interaction (see the application in Section 5.2 for illustration). Finally, we derive when an updating rule can be represented by the same model at all prior beliefs (a prior-independent representation) and when an updating rule or forecast can be represented by a model with a correct prior (a correct prior representation). The former yields insight into whether the extent and/or direction of bias varies with the prior, while the latter highlights the form of misspecification necessary to generate the given bias.

4.1 Sophisticated-Prospectively Correct Representations

When an agent exhibits bias at a future decision point, a common question is how she anticipates this bias. The two cases typically explored in the literature are that an agent is sophisticated—she accurately anticipates her future bias—or she is naive—she believes she will have no (or less) future bias. An agent who is aware of her retrospective bias has an accurate forecast. Therefore, a *sophisticated-prospectively correct* (sophisticated-PC) model will induce this forecast.

Definition 8 (Sophisticated-PC Model)**.** *Admissible model* $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ *is* sophisticated-PC *at* p *if it induces a biased updating rule,* $h(\cdot,p) \neq h_B(\cdot,p)$ *with* $\mu_1(\cdot;p)$*-positive probability, and the accurate belief forecast* $\rho_h(\cdot;p)$*.*

We are interested in when a given biased updating rule can be represented by a sophisticated-PC model. From Theorem 1, we know that an updating rule and belief forecast can be jointly represented if and only if the forecast is plausible and the forecast and updating rule satisfy no unexpected beliefs. When the forecast is

accurate, no unexpected beliefs is trivially satisfied. Therefore, the necessary and sufficient condition for an updating rule to have a sophisticated-PC representation is that the accurate forecast is plausible. It immediately follows from Theorem 1 that when such a representation exists, it is essentially unique.

Proposition 1 (Sophisticated-PC Representation)**.** *Consider an updating rule h that is biased at prior p. There exists an admissible sophisticated-PC model $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ that represents h at p if and only if the accurate forecast $\rho_h(\cdot;p)$ is plausible at p. When such a representation exists, it is essentially unique and, for each $\omega_i \in \Omega$, satisfies*

$$\hat{\mu}_i(Z;p) = \frac{1}{p_i}\int_Z h(z,p)_i \, d\mu(z;p) \tag{7}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h(p)$.

The example in Appendix C.1 illustrates how to determine whether a sophisticated-PC representation exists and construct it.[14]

The requirement that the accurate forecast is plausible is quite restrictive. Recall that if ρ_h is plausible, it must satisfy $\int_{\Delta\Omega} x_i \, d\rho_h(x;p) = p_i$ for all i. By change of variables, this becomes $\int_{\mathcal{Z}} h(z,p)_i \, d\mu(z;p) = p_i$. So the accurate forecast is plausible only if the biased updating rule averages to the prior under the *true* signal distribution (as opposed to under some *misspecified* signal distribution). This relates to the Bayes-plausibility condition in Kamenica and Gentzkow (2011) which, in our notation, requires plausibility with respect to the Bayesian updating rule, i.e. $\int_{\mathcal{Z}} h_B(z,p)_i \, d\mu(z;p) = p_i$.

Despite this restrictive condition, a sophisticated-PC representation exists for some forms of retrospective bias. Such a representation must preserve the center of mass of beliefs but can otherwise arbitrarily distort the spread of these beliefs. This makes it possible to represent retrospective biases that distort the variance of posterior beliefs, such as the geometric model of under/overreaction from Example 1 (see the example in Appendix C.2). On the other hand, retrospective biases that distort the mean of posterior beliefs, such as partisan bias, can never have such a

[14] In Bohren and Hauser (2024) we discuss how the sophisticated-PC property is essentially equivalent to a much more demanding property—introspection-proofness—which requires the predicted distribution of the signal is equal to the true distribution.

TABLE 1. Properties of Updating Rule Representations

Retrospective Bias	Updating Rule	Soph-PC Rep.	Naive-PC Rep.	Prior-Indep. Rep.
Over/ undereaction	Example 1(a)	Y	N	N
	Example 1(b)	N	Y	Y
Partisan bias	Example 1(c) posterior	N	Y	N
	Example 1(c) signal LR	N	Y	Y
Confirmation bias	Rabin and Schrag (1999)	N	Y	N
	Section 5.1	N	N	N
Cognitive noise	Example 1(d)	N*	N	N
Base-rate neglect	Example 1(e)	N	Y	N
Complexity	Example 1(f)	Y	N	Y

*Except for the case where $p_d = p$.

representation (see the example in Appendix C.3). A sophisticated-PC model also requires a certain amount of complexity in how the updating rule distorts beliefs. This prevents many simple updating rules from having such a representation, such as the updating rule implied by the canonical Grether regressions and commonly used in empirical work (Grether (1980); see Example 1(b)). Table 1 outlines which of the updating rules in Example 1 have a sophisticated-PC representation with a correct prior. A wider range of updating rules can be represented by a sophisticated-PC model with a misspecified prior.

Similar restrictions have been used to pin down prospective beliefs for specific non-Bayesian updating rules. For example, the processing-consistency property in Benjamin et al. (2016) requires an agent to correctly anticipate how she will process information. They define this property with respect to how an agent anticipates versus actually groups multiple signals for processing. In contrast, our condition applies to a single signal (or a fixed grouping of signals): it requires an agent to correctly anticipate her belief distribution after observing this signal. Conceptually similar approaches have also been used in the misspecified model literature to construct plausible restrictions on the space of misspecified models. For example, Spiegler (2016) uses a similar condition to connect a misspecified causal graph—as opposed to an updating rule—to a misspecified model. He imposes this condition on each link

of the graph to pin down a misspecified probability distribution over the outcome of interest. Mailath and Samuelson (2020) study a model of omitted variable bias, where the set of omitted variables together with a sophisticated-PC condition pin down the misspecified model.

4.2 Naive-Prospectively Correct Representations

We next develop the notion of a *naive-prospectively correct* (naive-PC) model to capture settings in which an agent does not have any inherent prospective bias but also does not anticipate her future retrospective bias. The agent naively predicts that she will update beliefs correctly in the future, but when the information arrives, she interprets it with bias. Such an agent has a belief forecast that is accurate with respect to the Bayesian updating rule.

Definition 9 (Naive-PC Model)**.** *Admissible model* $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ *is* naive-PC *at prior* p *if it induces a biased updating rule,* $h(\cdot,p) \neq h_B(\cdot,p)$ *with* $\mu_1(\cdot;p)$*-positive probability and the accurate belief forecast* $\rho_B(\cdot;p)$ *with respect to the Bayesian updating rule.*

Before information arrives, a naive-PC agent makes the same decisions as a correctly specified agent; her misspecification only alters behavior after observing the signal.

Again we are interested in when a given updating rule has a naive-PC representation. Forecast ρ_B is plausible since it is generated by the correctly specified model. Therefore, from Theorem 1, no unexpected beliefs with respect to ρ_B is the necessary and sufficient condition for an updating rule to have a naive-PC representation. It immediately follows from Theorem 1 that when such a representation exists, it is essentially unique and defined by (8).

Proposition 2 (Naive-PC Representation)**.** *Consider an updating rule* h *that is biased at prior* p*. There exists an admissible naive-PC model* $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ *that represents* h *at* p *if and only if* h *and* ρ_B *satisfy no unexpected beliefs at* p*. When such a representation exists, it is essentially unique and, for each* $\omega_i \in \Omega$*, satisfies*

$$\hat{\mu}_i(Z;p) = \frac{1}{p_i}\int_Z h(z,p)_i d\rho_B(h(z,p);p) \tag{8}$$

for any measurable set of signal realizations $Z \in \mathcal{F}_h(p)$.[15]

The example in Appendix C.1 illustrates how to determine whether a naive-PC representation exists and construct it.

The requirement that the updating rule satisfies no unexpected beliefs with respect to ρ_B is not particularly strong. With a sufficiently rich signal space, it holds for many commonly used updating rules, including most in Example 1 (see Table 1). Therefore, in contrast to the sophisticated representation, a naive-PC representation broadly exists for many forms of retrospective bias. As we show in Appendix C.3, a common partisan bias updating rule which did not have a sophisticated-PC representation does have a naive-PC representation.

The naive-PC belief forecast is analogous to common naiveté assumptions made in many behavioral models (e.g., models of time inconsistency (O'Donoghue and Rabin 1999)). It has been used to pin down prospective beliefs in models of biased individual learning (e.g., base rate neglect (Benjamin et al. 2019) and social learning (e.g., partisan bias and overreaction (Bohren and Hauser 2021)). It has also been informally used in less detailed behavioral models (e.g., Benjamin et al. (2016)). Therefore, formalizing how to capture a naive-PC belief forecast in a misspecified model shows that we can consistently and rigorously impose such a property.

Taken together, both the sophisticated- and naive-PC representations of an updatnig rule pin down a belief forecast with respect to the correctly specified model. But the condition for a sophisticated-PC representation to exist is much more restrictive than that for a naive-PC representation. In Section 5.2, we compare how retrospective over- and underprecision impact search behavior in these two representations.

4.3 Retrospectively Correct Representations

We next consider settings in which prospective bias is the primary focus, and derive a representation that shuts down retrospective bias in order to isolate its impact. In a *retrospectively correct* (RC) model, an agent has a biased belief forecast but correctly interprets signals using the Bayesian updating rule.

Definition 10 (RC Model)**.** *Admissible model* $\{\hat{\mu}_i(\cdot;p)\}_{\omega_i\in\Omega} \in \mathcal{M}(p)$ *is* RC *at prior* p *if it induces the Bayesian updating rule* $h_B(\cdot,p)$ *and a biased belief forecast* $\hat{\rho}(\cdot;p) \neq$

[15] Alternatively, one could write this representation as $\hat{\mu}_i(Z;p) = \mu_i(\{z : h_B(z,p) \in h(Z,p)\};p)$.

TABLE 2. Properties of Belief Forecast Representations

Prospective Bias	Belief Forecast	RC Rep.	Correct Prior Rep.
Over/underprecision	Example 2(a)	Y	Y
Partisan Bias	Example 2(b)	Y	N
Complexity	Example 2(c)	N	Y

$\rho_B(\cdot; p)$.

A misspecified agent with a RC model makes the same decisions as a correctly specified agent after information arrives, but can behave differently ex-ante. When a belief forecast has a RC representation immediately follows from Theorem 1.

Proposition 3. *Consider a belief forecast $\hat{\rho}$ that is biased at prior p. There exists an admissible RC model $\{\hat{\mu}_i(\cdot; p)\}_{\omega_i \in \Omega} \in \mathcal{M}(p)$ that represents $\hat{\rho}$ at p if and only if $\hat{\rho}$ is plausible at p and h_B and $\hat{\rho}$ satisfy no unexpected beliefs at p. When such a representation exists, it is essentially unique and satisfies (6) setting $h(\cdot, p) = h_B(\cdot, p)$.*

This establishes that many belief forecasts are consistent with Bayesian updating. An agent can have very biased predictions about her future beliefs, but still update correctly after observing the signal. Therefore, the misspecified model approach can be used to capture prospective biases without needing to also allow for retrospective bias. As shown in Table 2, two of the three belief forecasts in Example 2 have a RC representation. The example in Appendix C.1 illustrates how to determine whether a RC representation exists and construct it.

4.4 Other Properties

Prior-Independent Representations. In Theorem 1, the representation can vary with the prior belief. This is natural when the true model varies with the prior, but it can also occur when the true model does not. Therefore, another property that sheds light on the structure of bias is whether an updating rule can be represented by the same model at all prior beliefs—that is, it has a *prior-independent representation*. In Appendix D.4 we derive a necessary and sufficient condition for an updating rule to have such a representation. Many well-known parameterizations of common biases have prior-independent representations. As shown in Table 1, this includes many of

the updating rules in Example 1.

Correct Prior Representations. Some biases can only be represented by a misspecified model with an incorrect prior. Therefore, whether an updating rule or forecast can be represented by a model with a correct prior also provides insight into the structure of the bias. This is more relevant for belief forecasts, since, through plausibility, the requirement of a correct prior places more structure on the prospective bias than on the retrospective bias. Table 2 highlights whether the forecasts in Example 2 have such a *correct prior representation*.

5 Applications

The following two applications demonstrate the results from Sections 3 and 4. The first shows how a misspecified model representation can be used to study learning when an agent updates with confirmation bias. The second shows how the decomposition clarifies the impact that a misspecified model has on search behavior.

5.1 Representing Confirmation Bias

This application explores confirmation bias in belief-updating, in that an agent misinterprets information to confirm her current belief. Rabin and Schrag (1999) show that a particular updating rule exhibiting confirmation bias leads to *incorrect learning* (i.e., with positive probability the agent places probability one on the incorrect state). We use the representation in Theorem 1 to show that, in an individual learning setting, incorrect learning continues to emerge for a broad class of updating rules that exhibit confirmation bias, regardless of the belief forecast. However, in a social learning setting, the asymptotic learning outcomes depend crucially on the chosen belief forecast—different representations lead to different learning predictions.

Individual Learning. In Rabin and Schrag (1999), an agent observes an i.i.d. sequence of binary signals $z_t \in \{l, r\}$ for $t = 1, 2, ...$, where each signal matches the unknown state $\omega \in \{L, R\}$ with probability $\theta > 1/2$. If the signal contradicts her current belief (i.e., she believes state L is more likely and observes signal r, or vice versa), then with probability $q \in (0, 1)$ the agent misinterprets the signal and updates to the Bayesian posterior following the opposite signal realization. Letting p denote the current belief that the state is R, m denote a misinterpreted signal realization and, in a slight abuse of notation, l and r denote correctly interpreted signal realizations,

this corresponds to $h(z,p) = h_B(r,p)$ when $z = m$ and $p > 1/2$, $h(z,p) = h_B(l,p)$ when $z = m$ and $p \leq 1/2$, and $h(z,p) = h_B(z,p)$ otherwise. They show that for high q, incorrect learning occurs with positive probability. This parameterization makes several strong assumptions. First, misinterpreted signals are interpreted as the opposite signal. Second, the severity of the confirmation bias—the frequency q that a signal is misinterpreted and the degree to which it is slanted—is independent of the current belief: an agent exhibits the same bias regardless of whether she believes one state is a lot or a little more likely than the other.

We use our framework to show that similar learning outcomes obtain for a broad class of updating rules with confirmation bias, where we allow the probability of misinterpreting the signal and the slant of the misinterpreted signal to vary with the current belief. Specifically, if the signal contradicts the agent's current belief p, then with probability $q(p) \in [0,1]$ she misinterprets it and slants it by weight $\nu(p) \in [0,1]$ towards the Bayesian posterior for the opposite realization. Again letting m denote a misinterpreted signal realization, the following updating rules captures these features:

$$h(z,p) = \begin{cases} (1-\nu(p))h_B(l,p) + \nu(p)h_B(r,p) & z = m \text{ and } p > 1/2 \\ \nu(p)h_B(l,p) + (1-\nu(p))h_B(r,p) & z = m \text{ and } p \leq 1/2 \\ h_B(z,p) & z \in \{l,r\}. \end{cases} \tag{9}$$

For technical reasons, we assume that ν and q are continuous and symmetric at certainty, $\nu(0) = \nu(1)$ and $q(0) = q(1)$. The model of Rabin and Schrag (1999) corresponds to the case where $q(p)$ is constant and $\nu(p) = 1$.

Theorem 1 makes it straightforward to represent this more general updating rule as a misspecified model and apply the criterion for characterizing asymptotic learning from Bohren and Hauser (2021). This updating rule satisfies Bayes-feasibility, and therefore, by Lemma 1, has a misspecified model representation. Moreover, any misspecified model representation satisfies the assumptions in Bohren and Hauser (2021), so we can use their criterion to characterize asymptotic learning outcomes. Finally, all representations reduce to the same criterion, and therefore, lead to the same learning outcomes; the belief forecast does not impact asymptotic learning. This

leads to the following result.[16]

Proposition 4. *There exist cut-offs $\overline{q} \in (0,1)$ and $\overline{\nu} : [0,1] \to (0,1]$, with $\overline{\nu}(q(1)) < 1$ for $q(1) > \overline{q}$, such that (i) for $q(1) > \overline{q}$ and $\nu(1) > \overline{\nu}(q(1))$, both incorrect and correct learning arise with positive probability, and with probability one, one of these two outcomes arise; (ii) for $q(1) < \overline{q}$ or $\nu(1) < \overline{\nu}(q(1))$, almost surely learning is correct.*

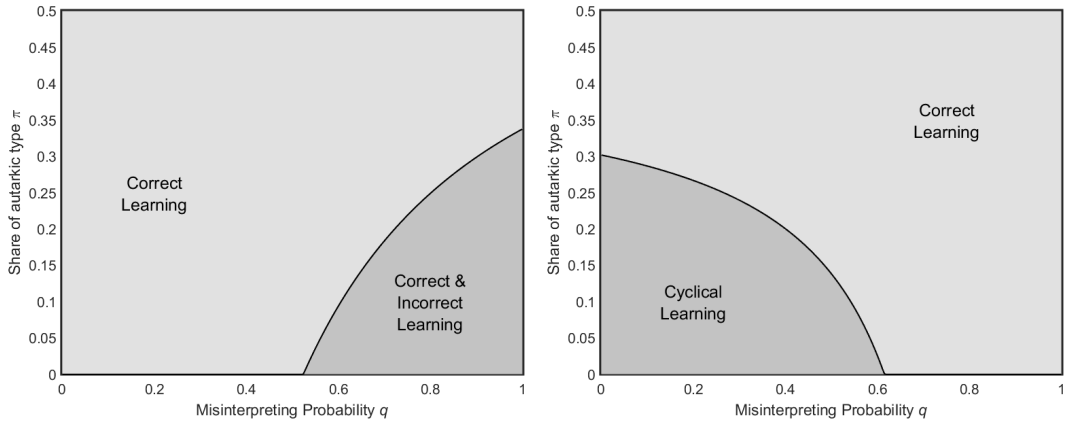
This establishes that the stark parameterization of confirmation bias in Rabin and Schrag (1999) does not drive the possibility of incorrect learning: when confirmation bias is sufficiently severe, beliefs become entrenched on the incorrect state regardless of the exact parameterization. See Appendix B for the analysis.

Social Learning. We next show that in a social learning environment where agents use the confirmation bias updating rule specified above, the choice of belief forecast *does* impact the asymptotic learning characterization. Suppose that a sequence of agents learn from a private signal and the action choices of prior agents. In particular, each agent t observes a private signal $z_t \in \{l, m, r\}$ as outlined above, then chooses an action $a_t \in \{L, M, R\}$, where payoffs are such that action R is optimal following $p > \overline{p} \in (1/2, 1)$, action L is optimal following $p < 1 - \overline{p}$, and action M is optimal for $p \in (1 - \overline{p}, \overline{p})$. Suppose that a share $1 - 2\pi$ of agents observe the action history $h_t = \{a_1, ..., a_{t-1}\}$, while share 2π are autarkic and only observe their private signal, with share π having a prior belief $p_A \in (1/2, \overline{p})$ and share π having a prior belief $1 - p_A < 1/2$. Both autarkic and non-autarkic types use the updating rule described above, but are not aware that anyone misinterprets signals. In particular, non-autarkic types correctly infer autartkic types' signals, but apply the same updating rule to this signal as they apply to their own signal. Assume that non-autarkic agents have a correct belief about the share of autarkic types.

This updating rule has multiple misspecified model representations. Social learning requires inference about sets of signals that map to a given action, which depends on both the belief forecast and the updating rule. Hence, the choice of representation affects the predicted learning outcomes.[17] To illustrate our point, we focus on

[16] We present this result for symmetric signals as in Rabin and Schrag (1999); it immediately extends to asymmetric signals.

[17] Note that this updating rule has a naive-PC representation only when $\nu(p) \in \{0, 1\}$, as otherwise the posterior following signal m does not arise in the correctly specified model, and does not have a



(a) Rep. #1: $\hat{\rho}(p) = 0.01$ (b) Rep. #2: $\hat{\rho}(p) = 0.5$

FIGURE 1. Social learning outcomes depend on the chosen representation ($\theta = 0.65$)

a parameterization of the updating rule in which agents interpret signal m as uninformative: $h(m, p) = p$. Fig. 1 illustrates how the learning outcomes depend on the probability of misinterpreting a signal q and the share of autarkic types π for two representations. The representation in panel (a) has a belief forecast that places very low probability on uninformative information, $\hat{\rho}(p) = 0.01$. The learning outcome is similar in spirit to the individual learning setting: incorrect learning arises when the signal is misinterpreted with sufficiently high probability. The representation in Panel (b) has a belief forecast that places higher probability on uninformative information—$\hat{\rho}(p) = 0.5$. This leads to starkly different learning outcomes: *cyclical learning* (i.e., beliefs fail to converge) arises for a sufficiently low share of autarkic types and probability of misinterpreting the signal.

5.2 Misspecified Search

This application explores how a misspecified model of the signal variance impacts search decisions. Following Theorem 1, we decompose the misspecified model into the prospective and retrospective biases it encodes, then use Propositions 1 to 3 to separately study the impact of each on search behavior. Whether the bias encoded in the misspecified model emerges ex-ante versus ex-post to information arrival plays a key role in determining whether excess or insufficient search occurs.

sophisticated-PC representation for any $q(p) > 0$.

Set-Up. A firm considers whether to adopt one of two new technologies, $j \in \{1, 2\}$. Technology j has either low or high value, $\omega_j \in \{L, H\}$, each equally likely. Values are independently drawn and unobserved. The firm learns about these technologies sequentially. In each of two periods, it chooses whether to search a new technology (if an unsearched option remains) or to adopt one of the technologies it has already searched. Without loss of generality, assume that technology 1 is searched first. When the firm searches technology j, it draws a signal z_j from normal distribution $N(1, 1)$ when $\omega_j = H$ and $N(-1, 1)$ when $\omega_j = L$. The signals are independent across technologies. The firm has a misspecified model of the signal process: it believes the signal is drawn from normal distribution $N(1, \hat{\sigma}^2)$ when $\omega_j = H$ and $N(-1, \hat{\sigma}^2)$ when $\omega_j = L$, with $\hat{\sigma} \neq 1$. In other words, it correctly perceives the mean but misperceives the variance: it overestimates it when $\hat{\sigma} > 1$ and underestimates it when $\hat{\sigma} < 1$. It correctly believes that the signals are independent across technologies. The firm receives a payoff of 1 from adopting a high value technology and 0 from adopting a low value technology or not adopting any technology. It costs the firm $c = 0.1$ to search each technology.

Decomposition. Decomposing the misspecified model into the induced updating rule and belief forecast isolates the retrospective and prospective biases it encodes. Letting $\phi(z|m, \sigma)$ denote the pdf of the normal distribution with mean m and variance σ^2, the misspecified model induces updating rule $h(z) = \frac{\phi(z|1,\hat{\sigma})}{\phi(z|1,\hat{\sigma})+\phi(z|-1,\hat{\sigma})}$, where $h(z)$ is the subjective probability that the technology is high value after observing realization z and we suppress the dependence on the prior since it is fixed. This is the geometric parameterization of under/overreaction in Example 1(b) with $\alpha = 1/\hat{\sigma}^2$. Thus, if the agent overestimates the variance, $\hat{\sigma}^2 > 1$, she retrospectively underreacts and if she underestimates the variance, $\hat{\sigma}^2 < 1$, she retrospectively overreacts. The misspecified model induces belief forecast $\hat{\rho}(x) = \hat{\mu}(h^{-1}(x))$ (in cdf form), where $\hat{\rho}$ is the subjective distribution over the posterior belief that the technology is high value, $\hat{\mu}$ denotes the cdf of an equally weighted mixture distribution of $N(1, \hat{\sigma}^2)$ and $N(-1, \hat{\sigma}^2)$, and again we suppress the dependence on the prior.[18] When the agent overestimates the variance, $\hat{\sigma}^2 > 1$, this forecast puts insufficient weight on precise posteriors and

[18]Since $h(z)$ is increasing in z, we can write cdf $\hat{\rho}$ in terms of cdf $\hat{\mu}$, i.e., $\hat{\rho}(x) \equiv Pr(\{z : h(z) \leq x\}) = \hat{\mu}(h^{-1}(x))$.

excess weight on intermediate posteriors, leading to prospective underprecision as in Example 2(a), and analogously prospective overprecision for $\hat{\sigma}^2 < 1$. Thus, this misspecified model puts structure on the relationship between the prospective and retrospective bias: retrospective overreaction is coupled with prospective overprecision, and similarly for underreaction and underprecision. Fig. 4 in Appendix B plots this updating rule and forecast, as well as the correctly specified analogues $h_B(z)$ and $\rho_B(x) = \mu(h_B^{-1}(x))$ and the accurate forecast $\rho_h(x) = \mu(h^{-1}(x))$.

Search Behavior. The firm always searches the first technology since $c < 0.5$. After observing signal realization z_1, it searches the second technology if

$$c < \int_{h(z_1)}^{1} (x - h(z_1))\, d\hat{\rho}(x), \tag{10}$$

where we assume the firm does not search when indifferent. Since the right hand side of (10) is decreasing in $h(z_1)$, and $h(z_1)$ is increasing in z_1, we can express this decision as a cut-off z^* on the signal such that the firm searches the second technology iff $z_1 < z^*$. If it does not search the second technology, the firm adopts the first technology, and if it does, it adopts the technology with the higher signal.

As can be seen in (10), both retrospective bias and prospective bias impact search behavior. When the firm has an overprecise forecast, it overestimates the likelihood of receiving a very precise signal when it searches the second technology, which leads it to overestimate the benefit of search and thereby search the second technology too often. But in this case, the firm also overreacts to the first signal: following a high first signal, it overestimates the probability that the first technology is good—counteracting the forecast bias—and following a low first signal, it underestimates the probability that the first technology is good—exacerbating the forecast bias. Thus, the interaction between the prospective and retrospective bias will determine whether the firm searches too often or not often enough, compared to a correctly specified firm.

Fig. 2 shows that the firm engages in insufficient search, regardless of whether the misperceived variance induces overprecision and overreaction ($\hat{\sigma} < 1$) or underprecision and underreaction ($\hat{\sigma} > 1$). Relative to the correctly specified model, the firm sets a lower signal threshold to adopt the first technology without searching the second. The extent of this deviation from optimal search is increasing in the difference between the true and misperceived variance: a larger bias in either direction leads to a

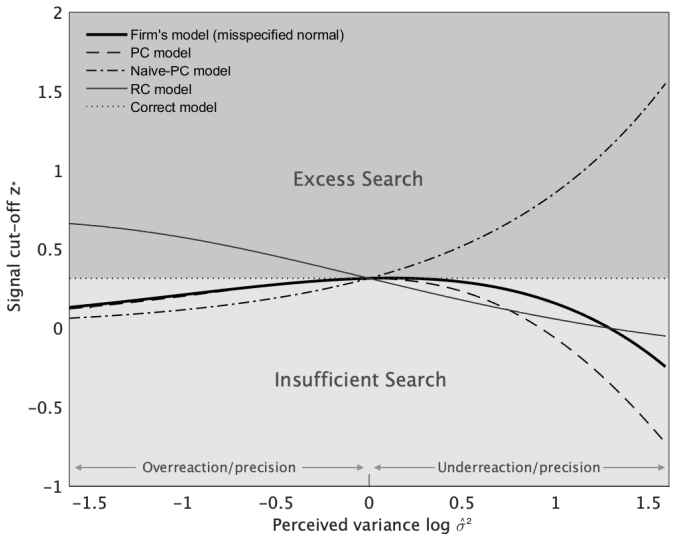


FIGURE 2. Impact of retrospective versus prospective bias on search behavior

higher probability of inefficiently stopping. We next study the role of the prospective versus retrospective bias in driving this inefficiency; the results from Section 4 allow us to separately shut down each bias and isolate the impact of the other.

The Impact of Retrospective Bias. We first examine how retrospective over- or underreaction impact search behavior by comparing the firm's behavior to the RC model. This model induces the Bayesian updating rule h_B and the same forecast $\hat{\rho}$ as the firm's model.[19] First consider the case where the firm underestimates the variance, $\hat{\sigma} < 1$, thereby exhibiting retrospective overreaction and prospective overprecision. Fig. 2 shows that the RC model leads to excess search, while the firm's model leads to insufficient search. Thus retrospective overreaction drives the insufficient search—absent such overreaction, prospective overprecision generates excess search. In contrast, when the firm overestimates the variance, $\hat{\sigma} > 1$ (retrospective underreaction and prospective underprecision), both the RC model and the firm's model lead to insufficient search. For small levels of bias, the RC model is more inefficient than the firm's model; thus, retrospective underreaction mitigates the efficiency loss stemming from prospective underprecision. For more severe bias, the

[19]Given that h_B and $\hat{\rho}$ satisfy no unexpected beliefs (h_B induces all posteriors in $(0,1)$, which is the support of $\hat{\rho}$) and $\hat{\rho}$ is plausible (since it is induced by a misspecified model), by Proposition 3, the RC model exists and is unique. See Appendix B for details.

firm's model is more inefficient; thus, retrospective underreaction exacerbates the efficiency loss stemming from prospective underprecision. Taken together, this shows that retrospective and prospective bias can either offset or amplify each other, depending on the direction and severity of the bias. In either case, prospective bias alone also leads to inefficient behavior, demonstrating the importance of considering the impact of prospective bias in addition to the more oft studied retrospective biases.

The Impact of Prospective Bias. We next examine how prospective bias impacts search behavior by comparing the firm's behavior to that in the two prospectively correct representations. Both induce the same updating rule h as the firm's model; the naive-PC representation induces the forecast from the correctly specified model, ρ_B, while the sophisticated-PC representation induces the accurate forecast, ρ_h.[20] Fig. 2 plots search behavior for each. When the firm underestimates the variance, search behavior in both prospectively correct models is similar to the firm's model. Intuitively, for small $\hat{\sigma}$, the firm maps any positive signal to an almost-one probability of high quality and any negative signal to an almost-zero probability of high quality. Therefore, the signal cut-off z^* approaches zero independently of the belief forecast, leading to insufficient search. In fact, overreaction on its own—combined with the failure to anticipate this overreaction in the naive-PC model—leads to even less search than the firm's model. Thus, the prospective overprecision in the firm's model offsets the inefficiently low search generated by retrospective overreaction. Since search behavior in the firm's model and the sophisticated-PC model are near identical, making the firm aware of its retrospective overreaction would not mitigate its insufficient search. When the firm overestimates the variance, a very high signal is needed to move the posterior away from the prior. In the naive-PC representation, the firm does not anticipate this and searches too often. In the sophisticated-PC representation, the firm anticipates this and searches too little. The same is true for the firm's model, but less so because it places lower probability on low signals when the state is low, and therefore, ascribes a higher value to search.

[20] Given that h and ρ_B satisfy no unexpected beliefs (h induces all posteriors in $(0,1)$, which is the support of ρ_B), by Proposition 2 the naive-PC representation exists and is unique. Given that ρ_h is plausible, by Proposition 1 the sophisticated-PC representation exists and is unique. See Appendix B for details.

Discussion. Taken together, these results show that whether bias emerges prospectively versus retrospectively leads to qualitatively different predictions about search decisions. Therefore, the timing of when a bias emerges has important implications for behavior. Moreover, the prospective and retrospective bias a model encodes can either offset or amplify each other. This has important implications for selecting policy interventions, including which bias is more important to target and whether targeting one bias but not the other will lead to further inefficiency.

These results also show how the choice of approach impacts the analysis. An economist who uses the non-Bayesian updating rule approach may view overreaction paired with the naive-PC forecast as the natural model to use, while an economist who uses the misspecified model approach may view the misspecified variance model as the natural set-up. Since these models lead to qualitatively different behavior, the choice of approach will impact the conclusions the economist reaches.

6 Discussion and Conclusion

Time Consistency. Time inconsistency is a key property of many dynamic behavioral models. In terms of belief distortions, it is an inherent feature of certain biases (e.g. confirmation bias or disbelief in the law of large numbers (Benjamin et al. 2016)). Therefore, any representation of such biases will exhibit time inconsistency in that the models an agent anticipates she will use in future periods differ from the models she actually uses. This notion of bias in anticipated versus actual model is conceptually distinct from our notion of prospective versus retrospective bias in anticipated versus actual processing of a signal within a given model. A straightforward extension of our framework that combines a prior-dependent representation with a failure to anticipate how the model changes with the current belief can capture time inconsistency (see Appendix D.2). Prior-dependent models do not always lead to time inconsistency. When the agent accurately anticipates how her model changes with her belief, she will be time consistent. For example, a correctly specified model that varies with the prior—as in active learning environments—is prior-dependent but clearly also time consistent.

Conclusion. We link two approaches commonly used to study biases in belief formation: the non-Bayesian approach and the misspecified model approach. Our main result decomposes a misspecified model into the two forms of bias it encodes—

retrospective bias via the updating rule and prospective bias via the belief forecast—and highlights the belief formation restrictions implicit in using the misspecified model approach. Moreover, it demonstrates how to uniquely represent an updating rule or belief forecast by suitably selecting the other component. Finally, we identify natural ways to construct such models that do not introduce additional bias. Taken together, these results provide a method to embed belief formation biases into economic decision problems. They also highlight the importance of eliciting the belief forecast as well as the (more commonly measured) updating rule in empirical work, as both components of belief formation play a key role in many economic settings.

References

Arrow, K. J. and J. R. Green (1973): "Notes on Expectations Equilibria in Bayesian Settings," *Institute for Mathematical Studies in the Social Sciences Working Papers*.

Ba, C. (2024): "Robust Model Misspecification and Paradigm Shifts," *PIER Working Paper*.

Ba, C., A. Bohren, and A. Imas (2024): "Over- and Underreaction to Information: the Role of Complexity in Belief-Updating," *Working paper*.

Benjamin, D., A. Bodoh-Creed, and M. Rabin (2019): "Base-rate neglect: Foundations and implications," .

Benjamin, D. J. (2019): "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2, 69–186.

Benjamin, D. J., M. Rabin, and C. Raymond (2016): "A Model of Nonbelief in the Law of Large Numbers," *Journal of the European Economic Association*, 14, 515–544.

Bohren, J. A. and D. N. Hauser (2021): "Learning with heterogeneous misspecified models: Characterization and robustness," *Econometrica*, 89, 3025–3077.

——— (2023): "Optimal Lending Contracts with Retrospective and Prospective Bias," *AEA Papers and Proceedings*.

——— (2024): "Introspection-Proof Misspecified Models," .

Chambers, C. P. and N. S. Lambert (2021): "Dynamic belief elicitation," *Econometrica*, 89, 375–414.

Chauvin, K. P. (2020): "Euclidean properties of bayesian updating," .

Cripps, M. W. (2018): "Divisible Updating," .

de Clippel, G. and X. Zhang (2022): "Non-bayesian persuasion," *Journal of Political Economy*, 130, 2594–2642.

Epstein, L. G., J. Noor, and A. Sandroni (2010): "Non-Bayesian Learning," *The B.E. Journal of Theoretical Economics*, 10.

Espitia, A. (2021): "Confidence and Organizations," .

Esponda, I. (2008): "Behavioral equilibrium in economies with adverse selection," *American Economic Review*, 98, 1269–91.

Esponda, I. and D. Pouzo (2016): "Berk–Nash equilibrium: A framework for modeling agents with misspecified models," *Econometrica*, 84, 1093–1130.

Esponda, I., D. Pouzo, and Y. Yamamoto (2021): "Asymptotic behavior of Bayesian learners with misspecified models," *Journal of Economic Theory*, 195, 105260.

Eyster, E. and M. Rabin (2005): "Cursed Equilibrium," *Econometrica*, 73, 1623–1672.

——— (2010): "Naive Herding in Rich-Information Settings," *American Economic Journal: Microeconomics*, 2, 221–243.

Frick, M., R. Iijima, and Y. Ishii (2020): "Misinterpreting Others and the Fragility of Social Learning," *Econometrica*, 88, 2281–2328.

——— (2023): "Belief Convergence under Misspecified Learning: A Martingale Approach," *Review of Economic Studies*, 90, 781–814.

——— (2024): "Welfare Comparisons for Biased Learning," *American Economic Review*, 114, 1612–1649.

Fudenberg, D. and G. Lanzani (2023): "Which misspecifications persist?" *Theoretical Economics*, 18, 1271–1315.

Fudenberg, D., G. Lanzani, and P. Strack (2021): "Limit points of endogenous misspecified learning," *Econometrica*, 89, 1065–1098.

Fudenberg, D., G. Romanyuk, and P. Strack (2017): "Active learning with a misspecified prior," *Theoretical Economics*, 12, 1155–1189.

Grether, D. M. (1980): "Bayes rule as a descriptive model: The representativeness heuristic," *The Quarterly journal of economics*, 95, 537–557.

He, K. (2022): "Mislearning from censored data: The gambler's fallacy and other

correlational mistakes in optimal-stopping problems," *Theoretical Economics*, 17, 1269–1312.

He, K. and J. Libgober (2024): "Evolutionarily stable (mis) specifications: Theory and applications," *PIER Working Paper*.

He, X. D. and D. Xiao (2017): "Processing consistency in non-Bayesian inference," *Journal of Mathematical Economics*, 70, 90–104.

Heidhues, P., B. Koszegi, and P. Strack (2018): "Unrealistic Expectations and Misguided Learning," *Econometrica*, 86, 1159–1214.

Jakobsen, A. M. (2023): "Coarse bayesian updating," .

Jehiel, P. (2005): "Analogy-based expectation equilibrium," *Journal of Economic Theory*, 123, 81–104.

Kamenica, E. and M. Gentzkow (2011): "Bayesian persuasion." *American Economic Review*, 2590–2615.

Karni, E. (2020): "A mechanism for the elicitation of second-order belief and subjective information structure," *Economic Theory*, 69, 217–232.

Lanzani, G. (2024): "Dynamic concern for misspecification," .

Le Yaouanq, Y. and P. Schwardmann (2022): "Learning About One's Self," *Journal of the European Economic Association*, 20, 1791–1828.

Levy, G., R. Razin, and A. Young (2022): "Misspecified Politics and the Recurrence of Populism," *American Economic Review*, 112, 928–62.

Libgober, J. (2024): "Identifying Wisdom of the Crowd: A Regression Approach," .

Mailath, G. J. and L. Samuelson (2020): "Learning under diverse world views: Model-based inference," *American Economic Review*, 110, 1464–1501.

Molavi, P. (2024): "The Empirical Content of Bayesianism," .

Mullainathan, S. (2002): "Thinking through categories," .

Nyarko, Y. (1991): "Learning in Misspecified Models and the Possibility of Cycles," *Journal of Economic Theory*, 55, 416–427.

O'Donoghue, T. and M. Rabin (1999): "Doing It Now or Later," *American Economic Review*, 89, 103–124.

Prelec, D. and J. McCoy (2022): "General identifiability of possible world models for crowd wisdom," .

Rabin, M. and J. L. Schrag (1999): "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 114, 37–82.

Schwartzstein, J. (2014): "Selective Attention and Learning," *Journal of the European Economic Association*, 12, 1423–1452.

Shmaya, E. and L. Yariv (2016): "Experiments on decisions under uncertainty: A theoretical framework," *American Economic Review*, 106, 1775–1801.

Spiegler, R. (2016): "Bayesian networks and boundedly rational expectations," *Quarterly Journal of Economics*, 131, 1243–1290.

——— (2020): "Behavioral implications of causal misperceptions," *Annual Review of Economics*, 12, 81–106.

Thaler, M. (2021): "Gender differences in motivated reasoning," *Journal of Economic Behavior & Organization*, 191, 501–518.

Woodford, M. (2020): "Modeling imprecision in perception, valuation, and choice," *Annual Review of Economics*, 12, 579–601.

A Proofs from Sections 3 and 4

In this section, except where noted we omit the prior p from the arguments of the forecast, updating rule, and misspecified model. As most results hold prior by prior, we establish them for an arbitrary interior prior $p \in \Delta(\Omega)$.

Proof of Lemma 1. (If) Let $\mathcal{M}_{uc}$ denote the set of signal distributions that are mutually absolutely continuous with respect to μ. Let $F \equiv \{x : x_i = \int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z),\ \hat{\mu} \in \mathcal{M}_{uc})\}$ and $\overline{S}(h)$ denote the closure of $S(h)$. We first show that $\overline{F} = \overline{S}(h)$, which implies that $S(h) = \text{rel int } F$ since both sets are convex, and then show that any prior that lies in the relative interior of F can be represented by a misspecified model. Consider any $x \in \overline{S}(h)$. Since the closure of the support of ρ_h, $\overline{\mathcal{X}}(h)$, is compact, $\overline{S}(h)$ is its convex hull. By Caratheodory's theorem there is a set of $K \leq N$ $a_i \in \overline{\mathcal{X}}(h)$ s.t. $\sum_{j=1}^{K} \lambda_j a_j = x$, $\lambda_j > 0$, $\sum_{j=1}^{K} \lambda_j = 1$.

It remains to establish that x lies in F, so we need to construct a $\hat{\mu}$, mutually absolutely continuous with respect to μ that satisfies $\int h(z)_i d\hat{\mu}(z) = x_i$. Fix $\varepsilon \in (0, \min_j\{\lambda_j\})$, and for each a_j take a collection of disjoint balls of radius $\delta < \frac{\varepsilon}{2K}$ around a_j, $B_\delta(a_j)$. The set of signals that map to this ball has positive measure.

Define a density by

$$\frac{d\hat{\mu}}{d\mu}(z) = \begin{cases} \frac{\lambda_j - \frac{\varepsilon}{2K}}{\mu(h^{-1}(B_\delta(a_i)))} \text{ if } z \in h^{-1}(B_\delta(a_i)) \\ \frac{\varepsilon}{2\mu(\mathcal{Z} \setminus h^{-1}(\bigcup_{j=1}^K B_\delta(a_j)))} \text{ o.w.} \end{cases}$$

if $\mu(\mathcal{Z} \setminus h^{-1}(\bigcup_{j=1}^K B_\delta(a_j))) > 0$, otherwise let $\frac{d\hat{\mu}}{d\mu}(z) = \frac{\lambda_j}{\mu(h^{-1}(B_\delta(a_i)))}$ if $z \in h^{-1}(B_\delta(a_i))$. This is by construction a non-negative measurable function. With respect to this density, $|\int_{\mathcal{Z}} h(z)_i d\hat{\mu}(z) - x_i| \leq \varepsilon$, so $x \in \overline{F}$. By standard argument any point in F is in the closure of $S(h)$, so these two sets are the same and we can work directly with points in F.

Now we show that h can be represented. Consider the vector $m \in \Delta(\Omega)$ where $m_i = \int_{\mathcal{Z}} h(z)_i \, d\mu(z)$, the expected value of the misspecified posterior under the true unconditional distribution, which exists, and lies in F. Since the prior p is in the relative interior, there exists an $\varepsilon > 0$ s.t. $q = (1+\varepsilon)p - \varepsilon m \in F$. Moreover, there exists a probability distribution $\gamma \in \mathcal{M}_{uc}$ absolutely continuous with respect to ν s.t. $q_i = \int_{\mathcal{Z}} h(z)_i \, d\gamma(z)$. Consider the compound lottery where with probability $\frac{1}{1+\varepsilon}$ the signal z is drawn from γ and with complementary probability it is drawn from μ. Call this measure $\hat{\mu}$. Then $\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z) = p_i$. Finally, suppose that there was a set Z with ν-positive measure where for all $z \in Z$, $\frac{d\mu_i}{d\nu}(z) > 0$ but $\frac{d\hat{\mu}_i}{d\nu}(z) = 0$. This set occurred with positive probability under μ so it must occur with positive probability under $\hat{\mu}$ by construction. This is a contradiction. Therefore, we can represent this with a misspecified model.

(Only If) Take a measure $\hat{\mu} \in \mathcal{M}_{uc}$. This induces a full support distribution over $\operatorname{supp} \rho_h$, denoted $\hat{\rho}_{\hat{\mu}} \equiv \hat{\mu} \circ h^{-1}$. Let $m_i = \int_{\mathcal{Z}} h(z)_i d\hat{\mu}(z)$. Suppose m was not on the relative interior. Then there exists a hyperplane that properly supports $S(h)$ at m, $v \in \mathbb{R}^N$ s.t. $v \cdot m \geq v \cdot s$ for all $s \in S(h)$, strict for any s on the relative interior. But then, since the relative interior is non-empty, any point on the relative interior can be written as the convex combination of points in the support (implying at least one of these points is not on the hyperplane), and any neighborhood of that point occurs with positive probability, $v \cdot m = \int v \cdot s \, d\hat{\rho}_{\hat{\mu}}(s) < v \cdot m$ by the full support assumption. This is a contradiction. □

Proof of Lemma 2. (If) Fix a plausible forecast $\hat{\rho}$ and the associated measurable function $\phi : \mathcal{Z} \to \Delta(\Omega)$ such that $\mu_1(\phi^{-1}(\cdot))$ and $\hat{\rho}$ are mutually absolutely continuous

(by Definition 2 such a function exists). Let $\rho_\phi = \mu \circ \phi^{-1}$. Define the measure $\hat{\mu}(Z) = \int_Z \frac{d\hat{\rho}}{d\rho_\phi}(\phi(z))\, d\mu(z)$. Note that $\int_{\mathcal{Z}} \phi(z)_i d\hat{\mu}(z) = \int_{\Delta(\Omega)} x_i d\hat{\rho}(x) = p_i$ so $\hat{\mu}_i(Z) = \frac{1}{p_i}\int_Z \phi(z)_i\, d\hat{\mu}(z)$ is a subjective model with unconditional signal distribution $\hat{\mu}$. This subjective model has forecast $\hat{\rho}$ by construction of $\hat{\mu}$ and the change of variables formula. (Only If) Fix a subjective model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$. Let $h(z)$ be the updating rule defined by Bayes rule with respect to this model. Then if $\hat{\rho}(X) = \hat{\mu}(h^{-1}(X))$ is a forecast, it is, by definition, the forecast represented by the subjective model. By construction, $h(z)$ is a measurable function s.t. $\hat{\rho}(X) = 0$ if and only if $\rho_h(X) = 0$. So $\hat{\rho}$ is a forecast. Finally,

$$\int_{\Delta(\Omega)} x_i d\hat{\rho}(x) = \int_{\mathcal{Z}} h(z)_i d\hat{\mu}(z) = \int_{\mathcal{Z}} \frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{k=1}^N p_k \frac{d\hat{\mu}_k}{d\nu}(z)}\, d\hat{\mu}(z) = p_i \int_{\mathcal{Z}} d\hat{\mu}_i(z) = p_i,$$

for any $\omega_i \in \Omega$ so it is a plausible forecast. □

Intermediate Lemmas. Before proving Theorem 1, we first prove two lemmas. The following lemma establishes when a measure over the signal space can be part of a model representing a given updating rule.

Lemma 3. *(i) Updating rule h can be represented by an admissible model $\{\hat{\mu}_i(\cdot;p)\} \in \mathcal{M}(p)$ at prior p with unconditional signal distribution $\hat{\mu}(\cdot;p)$ iff for all $\omega_i \in \Omega$,*

$$\int_{\mathcal{Z}} h(z,p)_i\, d\hat{\mu}(z;p) = p_i. \tag{11}$$

If such a representation exists, then for any ω_i with $p_i > 0$, $\hat{\mu}_i(Z;p) = \frac{1}{p_i}\int_Z h(z,p)_i\, d\hat{\mu}(z;p)$ for any measurable set of signal realizations $Z \subset \mathcal{F}$. (ii) Updating rule h can be represented by an admissible model $\{\hat{\mu}_i(\cdot;p)\} \in \mathcal{M}(p)$ at prior p with conditional signal distribution $\hat{\mu}_j(\cdot;p) \in \mathcal{M}_{uc}$ in state ω_j iff

$$\int_{\mathcal{Z}} \frac{h(z,p)_i}{h(z,p)_j} d\hat{\mu}_j(z;p) = \frac{p_i}{p_j} \tag{12}$$

for all $\omega_i \in \Omega$. If such a representation exists, then for any ω_i with $p_i > 0$, $\hat{\mu}_i(Z) = \frac{p_j}{p_i}\int_Z \frac{h(z)_i}{h(z)_j}\, d\hat{\mu}_j(z)$ for any measurable set of signal realizations $Z \subset \mathcal{F}$.

Proof. Fix an updating rule h. **Part 1:** (If) Suppose h can be represented by a model with unconditional signal distribution $\hat{\mu}$. It follows from standard argument that beliefs must be a martingale, which implies $\int_{\mathcal{Z}} h(z)_i\, d\hat{\mu}(z) = p_i$. (Only If) Now

suppose that $\hat{\mu}$ is a measure with $\int_{\mathcal{Z}} h(z)_i \, d\hat{\mu}(z) = p_i$. Define conditional distributions $\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i \, d\hat{\mu}(z)$ for all $Z \in \mathcal{F}$. These are probability distributions, as $h(z)_i$ is non-negative and $\hat{\mu}_i(\mathcal{Z}) = 1$ by construction. It remains to show this model induces the posterior prescribed by h following each signal realization z. Since $\hat{\mu}_i$ is absolutely continuous with respect to $\hat{\mu}$, Bayes rule with respect to $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ and the properties of the Radon-Nikodym derivative imply that, μ-a.e.,

$$\frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\nu}(z)} = \frac{p_i \frac{d\hat{\mu}_i}{d\hat{\mu}}(z)}{\sum_{j=1}^N p_j \frac{d\hat{\mu}_j}{d\hat{\mu}}(z)} = h(z)_i,$$

so these distributions induce the posterior prescribed by h. Finally, for the above equation to hold, any misspecified model that represents h must solve

$$\begin{pmatrix} p_1/h(z)_1 & -p_2/h(z)_2 & 0 & \dots & 0 \\ p_1/h(z)_1 & 0 & -p_3/h(z)_3 & \dots & 0 \\ \vdots & & \ddots & & \\ p_1/h(z)_1 & 0 & \dots & 0 & -p_N/h(z)_N \\ p_1 & p_2 & \dots & p_{N-1} & p_N \end{pmatrix} \begin{pmatrix} \frac{d\hat{\mu}_1}{d\hat{\mu}}(z) \\ \frac{d\hat{\mu}_2}{d\hat{\mu}}(z) \\ \vdots \\ \frac{d\hat{\mu}_N}{d\hat{\mu}}(z) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

$\hat{\mu}$-a.s. Therefore, the conditional distributions are unique as the left-hand matrix is an $N \times N$ full-rank matrix.

Part 2. (If) Suppose h can be represented by a misspecified model with conditional signal distribution $\hat{\mu}_j$. Then, by standard argument, for any ω_i the likelihood ratios $h(z)_i/h(z)_j$ must be martingales with respect to $\hat{\mu}_j$ so $\int_{\mathcal{Z}} \frac{h(z)_i}{h(z)_j} d\hat{\mu}_j(z) = \frac{p_i}{p_j}$. (Only If) Now suppose that $\hat{\mu}_j$ is a measure that satisfies $\int_{\mathcal{Z}} \frac{h(z)_i}{h(z)_j} d\hat{\mu}_j(z) = \frac{p_i}{p_j}$ for updating rule h and all i. Define the misspecified model $\hat{\mu}_i(Z) = \int_Z \frac{p_j}{p_i} \frac{h(z)_i}{h(z)_j} d\hat{\mu}_j(z)$. This is a misspecified model that induces updating rule $h(z)$. Without loss of generality assume $j = 1$. Then any family of misspecified models with updating rule h and conditional signal distribution $\hat{\mu}_1$ must solve

$$\begin{pmatrix} 0 & \frac{p_2}{p_1}\frac{h(z)_1}{h(z)_2} & 0 & \dots & 0 \\ 0 & 0 & \frac{p_3}{p_1}\frac{h(z)_1}{h(z)_3} & \dots & 0 \\ \vdots & & \ddots & & \\ 0 & 0 & \dots & 0 & \frac{p_N}{p_1}\frac{h(z)_1}{h(z)_N} \\ \frac{1}{p_1} & -\frac{p_2}{p_1} & \dots & -\frac{p_{N-1}}{p_1} & -\frac{p_N}{p_1} \end{pmatrix} \begin{pmatrix} \frac{d\hat{\mu}}{d\hat{\mu}_1}(z) \\ \frac{d\hat{\mu}_2}{d\hat{\mu}_1}(z) \\ \vdots \\ \frac{d\hat{\mu}_N}{d\hat{\mu}_1}(z) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{pmatrix}$$

$\hat{\mu}_1$ a.s. so this model is unique since the left-most matrix is full-rank. □

Lemma 4 (Construction of Representation)**.** *Consider an updating rule h and a plausible belief forecast $\hat{\rho}$ that satisfy no unexpected beliefs, and let ρ_h be the accurate forecast for h. Then h and $\hat{\rho}$ are represented by the family of models $\{\hat{\mu}_i(\cdot, p)\}_{\omega_i \in \Omega, p \in \Delta(p)}$ with, for each $\omega_i \in \Omega$ and $p \in \Delta(\Omega)$,*

$$\hat{\mu}_i(Z; p) = \frac{1}{p_i} \int_Z h(z,p)_i \frac{d\hat{\rho}}{d\rho_h}(h(z,p); p)\, d\mu(z; p) \tag{13}$$

for any measurable set of signal realizations $Z \in \mathcal{F}$.

Proof. By assumption, $\hat{\rho}$ is absolutely continuous with respect to ρ_h, so $\frac{d\hat{\rho}}{d\rho_h}$ exists. For any Borel set X, define

$$\hat{\rho}_i(X) \equiv \int_X \frac{x_i}{p_i} \frac{d\hat{\rho}}{d\rho_h}(x) d\rho_h(x) = \int_{h^{-1}(X)} \frac{h(z)_i}{p_i} \frac{d\hat{\rho}}{d\rho_h}(h(z))\, d\mu(z)$$

where the second equality follows from change of variables. These are probability measures, and $\sum p_i \hat{\rho}_i(X) = \hat{\rho}(X)$. For any $Z \in \mathcal{F}$, define

$$\hat{\mu}_i(Z) \equiv \int_Z \frac{1}{p_i} h(z)_i \frac{d\hat{\rho}}{d\rho_h}(h(z))\, d\mu(z).$$

We are integrating a measurable function over a measurable set, so the model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ is indeed a family of measures over $(\mathcal{Z}, \mathcal{F})$. This is a probability measure as

$$\hat{\mu}_i(\mathcal{Z}) = \int_{\mathcal{Z}} \frac{1}{p_i} h(z)_i \frac{d\hat{\rho}}{d\rho_h}(h(z))\, d\mu(z) = \int_{\Delta(\Omega)} \frac{1}{p_i} x_i \frac{d\hat{\rho}}{d\rho_h}(x)\, d\rho_h(x) = 1.$$

Model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ clearly induces the the specified updating rule h, as $\frac{d\hat{\rho}}{d\rho_h}$ is non-zero a.s. over the support of ρ_h by mutual absolute continuity. It remains to show that this induces the desired forecast, i.e., $\hat{\mu} \circ h^{-1}(X) = \hat{\rho}(X)$ for any Borel set X. For any Borel set X, note that

$$\hat{\rho}_i(X) = \int_X \frac{x_i}{p_i} \frac{d\hat{\rho}}{d\rho_h}(x)\, d\rho_h(x) = \int_{h^{-1}(X)} \frac{h(z)_i}{p_i} \frac{d\hat{\rho}}{d\rho_h}(h(z))\, d\mu(z) = \hat{\mu}_i(h^{-1}(X)),$$

and therefore, $\hat{\mu}(h^{-1}(X)) = \sum p_i \hat{\rho}_i(X) = \hat{\rho}(X)$. This establishes that $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ induces the desired forecast. □

Proof of Theorem 1. Lemma 4 establishes sufficiency, since the model defined in (13) represents h and $\hat{\rho}$. We use Lemmas 2 and 3 to establish necessity and uniqueness. By Lemma 2, the forecast must be plausible. Suppose there exists a Borel set X such that $\rho_h(X) > 0$ but $\hat{\rho}(X) = 0$ and a subjective model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ that induces the desired forecast and updating rule exists. Let $Z = h^{-1}(X)$. Then by the mutual absolute continuity of the misspecified and correctly specified measures, $0 = \hat{\mu}(Z) = \mu(Z) = \rho_h(X) > 0$, which is a contradiction. Nearly identical logic implies that it's impossible for $\rho_h(X) = 0$ but $\hat{\rho}(X) > 0$. Therefore, ρ_h and $\hat{\rho}$ must be mutually absolutely continuous. Uniqueness of the representation on $\mathcal{F}_h$ follows from Lemma 3. Fix a model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ that represents h and $\hat{\rho}$. For any $Z \in \mathcal{F}_h$, the unconditional measure $\hat{\mu}(Z)$ must satisfy $\hat{\mu}(Z) = \hat{\rho} \circ h(Z)$. Since the model $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ induces $\hat{\mu}$ and h when restricted to the measurable space $(\mathcal{Z}, \mathcal{F}_h)$, this implies that

$$\hat{\mu}_i(Z) = \int_Z h(z)_i d\hat{\mu}(z) = \int_Z h(z)_i d\hat{\rho}(h(z)),$$

so these conditional measures are unique. Note that (13) is equal to (6) on $\mathcal{F}_h(p)$. □

Proof of Proposition 1. This result is immediate from Theorem 1. It follows from that result that for any given updating rule h and accurate forecast ρ_h, there exists a representation at p if and only if ρ_h is plausible and h and ρ_h satisfy no unexpected beliefs. Since ρ_h is mutually absolutely continuous with itself no unexpected beliefs is always satisfied, implying Proposition 1. □

Proof of Proposition 2. (If) The existence of a misspecified model with forecast ρ_B follows from Theorem 1, since ρ_B is plausible because it is the correctly specified forecast. For any Borel set X such that $Z = h^{-1}(X)$, note that $\hat{\mu}_i(Z) = \frac{1}{p_i} \int_Z h(z)_i d\rho_B(h(z)) = \mu_i(h_B^{-1}(X)) = \mu_i(h_B^{-1}(h(Z)))$ by construction of $\hat{\mu}_i$, (6), and the definition of ρ_B. (Only If) Let $\rho_B = \mu(h^{-1}(X))$ be the accurate Bayesian forecast. Suppose there exists a naive-PC representation $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ and there exists a Borel set X s.t. $\rho_B(X) > 0$ but $\hat{\rho}(X) = 0$. Then $\hat{\mu}(h^{-1}(X)) = 0$, which by absolute continuity implies $\mu(h^{-1}(X)) = 0$. But this implies that $\mu(h_B^{-1}(X)) = 0$ which is a contradiction. Similar logic applies when $\rho_B(X) = 0$ but $\hat{\rho}(X) > 0$. □

Proof of Proposition 3. This result follows immediately from Theorem 1. □

B Calculations from Section 5

Proof of Proposition 4. When $p > 1/2$, the true model we outline in Section 5.1 corresponds to

	l	m	r
L	$(1-q(p))\theta$	$q(p)\theta$	$1-\theta$
R	$(1-q(p))(1-\theta)$	$q(p)(1-\theta)$	θ

and when $p \leq 1/2$, it corresponds to

	l	m	r
L	θ	$q(p)(1-\theta)$	$(1-q(p))(1-\theta)$
R	$1-\theta$	$q(p)\theta$	$(1-q(p))\theta$

.

It is straightforward to extend the Bohren and Hauser (2021) framework to allow the signal misspecification to map two signals that induce the same true posterior to different misspecified posteriors. We first characterize the locally stable set $\Lambda(L)$. Let $\tilde{\nu} \equiv \nu(0)$ denote the slant at certainty (recall $\nu(1) = \nu(0)$ by assumption), and similarly, let $\tilde{q} \equiv q(0)$ denoted the misinterpretation probability at certainty. For $p < 1/2$, any model representing this updating rule satisfies

$$\frac{\hat{\mu}(m|R,p)}{\hat{\mu}(m|L,p)} = \frac{(1-\nu(p))h_B(r,p) + \nu(p)h_B(l,p)}{1-(1-\nu(p))h_B(r,p) - \nu(p)h_B(l,p)} * \frac{1-p}{p}.$$

At $p = 0$, this simplifies to

$$\frac{\hat{\mu}(m|R,p)}{\hat{\mu}(m|L,p)} = \frac{(1-\tilde{\nu})\theta}{1-\theta} + \frac{(1-\theta)\tilde{\nu}}{\theta}.$$

Therefore, the local stability of correct learning is determined by the sign of

$$\gamma(0,L) \quad = \quad (1-\tilde{q})(1-\theta)\log\left(\frac{\theta}{1-\theta}\right) + \tilde{q}(1-\theta)\log\left(\frac{(1-\tilde{\nu})\theta}{1-\theta} + \frac{\tilde{\nu}(1-\theta)}{\theta}\right) + \theta\log\left(\frac{1-\theta}{\theta}\right).$$

At $\tilde{q} = 0$, agents have a correctly specified model at certainty, so $\gamma(0,L) < 0$. As $\tilde{q}$ increases, more weight is placed on the second term and less weight is placed on the first term. The second term is less than the first term; therefore, $\gamma(0,L)$ is decreasing in $\tilde{q}$. Therefore, for all $\tilde{q}$ and $\tilde{v}$, $\gamma(0,L) < 0$ and correct learning is locally stable, $0 \in \Lambda(L)$.

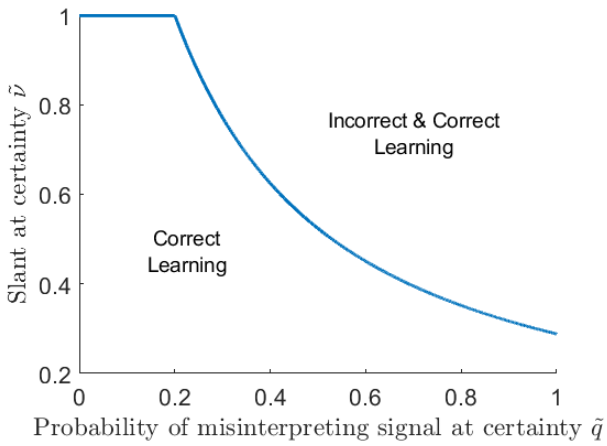


FIGURE 3. Confirmation Bias ($\theta = 0.625$)

Similarly, the local stability of incorrect learning is determined by the sign of

$$\gamma(\infty, L) \quad = \quad \theta(1-\tilde{q}) \log\left(\frac{1-\theta}{\theta}\right) + \theta\tilde{q} \log\left(\frac{(1-\tilde{\nu})(1-\theta)}{\theta} + \frac{\tilde{\nu}\theta}{1-\theta}\right) + (1-\theta) \log\left(\frac{\theta}{1-\theta}\right).$$

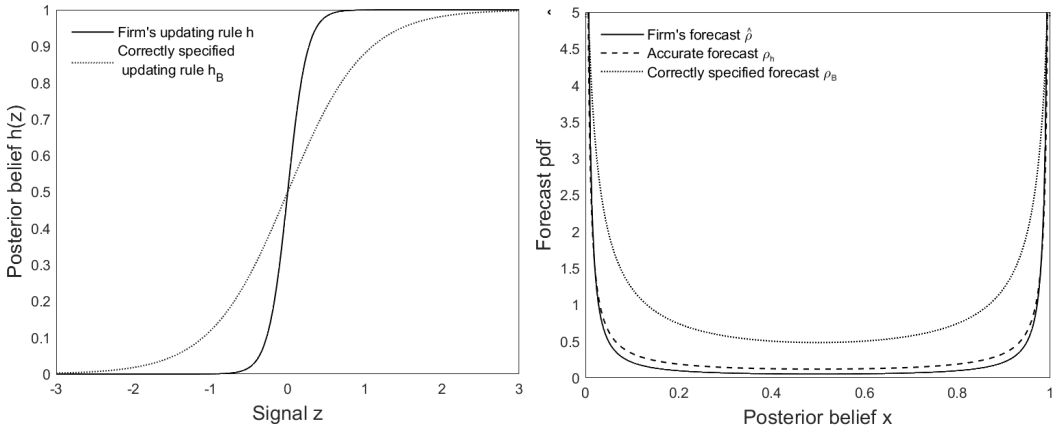
At $\tilde{q} = 0$ or $\tilde{\nu} = 0$, agents have a correctly specified model, so $\gamma(\infty, L) < 0$. As $\tilde{q}$ increases, more weight is placed on the second term and less weight is placed on the first term. The second term is greater than the first term; therefore, $\gamma(\infty, L)$ is increasing in $\tilde{q}$. Similarly, the second term is increasing in $\tilde{\nu}$, and therefore, so is $\gamma(\infty, L)$. At $\tilde{q} = 1$ and $\tilde{\nu} = 1$,

$$\gamma(\infty, L) = \log\left(\frac{\theta}{1-\theta}\right) > 0. \tag{14}$$

Therefore, the desired cutoffs $\overline{q} \in (0,1)$ and $\overline{\nu}(\tilde{q}) \in (0,1)$ exist such that incorrect learning is locally stable for $\tilde{q} > \overline{q}$ and $\tilde{\nu} > \overline{\nu}(\tilde{q})$. The construction of $\Lambda(R)$ is analogous. Given that there is a single type, mixed learning does not arise. Therefore, by Theorem 4 in Bohren and Hauser (2021), $\Lambda(\omega)$ fully characterizes asymptotic learning outcomes. □

Calculations from Section 5.2.

Derivation of the misspecified normal belief forecast. From the normal distribution,



(a) Updating Rule: Overreaction (b) Belief Forecast: Overprecision

FIGURE 4. Decomposing Retrospective and Prospective Bias ($\hat{\sigma}^2 = 0.25$)

$h^{-1}(x)$:

$$h(z) = \frac{e^{-\frac{1}{2}\left(\frac{z-1}{\hat{\sigma}}\right)^2}}{e^{-\frac{1}{2}\left(\frac{z-1}{\hat{\sigma}}\right)^2} + e^{-\frac{1}{2}\left(\frac{z+1}{\hat{\sigma}}\right)^2}} = x$$
$$\iff h^{-1}(x) = -\frac{\hat{\sigma}^2}{2}\log\left(\frac{1-x}{x}\right).$$

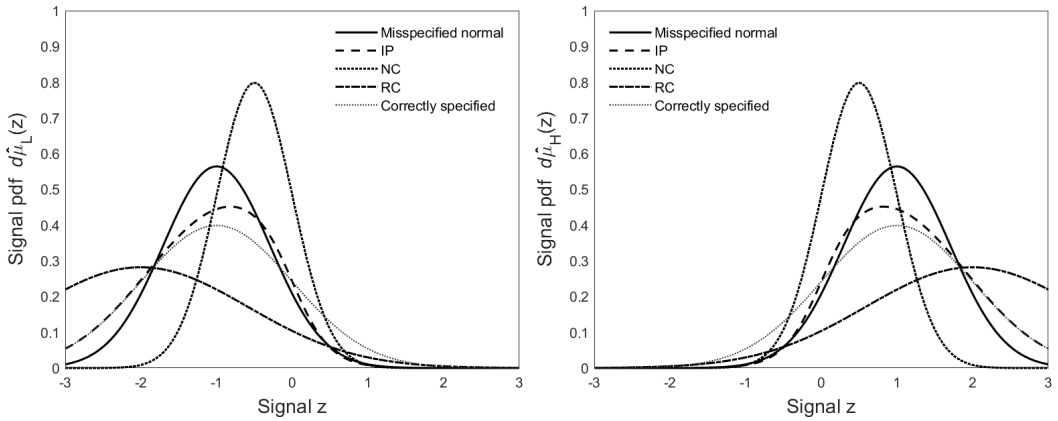
Because z is the mixture of two normal distributions, its subjective unconditional cdf is given by $\hat{\mu}(z) = (\Phi(z|1,\hat{\sigma}) + \Phi(z|-1,\hat{\sigma}))/2$, where $\Phi(z|m,\sigma)$ denotes the cdf of the normal distribution with mean m and variance σ^2. From $\hat{\rho}(x) = \hat{\mu}(h^{-1}(x))$, computing the pdf $d\hat{\rho}(x)$ is straightforward:

$$d\hat{\rho}(x;p) = d\hat{\mu}(h^{-1}(x,p)) \times \frac{\hat{\sigma}^2}{2}\left(\frac{1}{x-x^2}\right),$$

where $d\hat{\mu}(z) = (\phi(z|1,\hat{\sigma}) + \phi(z|-1,\hat{\sigma}))/2$ denotes the subjective unconditional pdf of z. Fig. 4 plots this forecast for $\hat{\sigma}^2 = 0.25$, as well as the correctly specified forecast which corresponds to $\hat{\sigma}^2 = 1$.

Derivation of the sophisticated-PC representation. The sophisticated-PC forecast has cdf $\rho_h(x) = \mu(h^{-1}(x))$ and pdf

$$d\rho_h(x) = d\mu(h^{-1}(x)) \times \frac{\hat{\sigma}^2}{2}\left(\frac{1}{x-x^2}\right).$$



(a) Perceived signal distribution in state L (b) Perceived signal distribution in state H

FIGURE 5. Misspecified model representations ($\hat{\sigma}^2 = 0.5$)

A misspecified model representation exists if ρ_h is plausible, which holds at $p = 1/2$: This representation corresponds to $\hat{\mu}_H^{IP}(z) = 2\int_{-\infty}^{z} h(u)\, d\mu(u)$ and $\hat{\mu}_L^{IP}(z) = 2\int_{-\infty}^{z}(1 - h(u))\, d\mu(u)$. Fig. 4 plots the accurate forecast and Fig. 5 plots the pdfs of $\hat{\mu}_H^{IP}$ and $\hat{\mu}_L^{IP}$.

Derivation of the naive-PC representation. The correctly specified forecast has cdf $\rho_B(x) = \mu(h_B^{-1}(x))$ and pdf

$$d\rho_B(x) = d\mu(h_B^{-1}(x)) \times \frac{1}{2}\left(\frac{1}{x - x^2}\right).$$

A misspecified model representation exists if h and ρ_B satisfy no unexpected beliefs, which is satisfied since both models induce all beliefs $x \in (0, 1)$. This representation corresponds to: $\hat{\mu}_H^{NC}(z) = 2\int_{-\infty}^{z} h(u) d\rho_B(h(u))$ and $\hat{\mu}_L^{NC}(z) = 2\int_{-\infty}^{z}(1 - h(u)) d\rho_B(h(u))$. Note that we need to be careful here in terms of how we take the integral. It is not $(d\rho_B)(h(z))$, but $d(\rho_B(h(z)))$. Explicitly, we have:

$$d(\rho_B(h(z))) = d\mu\left(\frac{\sigma^2}{\hat{\sigma}^2} z\right) \times \frac{\sigma^2}{\hat{\sigma}^2}.$$

Fig. 5 plots the pdfs of $\hat{\mu}_H^{NC}$ and $\hat{\mu}_L^{NC}$.

Derivation of the retrospectively-correct representation. The RC updating rule is h_B. A misspecified model representation exists if h_B and misspecified normal forecast

$\hat{\rho}$ satisfy no unexpected beliefs, which is satisfied since all beliefs $x \in (0,1)$ are induced by both (we already know $\hat{\rho}$ is plausible, which is the other condition). The representation corresponds to $\hat{\mu}_H^{RC}(z) = 2\int_{-\infty}^{z} h_B(u)d\hat{\rho}(h_B(u))$ and $\hat{\mu}_L^{RC}(z) = 2\int_{-\infty}^{z}(1-h_B(u))d\hat{\rho}(h_B(u))$, where

$$d(\hat{\rho}(h_B(z))) = d\hat{\mu}\left(\frac{\hat{\sigma}^2}{\sigma^2}z\right) \times \frac{\hat{\sigma}^2}{\sigma^2}.$$

Fig. 5 plots the pdfs of $\hat{\mu}_H^{RC}$ and $\hat{\mu}_L^{RC}$.

C Additional Examples

C.1 Discrete Signal Example

The following example illustrates the results from Section 3. Consider binary state space $\Omega = \{L, R\}$ with a uniform prior and signal space $\mathcal{Z} = \{z_1, z_2, z_3, z_4\}$. Given the focus on a uniform prior, we suppress p as an argument of the updating rule, forecast, and model. Moreover, in a slight abuse of notation given the binary state space, we define the updating rule as the probability assigned to state R after observing each signal, i.e. $h(z) = Pr(R|z)$ for each $z \in \mathcal{Z}$ and the belief forecast as a distribution $\hat{\rho}(x)$ over set of probabilities x that the state is R. Note Definition 2 requires $|\operatorname{supp}\hat{\rho}(x)| \leq 4$. In this set-up, a model corresponds to a pair of vectors $\{\hat{\mu}_L, \hat{\mu}_R\}$, where each vector specifies a subjective probability $m_{\omega,k}$ for each signal z_k in each state ω, i.e. $\hat{\mu}_\omega = (m_{\omega,1}, m_{\omega,2}, m_{\omega,3}, m_{\omega,4})$ with $\sum_{k=1}^{4} m_{\omega,k} = 1$.

Illustration of Lemma 1. A Bayes-feasible updating rule maps at least one signal to a posterior above the prior and one signal to a posterior below the prior, i.e. $\min_z h(z) < 0.5 < \max_z h(z)$. Given a Bayes-feasible updating rule h, any solution $(m_1, m_2, m_3, m_4) \in \Delta^3$ to $\sum_{k=1}^{4} h(z_k)m_k = 0.5$ pins down a model that represents h, with $m_{R,k} = 2h(z_k)m_k$ and $m_{L,k} = 2(1-h(z_k))m_k$ for $k = 1, ..., 4$.[21] Aside from knife-edge cases, there are multiple solutions, and therefore, multiple representations. For example, if $h(z_1) = 0.1$, $h(z_2) = 0.2$, $h(z_3) = 0.8$ and $h(z_4) = 0.9$, then $(0.2, 0.3, 0.3, 0.2)$ and $(0.1, 0.4, 0.4, 0.1)$ are both solutions (in fact, there are a continuum of solutions). Note that each model induces a unique belief forecast, which assigns probability $m_k = (m_{R,k} + m_{L,k})/2$ to posterior belief $h(z_k)$.

[21] To see that any such model represents h, note that it induces posterior belief $m_{R,k}/(m_{R,k} + m_{L,k}) = h(z_k)$ following signal realization z_k, and therefore, it induces the desired updating rule.

Illustration of Lemma 2. A forecast $\hat{\rho}$ is plausible if $\sum_{x \in \text{supp}\, \hat{\rho}(x')} x\hat{\rho}(x) = 0.5$. For example, the forecast $\hat{\rho}(x) = 0.5\mathbb{1}\{x \in \{x_1, 1-x_1\}\}$ for some $x_1 \in (0, .5)$ is plausible since $.5x_1 + .5(1-x_1) = 0.5$. One such model that represents this forecast is $m_{R,1} = x_1/2$, $m_{R,2} = x_1/2$, $m_{R,3} = (1-x_1)/2$ and $m_{R,4} = (1-x_1)/2$ in state R, and similarly for state L substituting $1-x_1$ for x_1.[22] This model induces an updating rule that maps $\{z_1, z_2\}$ to posterior x_1 and $\{z_3, z_4\}$ to $1-x_1$.[23] Alternatively, the model $m_{R,1} = x_1/3$, $m_{R,2} = x_1/3$, $m_{R,3} = x_1/3$ and $m_{R,4} = 1-x_1$ in state R, and similarly for state L substituting $1-x_1$ for x_1, also represents $\hat{\rho}$. This model induces a different updating rule: it maps $\{z_1, z_2, z_3\}$ to x_1 and z_4 to $1-x_1$. In fact, for any updating rule that assigns at least one signal to each posterior x_1 and $1-x_1$, it is possible to find a model that induces this updating rule and represents $\hat{\rho}(x)$. As discussed above, each updating rule induces a different retrospective bias. For example, if the correct model maps $\{z_1, z_2\}$ to posterior x_1, then mapping $\{z_1, z_2, z_3\}$ to x_1 slants information towards state L, whereas mapping $\{z_1, z_3\}$ to x_1 inverts the interpretation of z_2 and z_3.

Illustration of Theorem 1. Consider the plausible forecast $\hat{\rho}(x) = 0.5 * \mathbb{1}\{x \in \{x_1, 1-x_1\}\}$ for $x_1 \in (0, .5)$. Then any updating rule with $h(z) \in \{x_1, 1-x_1\}$ for all $z \in \mathcal{Z}$ satisfies no unexpected beliefs. Consider $h(z_1) = h(z_2) = x_1$ and $h(z_3) = h(z_4) = 1-x_1$. Given that $h(z)$ maps $\{z_1, z_2\}$ to the same posterior and similarly for $\{z_3, z_4\}$, the σ-algebra generated by $h(z)$ is $\mathcal{F}_h(p) = \{\emptyset, \{z_1, z_2\}, \{z_3, z_4\}, \mathcal{Z}\}$. From (6), h and $\hat{\rho}$ have an essentially unique representation at $p = 0.5$ that satisfies $\hat{\mu}_R(\{z_1, z_2\}) = x_1$ and $\hat{\mu}_R(\{z_3, z_4\}) = 1-x_1$ in state R and $\hat{\mu}_L(\{z_1, z_2\}) = 1-x_1$ and $\hat{\mu}_L(\{z_3, z_4\}) = x_1$ in state L. Applying Lemma 4, one such representation is $\hat{\mu}_i(z_k) = \left(\frac{\mu(z_k)}{\mu(z_1)+\mu(z_2)}\right)\hat{\mu}_i(\{z_1, z_2\})$ for $k = 1, 2$ and $\hat{\mu}_i(z_k) = \left(\frac{\mu(z_k)}{\mu(z_3)+\mu(z_4)}\right)\hat{\mu}_i(\{z_3, z_4\})$ for $k = 3, 4$.

[22] To see that this model represents $\hat{\rho}(x)$, recall that z_k induces posterior belief $m_{R,k}/(m_{R,k} + m_{L,k})$. This simplifies to posterior belief x_1 following z_1 and z_2 and $1-x_1$ following z_3 and z_4. Therefore, it induces forecast $\hat{\rho}(x_1) = \hat{\mu}(\{z_1, z_2\}) = (m_{R,1} + m_{L,1})/2 + (m_{R,2} + m_{L,2})/2 = 0.5$ and $\hat{\rho}(1-x_1) = \hat{\mu}(\{z_3, z_4\}) = 0.5$ by an analogous calculation, as desired.

[23] To motivate the notion of essential uniqueness (Definition 4), note that any $\alpha \in (0, 1)$ pins down a model that represents $\hat{\rho}(x)$ with $m_{R,1} = \alpha x_1$, $m_{R,2} = (1-\alpha)x_1$, $m_{R,3} = \alpha(1-x_1)$ and $m_{R,4} = (1-\alpha)(1-x_1)$ in state R, and similarly for state L substituting $1-x_1$ for x_1. For each α, the corresponding model induces the same updating rule. Therefore, all models in this class induce the same forecast and updating rule, and are considered equivalent under Definition 4.

Illustration of Proposition 1. Consider updating rule $h(z_1) = h(z_2) = x_1$ and $h(z_3) = h(z_4) = 1 - x_1$ for some $x_1 \in (0, 1)$. The accurate forecast corresponds to $\rho_h(x_1) = \mu(z_1) + \mu(z_2)$ and $\rho_h(1 - x_1) = \mu(z_3) + \mu(z_4)$, where μ is the correct unconditional model. A sophisticated-PC representation of h exists if this forecast is plausible, i.e., $x_1(\mu(z_1)+\mu(z_2))+(1-x_1)(\mu(z_3)+\mu(z_4)) = 0.5$. Note that this condition depends on the *true* signal measure; it is satisfied if either an equal mass of signals map to each posterior, $\mu(z_1) + \mu(z_2) = \mu(z_3) + \mu(z_4) = 0.5$, or the signal is perceived to be uninformative, $x_1 = 0.5$. To construct a sophisticated-PC representation of h, suppose the true unconditional signal distribution is $\mu(z) = (0.2, 0.3, 0.3, 0.2)$. From (7), the unique introspection-proof sophisticated-PC representation is $\hat{\mu}_1(z) = (.4(1 - x_1), .6(1 - x_1), .6x_1, .4x_1)$ and $\hat{\mu}_2(z) = (.4x_1, .6x_1, .6(1 - x_1), .4(1 - x_1))$.

Illustration of Proposition 2. Again consider updating rule $h(z_1) = h(z_2) = x_1$ and $h(z_3) = h(z_4) = 1 - x_1$ for some $x_1 \in (0, 1)$. Suppose the correct model induces updating rule $h_B(z_1) = h_B(z_2) = h_B(z_3) = x_1$ and $h_B(z_4) = 1 - x_1$. Then $\rho_B(x_1) = \mu(\{z_1, z_2, z_3\})$ and $\rho_B(1 - x_1) = \mu(z_4)$, where μ is the correct unconditional model. Given that the updating rule h induces set of posteriors $\{x_1, 1 - x_1\}$, which is equal to the support of ρ_B, h and ρ_B satisfy no unexpected beliefs. Therefore, a naive-PC representation of h exists. From (8), this representation is unique on $\mathcal{F}_h(0.5) = \{\{z_1, z_2\}, \{z_3, z_4\}, \mathcal{Z}\}$ and satisfies $\hat{\mu}_i(\{z_1, z_2\}) = \mu_i(\{z_1, z_2, z_3\})$ and $\hat{\mu}_i(\{z_3, z_4\}) = \mu_i(z_4)$ for $\omega_i \in \Omega$.

Illustration of Proposition 3. Suppose the correct model induces updating rule $h_B(z_1) = 0.1$, $h_B(z_2) = 0.2$, $h_B(z_3) = 0.8$, and $h_B(z_4) = 0.9$ and belief forecast $\rho_B = (0.1, 0.4, 0.4, 0.1)$ over posteriors $(0.1, 0.2, 0.8, 0.9)$. Consider a prospectively overprecise belief forecast $\hat{\rho} = (0.4, 0.1, 0.1, 0.4)$ over posteriors $(0.1, 0.2, 0.8, 0.9)$ that places more weight on the extreme posteriors and less weight on the interior posteriors relative to ρ_B. This forecast is plausible and satisfies no unexpected beliefs with respect to h_B. Therefore, a RC representation of $\hat{\rho}$ exists. From (6), it is equal to $\hat{\mu}_1 = (.72, .16, .04, .08)$ and $\hat{\mu}_2 = (.08, .04, .16, .72)$. Note this model is misspecified, since from h_B and ρ_B, the correctly specified model is $\mu_1 = (.18, .64, .16, .02)$ and $\mu_2 = (.02, .16, .64, .18)$.

C.2 Linear Under- and Overreaction

The following example illustrates the multiplicity of representations for the linear under- and overreaction updating rule from Example 1 (Epstein et al. 2010) and shows that it has a sophisticated-PC representation. A common updating rule for underreaction models the posterior belief as a weighted average of the Bayesian posterior and the prior,

$$h(z,p) = \alpha h_B(z;p) + (1-\alpha)p$$

for $\alpha \in [0,\infty)$ (such that $h(z,p)$ is a probability for any $z \in \mathcal{Z}$). We use Lemma 3 to find misspecified models that represent this updating rule.

First consider a sophisticated-PC misspecified model, i.e., $\hat{\mu} = \mu$. We show that such a model is pinned down by the true unconditional measure μ, the Bayesian updating rule h_B, and the bias parameter α, independent of the details of the information environment. When $\hat{\mu} = \mu$, $\hat{\mu}$ satisfies (11) as $\int_{\mathcal{Z}} h_B(z;p)_i \, d\hat{\mu}(z;p) = \int_{\mathcal{Z}} h_B(z;p)_i \, d\mu(z;p) = p_i$ by standard argument, and therefore, $\int_{\mathcal{Z}} (\alpha h_B(z;p)_i + (1-\alpha)p_i) \, d\hat{\mu}(z;p) = p_i$. In this case, the subjective distribution in state ω_i must be equal to:

$$\frac{d\hat{\mu}_i}{d\nu}(z;p) = \left[\frac{\alpha}{p_i} h_B(z;p)_i + (1-\alpha)\right] \frac{d\mu}{d\nu}(z;p).$$

In this representation, the agent correctly anticipates her future beliefs but retrospectively underreacts to the signal.

To construct an alternative representation, we need to put more structure on the information environment. Consider a setting with $|\Omega| = 2$, $\mathcal{Z} = [0,1]$, a uniform prior and true unconditional signal distribution, and $|h_B(z;p)_1 - \frac{1}{2}|$ symmetric about $z = 1/2$. Then the model that induces subjective unconditional pdf $d\hat{\mu}(z;p) = 3/2 - 6(z-1/2)^2$ (in a slight abuse of notation, using $d\hat{\mu}$ to denote the pdf) also satisfies $\int_{\mathcal{Z}} h_B(z;p)_i d\hat{\mu}(z;p) = 1/2$, and therefore, $\int_{\mathcal{Z}} (\alpha h_B(z;p)_i + (1-\alpha)/2) \, d\hat{\mu}(z;p) = 1/2$.[24] In this representation, the agent underestimates the frequency of extreme beliefs—she exhibits prospective underprecision.

C.3 Partisan Bias (distort posterior)

This example shows that a common parameterization of partisan bias does not have a sophisticated-PC representation when the prior is correctly specified but can when

[24] Recall that h_B and α are such that this density is never negative.

the prior is misspecified. It also has a naive-PC representation.

Consider binary state space $\Omega = \{\omega_1, \omega_2\}$ and updating rule $h(z,p)_2 = h_B(z,p)_2^\alpha$ for $\alpha \in (0,1)$ from Example 1. This updating rule exhibits ω_2-partisan bias: after any signal realization, the agent places higher probability on state ω_2 than a correctly specified agent. Let ρ_h denote the accurate distribution over the posterior belief that the state is ω_2. Then $\int_0^1 x\, d\rho_h(x;p) = \int_{\mathcal{Z}} h(z,p)_2\, d\mu(z;p)$ follows from a change of variables and a property of the accurate forecast. But $\int_{\mathcal{Z}} h(z,p)_2\, d\mu(z;p) > \int_{\mathcal{Z}} h_B(z,p)_2\, d\mu(z;p) = p_2$, where the equality follows from Bayes plausibility (i.e., h_B is plausible at p). Therefore, the accurate forecast cannot be plausible at p. This argument clearly applies more generally to any bias that systematically skews posterior beliefs in one direction. Note that with a misspecified prior $\hat{p} \neq p$, it is possible to find a model in which the updating rule averages to the misspecified prior and the Bayesian updating rule averages to the correct prior.

We need to place more structure on the information environment to show that this updating rule has a naive-PC representation. Consider signal space $\mathcal{Z} = [0,1]$, a uniform prior, and for notational simplicity suppress the dependence of the models on p. Suppose the true measures are $\mu_2(z) = z^2$ and $\mu_1(z) = 2z - z^2$ (in a slight abuse of notation, in cdf form). This induces unconditional measure $\mu(z) = (\mu_1(z)+\mu_2(z))/2 = z$, updating rule $h_B(z, 0.5)_2 = \frac{1}{1+d\mu_1/\mu_2(z)} = z$, and belief forecast $\rho_B(x; 0.5) = Pr(z : h_B(z, 0.5)_2 \leq x) = \mu(x) = x$ (in cdf form over the posterior belief that the state is ω_2). The accurate forecast is $\rho_h(x; 0.5) = Pr(z : h(z, 0.5) \leq x) = Pr(z : z \leq x^{1/\alpha}) = \mu(x^{1/\alpha}) = x^{1/\alpha}$. Since ρ_h and ρ_B have the same support, h and ρ_B satisfy no unexpected beliefs. Therefore, a naive-PC representation exists. It is unique since each signal maps to a unique posterior, and from (8), is equal to $\hat{\mu}_i(z) = \mu_i(z^\alpha)$ for $\omega_i \in \Omega$.

C.4 Beta Distribution Belief Forecasts

Suppose $\Omega = \{L, R\}$. Let p denote the prior probability of state R, $\mathcal{Z} = [0,1]$, $\mathcal{F}$ be the Borel σ-algebra, and the correctly specified model be a set of full support distributions over $\mathcal{Z}$. Consider the following parametric family of forecasts, where, in a slight abuse of notation, $d\hat{\rho}_\theta$ denotes the pdf of the forecast over the posterior x

that the state is R:

$$d\hat{\rho}_\theta(x;p) = \frac{x^{\theta-1}(1-x)^{\theta(1-p)/p-1}}{\Gamma(\theta)\Gamma(\theta(1-p)/p)/\Gamma(\theta/p)} \tag{15}$$

for $\theta > 0$.[25] This corresponds to the family of beta distributions with mean p. For any θ, this forecast is plausible since $\int_{\Delta(\Omega)} x\, d\hat{\rho}_\theta(x;p) = p$. To illustrate the multiplicity of representations, suppose $\theta = 1$ and $p = 0.5$, so that $d\hat{\rho}_1(x;0.5) = 1$ is the uniform forecast. For any $\gamma > 0$, the model with pdfs $d\hat{\mu}_R(z) = 2\gamma z^{2\gamma-1}$ and $d\hat{\mu}_L(z) = 2\gamma z^{\gamma-1} - d\hat{\mu}_R(z)$ represents $\hat{\rho}_1(x, 0.5)$.[26] From Bayes rule, this model induces updating rule $h(z, 0.5) = d\hat{\mu}_R(z)/(d\hat{\mu}_R(z) + d\hat{\mu}_L(z)) = z^\gamma$ that the state is R. Each value of γ captures a different level of retrospective bias: as γ decreases, the updating rule slants the interpretation of a given signal realization more in favor of state R.

D Additional Analysis and Discussion

D.1 Misspecified Prior

Let $\hat{p} \equiv (\hat{p}_1, ..., \hat{p}_N) \in \Delta(\Omega)$ denote the agent's subjective prior and assume it has full support. A misspecified prior corresponds to $\hat{p} \neq p$. Let $\mu_i(\cdot;\hat{p}) \in \Delta(\mathcal{Z})$ denote the true signal distribution conditional on state ω_i at subjective prior $\hat{p}$, where, given the motivation that action choices can influence the signal distribution, these true measures depend on the subjective prior (which guides ex-ante actions). Let $\mu(\cdot;\hat{p},p) \equiv \sum_{i=1}^N p_i \mu_i(\cdot;\hat{p})$ denote the true unconditional signal distribution at $\hat{p}$ and true prior p, where the average is taken with respect to the objective prior p. The accurate belief forecast depends on both the objective and subjective priors since it depends on $\mu(\cdot;\hat{p},p)$: $\rho_h(X;\hat{p},p) = \mu(\{z : h(z,\hat{p}) \in X\};\hat{p},p)$, while $h_B(z,\hat{p})$ depends on the subjective prior. The support of the accurate forecast at $\hat{p}$, $\mathcal{X}(h,\hat{p}) \equiv \operatorname{supp} \rho_h(\cdot;\hat{p},p)$, is independent of the objective prior by the mutual absolute continuity of $\mu_i(\cdot;\hat{p})$ and $\mu_j(\cdot;\hat{p})$.[27] The misspecified prior is a primitive of either

[25] Note that $\phi(z) = (z, 1-z)$ satisfies the mutually absolutely continuous condition in Definition 2, and therefore, this is indeed a valid belief forecast.

[26] To see this, note that the unconditional signal cdf is $\hat{\mu}(z;0.5) = z^\gamma$. Given $x = z^\gamma$, this induces forecast cdf $\hat{\mu}(x^{1/\gamma};0.5) = x$ which is the uniform forecast.

[27] From $\mu(\cdot;\hat{p},p) \equiv \sum_{i=1}^N p_i \mu_i(\cdot;\hat{p})$ and mutual absolute continuity, $\mu(\cdot;\hat{p},p)$ has the same support as each $\mu_i(\cdot;\hat{p})$ for any $p \in \Delta(\Omega)$. By (2), $\rho_h(\cdot;\hat{p},p)$ is defined with respect to $\mu(\cdot;\hat{p},p)$, and therefore, has the same support as any measure over posteriors defined with respect to $\mu_i(\cdot;\hat{p})$. Therefore, given that $\mu_i(\cdot;\hat{p})$ is independent of p, so is $\mathcal{X}(h,\hat{p})$.

approach—along with the misspecified signal distributions $\hat{\mu}_i(\cdot;\hat{p})$ or the updating rule and forecast $h(z,\hat{p})$ and $\hat{\rho}(x;\hat{p})$. There is a direct analogue of Theorem 1, substituting the misspecified prior for the correct prior. In particular, the forecast must be plausible with respect to the misspecified prior and the misspecified model is as in (6) given the misspecified prior.

D.2 Time Consistency

In this section we show how a prior-dependent representation can lead to time inconsistency in a dynamic version of our framework. Suppose state ω is drawn at the beginning of the game. An agent observes a sequence of conditionally i.i.d. signals drawn from μ_i when the realized state is ω_i. The agent uses an updating rule and a belief forecast that have a prior-dependent representation with model $(\hat{\mu}_i(\cdot;p))_{\omega_i\in\Omega}$ at prior p. When the agent has belief p, she believes that she will use the updating rule and forecast induced by model $(\hat{\mu}_i(\cdot;p))_{\omega_i\in\Omega}$ in all future periods. This leads to dynamically inconsistent behavior, as the agent's model of how to interpret information changes with her belief but she does not anticipate this. Therefore, the agent may wish to deviate from her ex-ante action strategy after observing the signal and updating her belief, and hence, her model.

D.3 Relaxing Admissibility

The admissible assumption requires that the misspecified model and the correctly specified model are mutually absolutely continuous. This in turn identifies the support of the signal distributions in any model that represents a belief forecast and an updating rule. In Theorem 1, the no unexpected beliefs condition ensures that the updating rule and belief forecast are consistent in terms of which beliefs they induce on this support. We could relax this assumption and provide a similar result without specifying a correctly specified model. Instead of using the correctly specified model to determine the support of the signal distributions, we could use the updating rule and forecast directly. In this case, the condition for a representation to exist can be weakened: there only needs to exist a set of signals on which the updating rule induces the beliefs in the support of the forecast, instead of a specific set pinned down by the correctly specified model.[28]

[28] Formally, no unexpected beliefs can be relaxed to "$\hat{\rho}$ is absolutely continuous with respect to $\nu \circ h^{-1}$". A variation of Theorem 1 then holds under the appropriate reformulation of Definition 3.

D.4 Prior-Independent Representations

The following definition formalizes the notion of a prior-independent representation.

Definition 11 (Prior-Independent Representation)**.** *An updating rule $h(z,p)$ has a* prior-independent representation *if there exists a subjective model $\{\hat{\mu}_i(\cdot)\}_{\omega_i\in\Omega} \in \Delta(\mathcal{Z})^N$ that is admissible at all $p \in \Delta(\Omega)$ and represents $h(z,p)$ at all $p \in \Delta(\Omega)$.*

This is an appealing property for biases in which an agent is inherently Bayesian but has a mistaken understanding of the signal that does not vary with her prior. For example, biases such as overreaction and optimism are not intrinsically linked to the agent's prior. In contrast, the property is conceptually at odds with biases in which the agent's prior influences her perception of information. For example, the prior is a key component of confirmation bias, and therefore, any representation of an updating rule exhibiting confirmation bias naturally varies with it.

The following proposition presents a necessary and sufficient condition for an updating rule to have a prior-independent representation. In particular, such a representation exists if and only if it is possible to factor the prior likelihood ratio p_j/p_i out of the posterior likelihood ratio $h(z,p)_j/h(z,p)_i$ for any pair of states.

Proposition 5 (Prior-Independent Representation)**.** *Fix an updating rule $h(z,p)$ that is Bayes-feasible at all $p \in \Delta(\Omega)$. Then $h(z,p)$ has a prior-independent representation if and only if $\frac{p_i h(z,p)_j}{p_j h(z,p)_i}$ is independent of p for all $p \in \Delta(\Omega)$, $z \in \mathcal{Z}$, and $\omega_i, \omega_j \in \Omega$. When this holds, then any model that represents $h(z,p)$ at prior p' also represents $h(z,p)$ at all other priors $p'' \in \Delta(\Omega)$.*

Proof. (If:) Fix an interior prior $p \in \Delta(\Omega)$. By Lemma 1, there exists a misspecified model $\{\hat{\mu}_i\}_{\omega_i\in\Omega}$ that represents $h(z,p)$ at p. Therefore, by Bayes rule, for μ-almost all z

$$\frac{h(z,p)_i}{h(z,p)_j} = \frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{p_j \frac{d\hat{\mu}_j}{d\nu}(z)}.$$

In this case, the forecast determines the set of beliefs that must be "rationalized", whereas in Theorem 1, the forecast and updating rule must be rationalized over the set of signals that can actually occur.

So the condition from Proposition 5 implies that

$$\frac{h(z,p')_i}{h(z,p')_j} = \frac{p'_i \frac{d\hat{\mu}_i}{d\nu}(z)}{p'_j \frac{d\hat{\mu}_j}{d\nu}(z)}$$

which is exactly the condition $h(z,p')$ must satisfy to be induced by $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ at p'.

(Only If:) Suppose that $h(z,p)$ admits a prior independent representation $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$. By Lemma 1, for every p, $h(z,p) \in S(h(\cdot,p))$. Moreover, by Bayes rule

$$\frac{h(z,p)_i}{h(z,p)_j} = \frac{p_i \frac{d\hat{\mu}_i}{d\nu}(z)}{p_j \frac{d\hat{\mu}_j}{d\nu}(z)},$$

so for any p,p'

$$\frac{p_j h(z,p)_i}{p_i h(z,p)_j} = \frac{p'_j h(z,p')_i}{p'_i h(z,p')_j}.$$

□

When this condition holds, then any model that represents an updating rule at some prior p can form a prior-independent representation.[29]

Many well-known parameterizations of common biases have prior-independent representations. As shown in the examples below, this includes the geometric under/overreaction and partisan bias (distort signal likelihood) updating rules in Example 1. Intuitively, any bias that distorts the true signal likelihoods $\frac{d\mu_i}{d\nu} / \sum_{\omega_j \in \Omega} \frac{d\mu_j}{d\nu}$ independently of the prior will have a prior-independent representation.

This result also establishes when an updating rule does not have a prior-independent representation. Many biases naturally vary with the prior, and updating rules that require a prior-dependent representation are essential for capturing the essence of these biases. For example, the direction of confirmation bias and the magnitude of base rate neglect depend on the prior; the versions of these updating rules in Example 1 require prior-dependent representations. The property is also at odds with some biases in which an agent is non-Bayesian. As shown in the examples below, the linear parameterization of over/underreaction in Example 1 only admits prior-dependent

[29] Whenever the updating rule has at least two representations at p, then trivially a prior-dependent representation also exists. To see this, consider two models that represent h at prior p and suppose both models also represent h at all priors. To form a prior-dependent representation, select one model to represent h at $p \in (0, 0.5]$ and the other model to represent h at $p \in [0.5, 1)$.

representations. Even though the bias parameter is independent of the prior, the additivity of the non-Bayesian updating rule differs structurally from the multiplicative form of Bayes rule and it can only be represented in a framework that imposes Bayesian updating by allowing the model to vary with the prior. Additionally, distorting the Bayesian posterior can link the magnitude of the bias to the prior. As shown in the examples below, the distort posterior version of partisan bias from Example 1 does not have a prior-independent representation. While prior-independent representations lend themselves to more straightforward dynamic analysis, prior-dependent representations are still tractable. For example, recent work in the misspecified learning literature establishes general convergence results when the model varies with the prior (Bohren and Hauser 2021; Frick et al. 2023).

Even when a prior-independent representation exists for a given updating rule, the unique model that represents a forecast-updating rule pair may not be prior-independent due to the dependence of the forecast on the prior. We next show that when an updating rule has a prior-independent representation, then pairing it with the naive-PC forecast results in a prior-independent representation.

Proposition 6 (Naive-PC and Prior-Independence)**.** *Fix an updating rule $h(z,p)$ that has a prior-independent representation and a naive-PC representation at some prior $p \in \Delta(\Omega)$. Then the naive-PC representation is prior-independent.*

This establishes a desirable property of the naive-PC model.

Proof. Fix a prior p where there exists a $\{\hat{\mu}_i\}_{\omega_i \in \Omega}$ that is an essentially unique representation of $h(z,p)$ and ρ_B at prior p. It follows from Proposition 5 that this induces $h(z,p)$ at every prior, as for any p' the likelihood ratio of the updating rule must be the likelihood ratio induced by Bayes rule with respect to the representation,

$$\frac{p'_j}{p'_i}\frac{h(z,p')_i}{h(z,p')_j} = \frac{p_j}{p_i}\frac{h(z,p)_i}{h(z,p)_j} = \frac{\frac{d\hat{\mu}_i}{d\nu}(z)}{\frac{d\hat{\mu}_j}{d\nu}(z)}.$$

By construction, this representation induces ρ_B at p', as for any Borel set X,

$$\rho_B(X;p') = \sum_{i=1}^{N} p'_i \mu_i(\{z : h_B(z) \in X\}) = \sum_{i=1}^{N} p'_i \hat{\mu}_i(h^{-1}(X)).$$

□

We already know that, by definition, the forecast induced by the naive-PC model is the same as the forecast induced by the correctly specified model in a one-period setting. In a dynamic setting with a sequence of signals, the forecast induced by the naive-PC model paired with an updating rule that has a prior-independent representation satisfies a stronger consistency property. While $\hat{\rho}(x;p)$ specifies the period-t forecast of the period-$(t+1)$ posterior belief, in a dynamic setting, one can also define the period-t forecast of the period-$(t+k)$ posterior belief for any $k>1$. The naive-PC representation of an updating rule with a prior-independent representation induces a period-t forecast over period-$(t+k)$ posterior beliefs that is equal to the period-t forecast of period-$(t+k)$ posterior beliefs in the correctly specified model.

Examples. We show that the updating rules modeling geometric overreaction and partisan bias (distort signal likelihood) in Example 1 have a prior-independent representation, while those modeling linear under/overreaction and partisan bias (distort posterior) do not.

Geometric overreaction. Suppose the correctly specified model does not depend on the prior. The geometric over- and underreaction updating rule from Example 1 corresponds to

$$\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_2}{p_1}\left(\frac{z}{1-z}\right)^{\alpha}.$$

It is straightforward to see that it is possible to factor out the prior from this expression, and therefore, it satisfies the condition in Proposition 5 and has a prior-independent representation.

Partisan Bias (distort signal likelihood). Consider the parameterization of partisan bias from Example 1 that distorts the signal likelihood:

$$\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_2}{p_1}\left(\frac{z^{\alpha}}{1-z^{\alpha}}\right).$$

Again, it is straightforward to see that it is possible to factor out the prior from this updating rule, and therefore, it has a prior-independent representation.

Linear overreaction. The linear over- and underreaction updating rule from Exam-

ple 1 does not satisfy the condition in Proposition 5, as

$$\frac{p_1}{p_2}\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_1}{p_2}\frac{\alpha h_B(z,p)_2 + (1-\alpha)p_2}{\alpha h_B(z,p)_1 + (1-\alpha)p_1}$$

clearly depends on the prior since $h_B(z,p)_2 \equiv \frac{p_2 z}{p_2 z + p_1(1-z)}$.

Partisan Bias (distort posterior). Similarly, in the model of partisan bias in Example 1 that distorts the posterior,

$$\frac{p_1}{p_2}\frac{h(z,p)_2}{h(z,p)_1} = \frac{p_1}{p_2}\left(\frac{h_B(z,p)_2}{h_B(z,p)_1}\right)^{\alpha} = \left(\frac{p_1}{p_2}\right)^{1-\alpha}\left(\frac{z}{1-z}\right)^{\alpha},$$

where the second equality follows from $h_B(z,p)_2 \equiv \frac{p_2 z}{p_2 z + p_1(1-z)}$. This expression also clearly depends on the prior.